

# ENHANCING CONSTRUCTION SAFETY MANAGEMENT THROUGH MULTIVARIABLE GREY MODEL ANALYSIS AND VARIABLE SELECTION OPTIMIZATION

Jian LIU<sup>1,2</sup>, Ye HE<sup>3</sup>, Rui FENG<sup>1,4</sup> , Qinlin CHU<sup>1</sup>

<sup>1</sup>*School of Resources and Safety Engineering, University of Science and Technology Beijing, 100083 Beijing, People's Republic of China*

<sup>2</sup>*Key Laboratory of High-Efficient Mining and Safety of Metal Mines of the Ministry of Education, University of Science and Technology Beijing, 100083 Beijing, People's Republic of China*

<sup>3</sup>*China Institute of Atomic Energy, 103413 Beijing, People's Republic of China*


<sup>4</sup>*Research Institute of Macro-Safety Science, University of Science and Technology Beijing, 100083 Beijing, People's Republic of China*

## Article History:

- received 9 January 2025
- accepted 20 May 2025

**Abstract.** In this study, a multivariable grey model (GM(1, N)) is employed to explore how different combinations of variables impact the accuracy of construction accident prediction, using a full permutation algorithm. The aim is to optimize variable selection and improve prediction accuracy. By conducting an exhaustive analysis of 511 potential combinations involving nine variables, it was observed that by integrating crucial external variables such as macroeconomic indicators and industry scale, the multivariable model achieved a prediction accuracy error rate of less than 0.5%, thereby significantly enhancing its information capture and forecasting precision. The analysis suggests that optimal predictive performance is achieved when the number of control variables is approximately four. Additionally, further research shows that increasing the dataset size significantly enhances the model's predictive capability. This study highlights the scientific rigor and precision of decision-making in preventing construction accidents and provides empirical evidence for construction safety management. The research in this paper not only enriches the connotation of the grey system prediction model theoretically, but also provides a data-driven decision support tool for urban construction and safety accident prevention in practice.

**Keywords:** multivariable grey model, construction safety management, variable selection optimization, prediction accuracy, data size effect.

 Corresponding author. E-mail: [fengr@ustb.edu.cn](mailto:fengr@ustb.edu.cn)

## 1. Introduction

The acceleration of global urbanization and the rapid development of the construction industry in recent times has led to an increase in construction accidents, becoming a significant societal issue that demands attention (Zhu et al., 2021). These accidents not only incur substantial economic losses but, more importantly, seriously threaten the safety of workers and the general public (Pinto et al., 2011). Among all industries, the fatality rate in the construction industry remains high (Yi et al., 2012), making it one of the most unsafe industries today (Alkaissy et al., 2020; Tam et al., 2004). This situation has raised widespread concerns, especially regarding the need to effectively prevent these accidents, presenting an urgent problem that needs to be solved (Chen et al., 2020).

Faced with the complexity of construction accidents and the challenges they present for prevention, there is a

growing emphasis on active accident management (Nini et al., 2020; Zhou et al., 2015), especially in the realm of accident prevention strategies. Research shows that the benefits of preventive measures far exceed the cost of accident prevention, with a ratio of approximately 3:1 (Ikpe et al., 2012). Therefore, it is imperative to provide data support for prevention strategies through accident prediction. Accurate predictions can not only help in identifying potential risk areas but also provide a scientific basis for formulating effective preventive measures (Fatemeh & Vedat, 2023; Wenli et al., 2023).

In 2020, there were a total of 407 construction accidents, representing 59.07% of the total incidents. These included 83 object strike accidents, accounting for 12.05%; 45 incidents involving lifting machinery injuries, making up 6.53%; 42 cases of earthwork and foundation pit col-

lapse, accounting for 6.53%; 26 construction equipment injury accidents, amounting to 3.77%; 22 electric shock accidents, comprising 3.19%; and 64 other accidents, making up 9.29% of the total (Ministry of Housing and Urban-Rural Development of the People's Republic of China, 2022). Compared with other high-risk industries, the construction sector possesses unique characteristics such as a high frequency of general accidents, a low occurrence of major accidents<sup>1</sup>, and diverse sources of hazards, making its safety management special. In view of this, strengthening research on construction accident prediction and integrating it into preventive strategies has become the focus of both academia and industry (Bing, 2022).

In history, methods for predicting construction accidents have included statistical analysis, machine learning techniques, expert systems, and so on. These methods each have their own advantages; Statistical analysis provides macro level data trend analysis, while machine learning offers complex pattern recognition capabilities. However, these methods typically require large datasets and high computational complexity. On the other hand, expert systems heavily rely on the subjective knowledge of experts, which may limit their objectivity and wide applicability. Grey system theory, as an alternative method, is particularly suitable for situations where uncertainty and limited information are common in construction accidents. Grey system theory, recognized as an effective tool for forecasting and decision-making for dealing with uncertainty and limited information, has been widely used in various fields, such as economics, sociology, and engineering (Chen & Tien, 1996; Hsiao & Liu, 2002; Hsu & Wen, 1998; Hsu, 2003; Lin & Yang, 2003; Song, 1992). For example, Deng (1989) applied grey system theory to analyze the safety status of construction projects, providing a new perspective for construction safety assessment. Chang et al. (1999) improved the modeling error of grey prediction by combining grey models. The grey prediction model has been applied to the global integrated circuit industry (Hsu, 2003), flood prediction (Trivedi & Singh, 2005), and vehicle mortality risk prediction (Mao & Chirwa, 2006), and has achieved good results. These studies have shown the potential of grey system theory in handling construction safety issues with uncertainty and limited data. Exploring the characteristics of accidents within the construc-

tion industry reveals significant potential for employing a grey prediction model in accident prediction (Y. Li & M. Li, 2015). Grey prediction is divided into univariate grey prediction and multivariate grey prediction (Cheng et al., 2023; Du et al., 2023; Lei & Wang, 2022; H. Wang & L. Wang, 2020; Xiong et al., 2021). While the univariate model demonstrates notable accuracy in predicting construction accident data with incomplete information and limited datasets (Sun & Liu, 2011), the prediction complexity of accident prediction, influenced by numerous interrelated factors, extends beyond the scope of a singular time series model (Chen et al., 2011; Tien, 2012; Wu & Chen, 2005). In order to capture these intricate relationships comprehensively and improve the accuracy and reliability of prediction, the adoption of a multivariate grey prediction model is an inevitable progression.

Multivariable grey prediction not only addresses single-factor prediction challenges but also synthesizes the interactions among multiple factors, thereby providing insights into more complex system dynamics. Ye et al. (2024) introduced a grey prediction model based on action time and intensity for China's food industry. Cheng et al. (2020) meanwhile, employed the GM(1, 3) model to simulate and predict clean energy consumption in China, identifying economic scale and population size as the main influencing factors. As an extension of grey prediction, the GM(1, N) model is suitable for forecasting scenarios characterized by limited data and incomplete information, influenced by numerous factors (Shanshan & Hazem, 2022). By discerning and extracting insights from incomplete information, the GM(1, N) model facilitates the discovery of patterns within complex and ever-changing practical problems, enabling relatively accurate predictions. Given these considerations, this study adopts a multivariate grey model (GM(1, N)) to address the limitations of traditional models.

Research and improvement of the GM(1, N) model mainly focuses on optimizing model parameters and structure. For example, Lao et al. (2021) analyzed optimal parameters by refining the background value, developing the DBGGM(1, N) model, and selecting population and GDP as influencing factors to predict China's energy and electricity consumption. Zeng (2018) improved parameters based on the fractional accumulation principle and minimized the average relative error of the system characteristic sequence, applying this approach to predict the output value of high-tech products in China and Guangdong Province. In addition, integrating other models to improve prediction accuracy is another research direction (Penghui et al., 2023). Li and Zhang (2024) combined the grey model with a neural network to create the NMGM(1, N) model, which accurately predicted China's per capita energy consumption by learning features directly from data samples. Some researchers have combined the grey model with a genetic algorithm, incorporating seasonal factors and time power terms. This adaptation addresses the insensitivity of the traditional multivariate grey model to seasonal fluctuations and nonlinear trends (Li et al., 2023). However, these

<sup>1</sup> According to the Regulations on Reporting, Investigation, and Handling of Production Safety Accidents, a general accident is defined as an incident resulting in fewer than 3 deaths, or fewer than 10 serious injuries, or direct economic losses totaling less than 10 million yuan. A major accident involves more than 3 but fewer than 10 fatalities, or more than 10 but fewer than 50 serious injuries, or direct economic losses ranging from more than 10 million yuan to less than 50 million yuan. A severe accident encompasses more than 10 but fewer than 30 deaths, or more than 50 but fewer than 100 serious injuries, or direct economic losses exceeding 50 million yuan but less than 100 million yuan. A particularly severe accident is one that results in more than 30 fatalities, causes over 100 serious injuries, or leads to direct economic losses of more than 100 million yuan.

optimizations primarily focus on parameter optimization and modeling mechanisms, with relatively limited research on variable selection for multivariate models. The GM(1, N) model includes one main variable representing system characteristics and N–1 variables representing influencing factors. Given the model's reliance on multiple variables, a more comprehensive analysis of variable selection is necessary for systems involving multiple influencing factors (Jianhong et al., 2024).

Most of the existing grey theory applications in construction safety mainly focus on micro-level analyses. However, from a macro perspective, understanding the overall trends and patterns of construction accidents across the entire industry and regions is crucial for formulating comprehensive safety policies and strategies. Macro-level research can provide insights into the relationship between the construction industry and the broader economic and social environment, which is essential for sustainable development. Firstly, it provides crucial information for policymakers and regulatory authorities to formulate and adjust safety policies and regulations. By understanding the potential trends of fatalities, they can set appropriate safety standards and allocate necessary resources more effectively, thus enhancing the overall safety level of the construction industry. Secondly, for construction companies, it serves as an important reference for them to improve their safety management systems. Predictions can help identify potential high-risk periods or projects ahead of time, enabling them to take proactive measures such as strengthening safety training, improving site supervision, and optimizing construction processes to reduce the occurrence of accidents and fatalities. Moreover, from a social perspective, it helps to raise public awareness of construction safety. The public can better understand the risks associated with construction activities and advocate for stronger safety measures, which in turn promotes a safety-conscious social environment.

In recent years, data-driven modeling and intelligent prediction approaches have gained significant traction in fields such as system optimization, classification, and safety forecasting. Various researchers have explored the integration of machine learning and feature selection techniques to improve predictive performance, interpretability, and system adaptability under uncertainty. Farghaly et al. (2020a) proposed a hybrid filter-based feature selection approach that combines Mutual Information (MI), Chi-square ( $\chi^2$ ), and Relief-F methods to automatically determine optimal thresholds for classification tasks. This method improves classification accuracy while reducing dimensionality and model complexity. Their strategy highlights the importance of quantifying feature relevance in prediction systems, which aligns conceptually with our exhaustive evaluation of variable combinations in GM(1, N) modeling based on relative error screening. In another study, Farghaly et al. (2020b) developed a hybrid associative classifier that integrates association rule mining with Support Vector Machines (SVM), enhanced by Sequential Forward

Selection (SFS) and Gini-index based pruning. Their work demonstrated that combining interpretable rules with robust learning algorithms can enhance performance and reduce redundancy. While their approach focuses on classification, the emphasis on input optimization and model interpretability resonates with our objective to evaluate the relative contribution of multiple variables in time-series grey models. In the energy optimization domain, El-messery et al. (2024) introduced a deep learning framework combining U-Net segmentation and CNN-based regression to estimate photovoltaic panel cooling efficiency from thermal images. The framework achieved high accuracy and supported non-invasive, real-time performance monitoring. Although the application domain and modeling paradigms differ, the study emphasizes the potential of data-driven predictive modeling in complex physical systems. Their work also reinforces the relevance of integrating data preprocessing, automated labeling, and multi-model comparison, which parallels our dynamic extension of GM(1, N) using sliding windows and permutation-based variable selection.

The purpose of this study is to develop a strategy for predicting construction accidents and selecting variables using the multivariate grey model (GM(1, N)), addressing the limitations of traditional univariate grey prediction models in complex data environments. By applying a full permutation algorithm and an exhaustive exploration strategy, 511 combinations generated by nine potential influencing variables are analyzed to predict the number of construction accident fatalities from 2017 to 2020. This approach seeks to identify the optimal number of prediction variables for the construction accident model. At the same time, considering the relationship between prediction accuracy and data size in the grey prediction model, this paper explores the optimal data size selection strategy during model construction to enhance the model's adaptability and accuracy. Through these comprehensive methods, this study not only optimizes variable combinations and improves prediction accuracy but also provides more scientific data to support the prevention of construction accidents, which is an essential step towards enhancing construction safety management and safeguarding the well-being of workers and the public.

## 2. Model construction

### 2.1. Data source

There are various indicators used to characterize accidents, such as the number of fatalities, the number of accidents, and the death rate per 100,000 people. Municipal housing engineering constitutes an important segment of the construction industry. This section provides statistics on accident data from the Chinese construction sector, issued by the Ministry of Housing and Urban-Rural Development of the People's Republic of China (Ministry of Housing and Urban-Rural Development of the People's Republic of China, 2022). It uses the death toll in housing municipal en-

gineering as the dependent variable to predict fatalities in construction accidents, thereby reflecting the state of construction safety. Many variables represent the development level of the construction industry, including indicators that reflect the overall construction economy, the scale of the construction industry, and the treatment of construction employees.

The data used in this study covers a ten-year period from 2010 to 2019. The data collection is focused on the Chinese construction market, which is one of the largest and most active construction markets globally, providing a rich and diverse dataset for analysis. The selection of this geographical scope is based on the availability and comprehensiveness of data, as well as the significance of the Chinese construction industry in the global context. By analyzing data from such a large and dynamic market, the results are expected to have broader applicability and representativeness.

Through correlation analysis between the collected variables and the number of fatalities caused by construction accidents, nine variables were identified with a grey correlation coefficient greater than 0.5. These variables fall into three categories: overall construction industry variables, construction project supervision variables, and survey and design industry variables, as shown in Figure 1. In traditional grey system theory applications, the selection of these influential variables has been primarily based on grey relational analysis, which measures the relationship between reference data and comparison data. However, this study introduces an innovative approach by employing a full permutation algorithm to identify the most significant variables affecting the model. This method allows for an exhaustive exploration of all possible combinations of variables, providing a more objective and comprehensive assessment of their impact on the model's predictive accuracy. The overall construction industry variables include the total output value, the number of enterprises,

and the number of employees. The total output value indicates the industry's economic contribution, the number of enterprise units represents enterprise distribution and market competition, and the number of employees reflects the industry's human resource scale. These variables reflect the economic vitality and market competition within the industry. The interaction of these factors influences engineering quality and safety standards. For example, intense market competition may lead some enterprises to neglect safety standards in efforts to cut costs, thus increasing the risk of accidents (Song et al., 2011). Table 1 shows the results of the ANOVA of the number of deaths versus year from 2000 to 2020, and the model corresponds to a p-value of < 0.0001, which is much smaller than the common significance level of 0.05, indicating that the independent variable as a whole has a significant effect on the dependent variable in the ANOVA. Figure 2 shows the corresponding heat map. The variables of the construction engineering supervision industry include operating income, the number of enterprises, and the number of registered practitioners. Operating income indicates the economic vitality of the supervision industry, the number of enterprises reflects the distribution of businesses in this sector, and the number of registered practitioners represents the professional workforce. These variables reflect the health and specialization of the industry, which directly affects the quality control of construction projects. An increase in the number of professionals contributes to better project quality, thereby reducing the accident rate (Gregersen et al., 2003; Simard & Marchand, 1994).

Table 1. ANOVA results for variable categories

Source	TSS	df	MS	F	p-value
Model	1023196.6933	3	341065.56443	22.90928	<0.001
Error	253090.25908	17	14887.6623	-	-
Total	1276286.95238	20	-	-	-

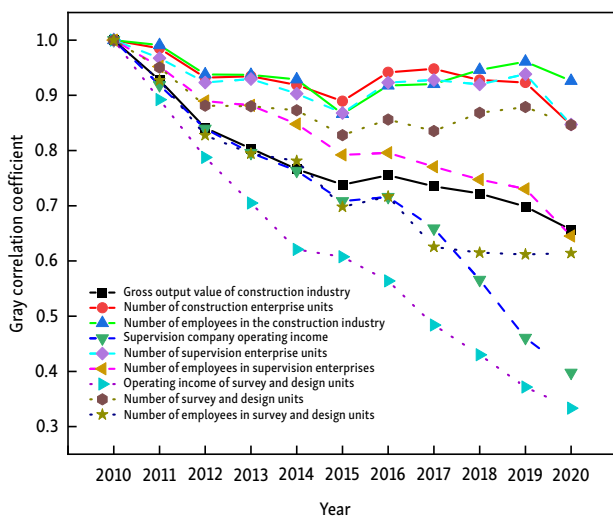
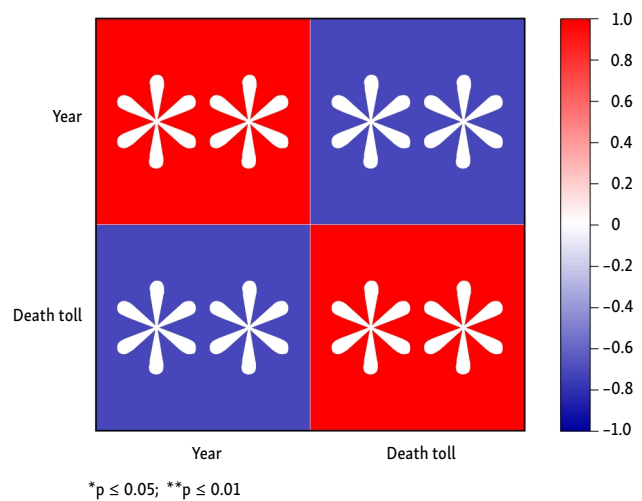


Figure 1. Grey correlation coefficients of nine predictor variables in three categories



\*p ≤ 0.05; \*\*p ≤ 0.01

Figure 2. Heatmap of correlation of deaths with year from 2000 to 2020

For the survey and design industry, the variables include operating income, the number of institutions, and the number of employees. Operating income reflects the industry’s economic benefits, the number of institutions indicates the market activity, and the number of employees shows the industry’s overall human resource scale. These variables demonstrate the scale and professional ability of the industry. High-quality design plays is crucial for construction safety, as excellent design can reduce structural defects and accident risks (Fonseca et al., 2014; Toole & Gambatese, 2008).

The dataset covers China’s construction industry statistics from 2010–2019, with the last year (2020) reserved for testing. We employed a rolling-window validation approach rather than random splitting to maintain temporal relationships. We have created Table 2 below.

**2.2. Construction of multivariate grey static model for construction accidents**

The GM(1, N) model is well-suited for prediction scenarios involving small data volumes and incomplete information, and it is affected by multiple factors. It is also widely used for predicting multivariate variables. In this model, “1” represents the single dependent variable, while “N” denotes the N independent variables involved, signifying the first order of the grey model with N variables. This model is suitable for analyzing the state of the system and the dynamics of variables, with a modeling and calculation process similar to the GM(1, 1) model. The modeling process is as follows.

Given a multivariate time-series dataset:

$$x_i^{(0)} = [x_i^{(0)}(1), x_i^{(0)}(2), \dots, x_i^{(0)}(p)] (p = 1, 2, \dots, n) \tag{1}$$

generate a cumulative sequence:

$$x_i^{(1)} = [x_i^{(1)}(1), x_i^{(1)}(2), \dots, x_i^{(1)}(p)]; \tag{2}$$

$$x_i^{(1)}(k) = \sum_{m=1}^k x_i^{(0)}(m), k = 1, 2, \dots, p, i = 1, 2, \dots, n. \tag{3}$$

The  $x_i^{(1)}$  sequence satisfies the following first-order linear differential equation model:

$$\frac{dx_1^{(1)}}{dt} + ax_1^{(1)} = b_1x_2^{(1)} + b_2x_3^{(1)} + \dots + b_{n-1}x_n^{(1)}. \tag{4}$$

According to the derivative definition of the derivative:

$$\frac{dx_1^{(1)}}{dt} = \text{Lim}_{\Delta t \rightarrow 0} \frac{x_1^{(1)}(t + \Delta t) - x_1^{(1)}(t)}{\Delta t}. \tag{5}$$

In a discrete form, the differential term can be written as:

$$\frac{\Delta x_1^{(1)}}{\Delta t} = \frac{x_1^{(1)}(k + 1) - x_1^{(1)}(k)}{k + 1 - k} = x_1^{(1)}(k + 1) - x_1^{(1)}(k). \tag{6}$$

Then, in Eqn (4),  $x_1^{(1)}$  takes the average of time  $k$  and  $k + 1$ , represented as  $\frac{1}{2}[x_1^{(1)}(k + 1) + x_1^{(1)}(k)]$ .

The discrete form of Eqn (4) is:

$$x_1^{(0)}(k + 1) + a \left[ \frac{1}{2} (x_1^{(1)}(k + 1) + x_1^{(1)}(k)) \right] = b_1x_2^{(1)}(k + 1) + \dots + b_{n-1}x_n^{(1)}(k + 1). \tag{7}$$

It is assumed that:

$$Y = \begin{bmatrix} x_1^{(0)}(2) \\ x_1^{(0)}(3) \\ \vdots \\ x_1^{(0)}(p) \end{bmatrix}; \tag{8}$$

$$B = \begin{bmatrix} -\frac{1}{2}[x_1^{(1)}(1) + x_1^{(1)}(2)] & x_2^{(1)}(2) & \dots & x_n^{(1)}(2) \\ -\frac{1}{2}[x_1^{(1)}(2) + x_1^{(1)}(3)] & x_2^{(1)}(3) & \dots & x_n^{(1)}(3) \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{2}[x_1^{(1)}(p-1) + x_1^{(1)}(p)] & x_2^{(1)}(p) & \dots & x_n^{(1)}(p) \end{bmatrix}; \tag{9}$$

$$\beta = \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_{n-1} \end{bmatrix}. \tag{10}$$

Then, Eqn (7) is abbreviated as:

$$Y = B^* \beta. \tag{11}$$

In this system of equations,  $Y$  and  $B$  represent the known quantities, while  $\beta$  serves as the undetermined parameter. Therefore, the least squares approximation can be obtained using the least squares method.

**Table 2.** Overview of variables and their associated categories

Feature	Description	Value/Range	Data Type	Time Period
Total observations	Number of yearly data points	10 years (2010–2019)	Time series	2010–2019
Dependent variable	Construction accident fatalities	500–900 deaths/year	Continuous	Annual
Independent variables	9 economic/industry indicators	See Table 3	Mixed (continuous, count)	Annual
Grey correlation threshold	Minimum correlation for inclusion	>0.5	–	–
Variable combinations	Total permutations analyzed	511 combinations	–	–

The solution can be determined as follows:

$$\hat{\beta} = (B^T B)^{-1} B^T Y = \begin{bmatrix} \hat{a} \\ \hat{b}_1 \\ \vdots \\ \hat{b}_{n-1} \end{bmatrix}. \quad (12)$$

By substituting the obtained  $\hat{\beta}$  back into Eqn (4), there are:

$$\frac{dx_1^{(1)}}{dt} + \hat{a} x_1^{(1)} = \hat{b}_1 x_2^{(1)} + \dots + \hat{b}_{n-1} x_n^{(1)}. \quad (13)$$

The discrete solution can be expressed as follows:

$$x_1^{(1)}(k+1) = \left[ x_1^{(0)}(1) - \frac{1}{a} \sum_{i=2}^n \hat{b}_{i-1} x_i^{(1)}(k+1) \right] e^{-\hat{a}k} + \frac{1}{a} \sum_{i=2}^n \hat{b}_{i-1} x_i^{(1)}(k+1). \quad (14)$$

Equation (14) represents the time response function model of the GM(1, N) model, serving as a concrete calculation formula for the grey prediction of the GM(1, N) model. The grey prediction model of the original series  $x_1^{(0)}$  is expressed as  $x_1^{(0)}(k+1) = x_1^{(1)}(k+1) - x_1^{(1)}(k)$ .

### 2.3. Construction of multivariate grey dynamic model for construction accidents

In the dynamic modeling approach of the multivariate grey prediction model, the model does not solely depend on all the original data. Instead, it takes into account multivariate factors that affect the prediction over time. Therefore, the model utilizes specific data sets to establish the initial prediction model and dynamically reconstructs it. By using part of the data set, the model can focus on the most representative information for the current prediction cycle. To update the data in the dynamic model, this process can be described using the following specific formula, ensuring that the model incorporates new data and replaces the old data at each step.

Define the data set within the sliding window, encompassing the data of  $n$  time units counted forward from the current time point  $t$ :

$$D_t = \{X_{t-n+1}, X_{t-n+2}, \dots, X_t\}. \quad (15)$$

Here,  $D_t$  denotes the data window at time  $t$ , and  $X_{t-i}$  represents the observed data at time  $t-i$ . Upon the arrival of new data  $X_{t+i}$ , update the data window by removing the oldest data  $X_{t-n+1}$  and adding the latest data:

$$D_{t+1} = D_t \cup \{X_{t+1}\} \setminus \{X_{t-n+1}\}. \quad (16)$$

This approach ensures that the window always retains the most recent  $n$  data points. This selective data usage strategy is similar to the sliding window method, where the data set progresses forward over time, constantly replacing the oldest data with new entries. This approach

not only improves the prediction's responsiveness and correlation but also provides a self-adjusting mechanism for the model to accommodate the potential complexities of nonlinear and multivariable systems.

The structure of the model needs to reflect the interplay among the multivariate variables and their impact on the prediction outcome. The modeling process of the GM(1, N) dynamic model, illustrated in Figure 3, encompasses four key steps: data preprocessing, parameter estimation, prediction output, and the iterative process of data updating. In the iterative update phase, old data is replaced with new data, ensuring that the model reflects the latest system state and the dynamic relationships between variables. When selecting a dynamic model, both the model's prediction efficiency and the influence of the dataset's size on prediction accuracy must be taken into account. Unlike static models, which rely solely on fixed datasets, optimizing dynamic models entails determining the optimal dataset size to maximize prediction accuracy.

The flexibility and updating ability of the model are crucial when implementing dynamic modeling for multivariate grey predictive models. Compared to traditional static models, dynamic models are more adept at handling rapidly changing environments and complex multivariable systems. Therefore, the dynamic modeling strategy for the multivariate grey prediction model is not merely about data elimination but focuses on optimizing the model's real-time performance and accuracy. In the context of basic experiments using MATLAB R2021b, second-hand workstations with minimal hardware configurations (costing approximately 200~300 \$) are adequate. For more demanding high-performance tasks, systems with multi-core CPUs, substantial RAM, dedicated GPUs, and SSD storage (ranging from 700~900 \$) are recommended. Remarkably, the program developed for this study is capable of perform-

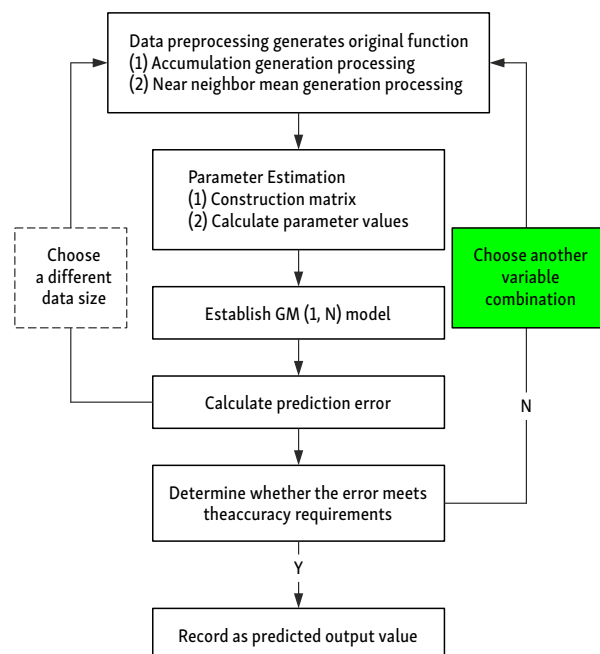


Figure 3. Dynamic modeling process of the GM(1, N) model

ing batch processing of data, achieving significant computational efficiency. It can process and generate results for over one hundred data points within approximately 6 seconds, thereby demonstrating both cost-effectiveness and high-speed processing capabilities. The implementation details of the GM(1, N) model are provided in Appendix.

## 2.4. Construction process of multivariate gray model for construction accidents

The proposed multivariable grey prediction framework GM(1, N) adopts a comprehensive and systematic architecture to enhance predictive accuracy, model robustness, and analytical depth, as illustrated in Figure 4. The modeling process begins with data input and preprocessing, wherein historical data from 2010 to 2019 are collected from authoritative sources such as government statistical yearbooks and industry safety reports. The dependent variable is defined as the annual number of fatalities in municipal housing engineering projects, while nine independent variables are rigorously selected based on grey relational analysis with a correlation threshold above 0.5, ensuring statistical significance and practical relevance. The data are normalized and structured into a multivariate time series format to eliminate scale discrepancies, reduce noise, and improve model stability. To exhaustively explore the input space, a full permutation algorithm is employed to generate 511 unique combinations of independent variables. Each combination is used to construct a static GM(1, N) model involving grey accumulation, least squares parameter estimation, and derivation of the time-response function to forecast the 2020 fatality trend. To address temporal variability and enhance adaptability, a dynamic GM(1, N) extension is introduced by incorporating a sliding time window mechanism that enables continuous parameter updating as new data become available. The evaluation and optimization module applies relative error metrics to assess model performance, retaining only those combinations with errors below 5% as optimal. Further statistical analyses are conducted to determine the ideal number of predictors, evaluate the influence of dataset size, and quantify the contribution of each variable. In the final output and interpretation phase, empirical findings are synthesized into actionable insights, identifying

dominant risk factors and offering evidence-based recommendations for construction safety governance, including targeted resource allocation and preventive strategy development. Through the integration of static and dynamic modeling with rigorous validation, this framework ensures high predictive precision, adaptability, and practical value in real-world applications.

## 3. Predictive results

In a previous study, the construction process of the multivariable grey prediction model (GM(1, N)) was extensively investigated. In order to leverage these findings to enhance construction safety, we can identify key influencing factors from the predictive results and formulate targeted preventive measures accordingly. In general, the purposes of macro prediction in the construction field can be divided into: preparation, planning, and hazard identification. During the preparation stage, through prediction, risk factors that may lead to serious accidents can be identified in advance, and targeted preventive measures can be formulated. Thus, during the pre-construction and construction processes, targeted technical improvements, safety training, equipment upgrades, and other measures can be taken to reduce the likelihood of accidents and reduce fatal accidents at their source. It helps construction enterprises and management departments determine key supervision areas and projects, increase safety investment and supervision efforts for high-risk projects, ensure that the construction process complies with safety standards, and effectively prevent accidents and protect the lives of workers. In terms of planning, it achieves optimized resource allocation. Based on the prediction results, safety resources can be allocated reasonably to improve resource utilization and make safety work more efficient and economical. During the hazard identification stage, targeted inspection strategies are formulated based on the prediction results, and the hazard identification plan is dynamically adjusted according to the changes in the predicted number of fatalities. If the risk of the number of fatalities in a certain stage is predicted to increase, the inspection intensity of that stage should be strengthened in a timely manner; conversely, the resource allocation should be optimized reasonably to avoid waste caused by excessive inspection. For instance, by analyzing the impact of different variable combinations on the predictive outcomes, we can pinpoint the factors that most significantly affect the number of construction accidents and fatalities, thereby taking measures to mitigate these risks during the planning and design phases. In this section, the focus shifts to applying this model to predict two key indicators: the number of construction accident fatalities and the total number of accidents. These two indicators serve as significant parameters for assessing the safety status of the construction industry. The number of fatalities directly impacts life safety, highlighting the critical importance of accurate prediction results for developing effective prevention strategies and

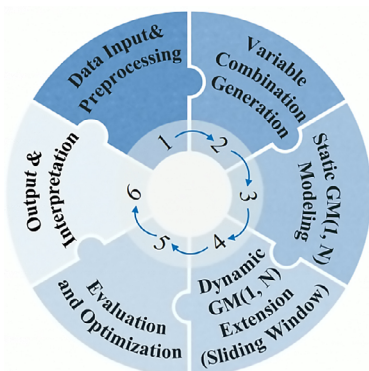


Figure 4. Workflow of the proposed GM(1, N) model

rescue plans. Additionally, it provides accident frequency data, offering insights into the overall effectiveness of the safety management system.

When making a prediction, choosing the optimal combination and number of predictor variables is crucial for improving prediction accuracy. By analyzing the effects of different variable combinations on prediction results, it is possible to identify the factors with the greatest impact on the number of construction accidents and deaths, thereby enhancing the model's explanatory power. Moreover, selecting the appropriate number of variables helps prevent model overfitting and ensures the generalization ability and practical value of the prediction results. Therefore, this section focuses on historical data from 2010 to 2019 and discusses how to predict the number of construction accident fatalities and accidents in 2020 by selecting the most appropriate number and combination of variables to attain improved predictive accuracy.

### 3.1. Prediction of fatalities in construction accidents

#### 3.1.1. Best prediction combination of construction accident fatalities

In this study, an exhaustive exploration strategy was adopted, and the full permutation algorithm generated 511 possible combinations of nine variables. Through a comprehensive analysis of these combinations, we can identify the key variables that have the greatest impact on the prediction of construction accident fatalities. This information is crucial for devising effective safety management measures, such as increasing regulatory oversight of enterprises with significant economic contributions to the construction industry and enhancing the quality of safety training for practitioners, thereby reducing accidents at their source. A combination of one variable is labeled as  $\text{Var}^{\text{one}}$ , a combination of two variables as  $\text{Var}^{\text{two}}$ , and so on, up to a combination of nine variables labeled as  $\text{Var}^{\text{nine}}$ . For example:

$$\text{Var}_1^{\text{one}} = X_1, \text{Var}_2^{\text{one}} = X_2, \dots, \text{Var}_1^{\text{two}} = X_1 X_2, \text{Var}_2^{\text{two}} = X_1 X_3.$$

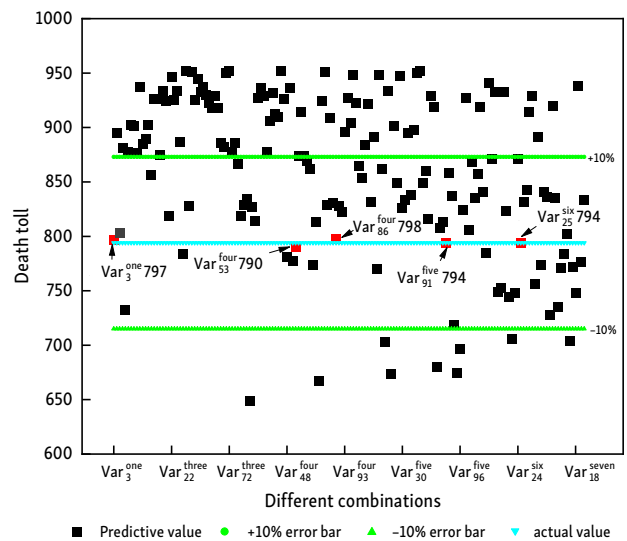
**Table 3.** Overview of variables and their associated categories

Variable	Definition	Unit	Associated Category
$X_1$	Total output value of the construction industry	10 billion yuan	Overall construction industry variables
$X_2$	Number of construction enterprise units	100	Overall construction industry variables
$X_3$	Number of employees in the construction industry	100,000 people	Overall construction industry variables
$X_4$	Operating income of construction project supervision	100 million yuan	Construction project supervision variables
$X_5$	Number of construction project supervision enterprises	10	Construction project supervision variables
$X_6$	Number of registered practitioners in construction project supervision enterprises	100 people	Construction project supervision variables
$X_7$	Operating income of survey and design units	10 billion yuan	Survey and design industry variables
$X_8$	Number of survey and design institutions	10	Survey and design industry variables
$X_9$	Number of employees at the end of the year in survey and design institutions	1000 people	Survey and design industry variables

To provide a clear overview of the variables used in our analysis and their relationships, we have created Table 3 below.

Using the data from these different combinations for the years 2010–2019 as simulated data, the study aimed to predict the number of construction accident deaths in 2020. The comparison between the predicted and actual values for different combinations with a relative error of less than 20% is shown in Figure 5. Table 4 presents a comprehensive summary of the calculated metrics, their formulas, and the values derived from the GM(1, N) model analysis.

When using the GM(1, N) model to predict the number of deaths in construction accidents in 2020, the model exhibited varying levels of accuracy. The actual number of deaths in construction accidents in 2020 was 794. In some prediction combinations, the model did not achieve the expected accuracy, with some prediction errors exceeding 10%. For example, the prediction error for the combination  $\text{Var}_{30}^{\text{two}}$  was as high as 18.01%. These results indicate



**Figure 5.** Combinations of variables with a relative error of less than 20% in predicting the number of construction accident fatalities in 2020



**Table 4.** Metrics calculated in experimental results

Metric	Formula/Definition
Relative error	$RE = \frac{\  \text{Actual} - \text{Predicted} \ }{\text{Actual}} \times 100\%$
Grey relational grade	$Y_i = \frac{1}{n} \sum_{k=1}^n \frac{\min \  x_0(k) - x_i(k) \  + \rho \max \  x_0(k) - x_i(k) \ }{\  x_0(k) - x_i(k) \  + \rho \max \  x_0(k) - x_i(k) \ }$
Mean absolute percentage error	$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{\  \text{Actual}_i - \text{predicted}_i \ }{\text{Actual}_i} \times 100\%$
Consistency function	$RV_k = \frac{r_k - R_{\min}}{R_{\max} - R_{\min}}$
Optimal data size	Empirical analysis of RE vs. data size

that the size of the prediction error may be influenced by many factors, such as the interaction effect between variables and changes in the external environment. Therefore, selecting the correct data and understanding the relationship between variables are crucial during the modeling process.

By adjusting the model with different combinations of variables and estimating the parameters, it is possible to better adapt to changes in the system state and improve prediction accuracy. For example, after adjustments, there were five combinations with a relative error of no more than 0.5%. In the prediction combinations  $\text{Var}_{91}^{\text{five}}$  and  $\text{Var}_{29}^{\text{six}}$ , the model's predicted values were completely consistent with the actual number of deaths. This indicates that when the model parameters are precisely adjusted to fit a specific combination of variables, the predictive ability of the model is significantly improved. This high degree of accuracy not only demonstrates the appropriateness of the model structure but also highlights its ability to capture the interactions between complex factors related to the number of fatalities in construction accidents when suitable variables are selected.

### 3.1.2. Optimal number of predictive variables for construction accident fatalities

After analyzing the prediction results based on the GM(1, N) model, it was observed that the accuracy of predictions changes with the number of variables included. Different combinations, from  $\text{Var}^{\text{one}}$  to  $\text{Var}^{\text{nine}}$ , represent increasing numbers of variables, with  $\text{Var}^{\text{one}}$  representing a single variable and  $\text{Var}^{\text{nine}}$  representing a combination of nine variables. The prediction results varied accordingly. For example, combinations  $\text{Var}^{\text{eight}}$  and  $\text{Var}^{\text{nine}}$  had relative errors above 70% and were unsuitable as predictor variables. In order to find the optimal number of predictor variables, the number of different combinations with relative errors below 20% was counted and categorized into different error intervals, as shown in Figure 6 and Table 5.

According to the prediction results, 164 out of 511 combinations had relative errors within 20%, with two combinations falling within the 0–5% error range. When

**Table 5.** Relative error distribution statistics for construction accident deaths using different numbers of predictor variables in 2020

Number of variables	0–5%	5–10%	10–20%	Total
1	2	0	1	3
2	0	1	8	9
3	5	2	27	34
4	14	6	28	48
5	9	12	18	39
6	5	10	6	21
7	6	2	2	10
Total	41	33	90	164

the number of variables increased to 2, there were no combinations with errors in the 0–5% range, and most errors were within the 10–20% range. For combinations of 3 to 4 variables, the prediction error was primarily within the 10–20% range. However, there was a significant increase in the number of combinations within the 0–5% error range, especially with four-variable combinations, which has the most predictions in the 0–5% range. Beyond 5 variables, the accuracy decreased compared to the four-variable combinations, especially in the lowest error interval. At the same time, comparing the prediction results of the four-variable combinations in Table 5 with the actual value, it was found that most four-variable combinations had small prediction errors, indicating that the model can provide more accurate predictions with four variables.

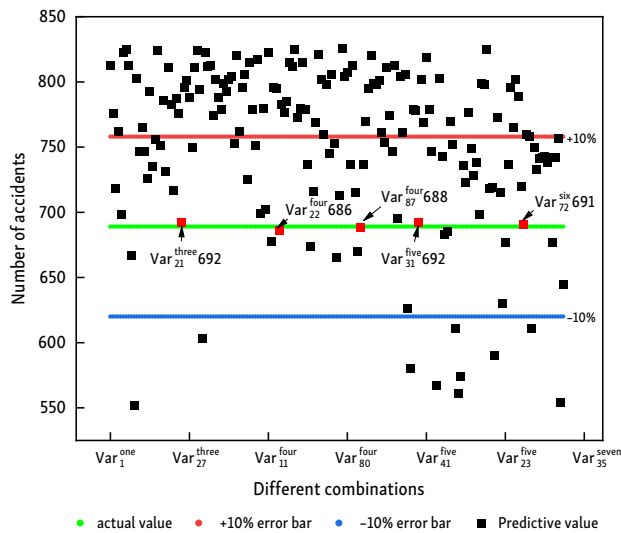
The number of selected variables is directly related to model complexity. Too many variables can lead to overfitting, while too few may fail to capture all important features of the data. In predicting the number of construction accident fatalities, using a combination of four variables achieves a balance by providing sufficient information to capture key factors while avoiding overfitting due to excessive variables. This balance reduces the computational cost of model training and prediction. The four-variable model accommodates the nonlinear relationships in the data while maintaining enough simplicity to avoid unnecessary complexity, making it the optimal choice for predicting construction accident fatalities in 2020.

## 3.2. Prediction of the number of construction accidents

### 3.2.1. Best prediction combination of the number of construction accidents

In order to comprehensively evaluate the impact of different variable combinations on the prediction of construction accidents, 511 combinations of nine variables were estimated using data from 2010 to 2019. The model then predicted the number of accidents in 2020. Figure 6 displays the combinations with a relative error of less than 20%. By comparing these predicted values with the actual number of accidents, it provides an intuitive assessment of which variable combinations are most effective in predicting construction accidents.

Among the 511 combinations, 173 have a relative error of less than 20% for predicting the number of accidents. The actual number of construction accidents in 2020 was 689. The five groups with the smallest prediction errors were 0.44% for  $\text{Var}_{21}^{\text{three}}$ , 0.44% for  $\text{Var}_{22}^{\text{four}}$ , 0.15% for  $\text{Var}_{87}^{\text{four}}$ , 0.44% for  $\text{Var}_{31}^{\text{five}}$  and 0.29% for  $\text{Var}_{72}^{\text{six}}$ . Similar to the prediction of accidents fatalities, the best prediction combinations often involve four variables, indicating that the model parameters are particularly well-suited for these combinations. When the number of variables in the



**Figure 6.** Variable combinations with a relative error of less than 20% in predicting the number of construction accidents in 2020

**Table 6.** Model parameters and settings

Parameter	Description
Data size ( $n$ )	Number of years included in the dataset for model training and validation
Dependent variable ( $x_1^{(0)}$ )	System characteristic variable (main predicted output)
Independent variables ( $N$ )	Influencing factors (e.g., industry scale, supervision metrics)
Grey relational coefficient ( $\gamma$ )	Threshold for variable inclusion
Model coefficients ( $a, b_2, b_3, \dots$ )	Parameters estimated via least squares
Background value ( $\alpha$ )	Weighting factor for cumulative sequence generation in grey modeling
Error threshold ( $\epsilon$ )	Maximum allowable relative error for model validation
Sliding window size ( $\omega$ )	Dynamic model update interval for real-time prediction

combination exceeds seven ( $\text{Var}^{\text{eight}}$  and  $\text{Var}^{\text{nine}}$  combinations), the relative error of the prediction rarely falls below 20%. This indicates that these models cannot fully reflect the changes in the number of accidents. Therefore, in order to explore the impact of the number of predictor variables on the accuracy of accident predictions, further analysis of empirical results is necessary to determine the optimal number of predictor variables. Table 6 shows the parameter summary table, which provides a comprehensive list of each parameter and can be traced directly back to the methods section.

### 3.2.2. Optimal number of predictive variables for construction accident fatalities

In Section 3.2.1, a detailed discussion was conducted on the results of predicting the number of construction accidents in 2020 using various combinations of variables, aiming to identify the optimal combinations in terms of performance. These combinations demonstrate exceptional prediction accuracy, providing strong evidence regarding which variable combinations are most effective in predicting the number of accidents. However, for further optimization of the predictive model to enhance its practicality and efficiency, it is crucial to determine the ideal number of variables for constructing the predictive model. Therefore, the following section will delve into a comprehensive analysis to understand how different numbers of variables affect the predictive capability of the model. Table 7 provides a summary of the counts of various combinations with a relative error below 20% across different error ranges.

The table presents error distribution data for predicting construction accidents using one to nine variables. It illustrates the frequency of prediction errors within the 0–5%, 5–10%, and 10–20% error ranges across different variable combinations, offering insights into each variable's predictive efficacy. Analysis reveals that the univariate model (one variable) generally yields substantial errors, especially within the 10–20% range, indicating insufficient predictive capacity for complex accidents. As the number of variables increases to two and three, the prediction accuracy improves but remains largely imprecise. Notably, the four-variable model demonstrates the highest performance, especially within the 0–5% error range, indicating its effectiveness in capturing accident complexities and providing

accurate predictions. However, when the number of variables exceeded four, performance in the lowest error interval (0–5%) did not see any significant improvement nor slight decrease. This may be because the additional variables may not improve the prediction performance of the model, but could instead lead to overfitting, making the model unable to generalize to new data effectively.

**Table 7.** Relative error distribution statistics of the number of accidents under different numbers of variables in 2020

Number of variables	0–5%	5–10%	10–20%	Total
1	2	0	4	6
2	1	5	6	12
3	3	6	32	41
4	11	7	35	53
5	5	9	18	32
6	6	3	9	18
7	1	8	0	9
8	0	1	1	2
Total	29	39	105	173

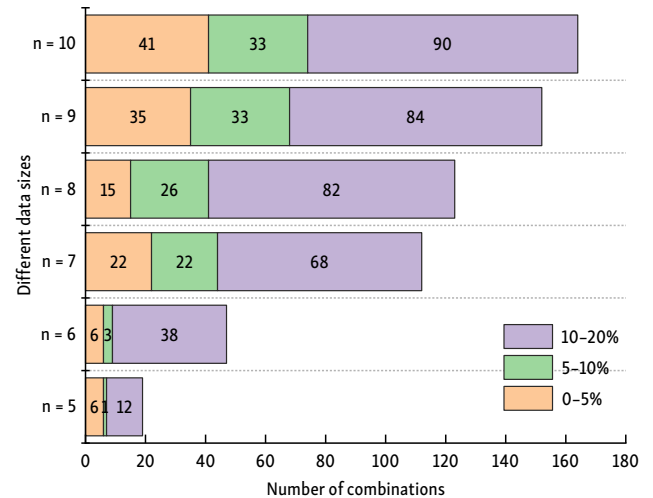
In conclusion, consistent with the prediction of fatalities in construction accidents, the four-variable model demonstrated the optimal balance in this study, providing the most effective approach to achieve high prediction accuracy while maintaining model simplicity.

## 4. Discussion

### 4.1. Effect of data size on the prediction

In the previous section, the data from 2010–2019 was utilized to predict the number of construction deaths in 2020, showcasing the prediction results based on a 10-year data span. This section will expand the data size to explore how different data sizes impact the accuracy of the prediction model. By incorporating data from diverse time series to expand the data size, the impact of time span on model performance was evaluated. The distribution of relative errors for various data sizes in 2020 is shown in Figure 7. Expanding the data size aids in comprehending the model's adaptability to different time scales and data densities. By comparing the model's performance across various data sizes, the most suitable data configuration to provide more accurate predictions can be discerned.

When discussing the impact of data size on predictive accuracy, we must consider not only how to improve the model's predictive power but also how to utilize this information to optimize safety management practices. By expanding the scope of data, we can gain a more comprehensive understanding of the trends in construction accidents, leading to more targeted policy-making and resource allocation decisions. When assessing the impact of data size on prediction accuracy, the number of predictions within the low error range (0–5%) increased as the data size increased, particularly notable within the 0–5%



**Figure 7.** Distribution of relative errors in fatalities across various data sizes

and 5–10% error ranges when the data size reached  $n = 7$ . This flexibility in data requirements is an advantage of the full permutation algorithm, which does not impose specific demands on the origin or characteristics of the data. This feature enhances the model's generalizability and could potentially apply to other countries with different datasets, thus broadening the applicability of our findings globally. This could be attributed to the availability of sufficient historical data, enabling the model to more effectively capture trends and periodicity in the data. Hence, larger data sizes may have a positive effect on the accuracy of the prediction model.

In constructing the grey prediction model, determining the appropriate data size is crucial. The increase in accuracy resulting from larger data size stems from the model's ability to use more historical information for learning the underlying patterns and periodicity in the data. In the case of the univariate grey prediction model, optimal performance is often observed with moderate data sizes, such as  $n = 7$  (Wang & Song, 2019). However, in the multivariate grey prediction model, although the number of prediction combinations with low errors significantly increases at the medium size of  $n = 7$  and gradually stabilizes, it continues to exhibit improvement, with the best performance observed at a larger data size, particularly at  $n = 10$ . This phenomenon may be attributed to the fact that in the univariate scenario, where the model only considers the variable itself, it struggles to capture all the system dynamics. Extensive data size may cause the model to focus on atypical historical fluctuations, thereby affecting prediction accuracy. Hence, the simplicity of the model in capturing and learning the main trends in the data is crucial. As the data size increases, beyond a certain critical point, additional noise may be introduced, leading to a reduction in prediction accuracy. Thus, the univariate model reaches equilibrium at moderate data sizes, achieving a balance between capturing the system's main trend and avoiding overfitting.

In contrast, the multivariate grey prediction model demonstrated its optimal performance over a span of 10 years. This suggests that by incorporating more relevant variables, the model can gain additional information from the increased amount of data, aiding in capturing more complex system dynamics and interactions between variables. The increased data size provides the multivariate model with richer information, allowing more precise parameter estimation and thus improving prediction accuracy. Moreover, the multivariate model exhibits improved noise filtration capabilities, improving its generalization ability on larger datasets. Currently, the largest data size limit in our study is 10 years based on the scope and quality of available data. Future studies could explore the possibility of extending the data size. However, careful consideration must be given to the relevance of the data and the adaptability of the prediction model to avoid compromising its generalization ability through overfitting.

Overall, the differences in data size selection between univariable and multivariable grey prediction models highlight the importance of balancing data size and model complexity during the modeling phase. Optimal data size selection requires considerations of data quality, system complexity, and model applicability. Therefore, when making predictions, exploring the optimal data size is crucial for improving both the predictive accuracy and reliability of the model. This study is constrained by a decadal dataset, which may limit the identification of long-term trends and complex system dynamics in the construction industry.

## 4.2. Variable importance assessment

In Section 4.1, the prediction of 2020 fatalities under different data sizes revealed significant disparities in the prediction results across different variable combinations (from

$\text{Var}_1^{\text{one}}$  to  $\text{Var}_1^{\text{nine}}$ ), indicating varying importance of individual variables for prediction. Subsequently, in the following analysis, the prediction results from 4.1 were used to count the frequency of each variable's occurrence with a relative error of below 5% across different data sizes, as shown in Figure 8. Because the multivariate grey prediction model exhibited stabilization at  $n = 7$ , statistical data were collected for sizes ranging from 7 to 10. This qualitative analysis aimed to assess the contribution of each variable to the prediction results by counting the occurrences of different variables with relative errors below 5%.

In assessing the impact of different variables on predictive outcomes, we have found that certain variables play a key role in reducing prediction errors. These findings can assist us in identifying areas that should be given focused attention in construction safety management, such as intensifying regulatory oversight in specific industries or enhancing the quality of safety training for certain positions. Based on the data presented in Figure 8, the variables exhibiting high importance in reducing prediction errors can clearly be identified. For example, as the data size increases, the number of occurrences of some variables also significantly increases. Moreover, the number of occurrences of  $X_5$ , representing the number of construction engineering supervision companies, increases from 6 times when  $n = 7$  to 32 times when  $n = 10$ . This suggests that the contribution of this variable to the model increases with the increase in the data size. Conversely,  $X_9$ , indicating the number of employees at the end of the survey and design agency, appears 18 times when  $n = 7$  and 17 times when  $n = 10$ , indicating it is not very sensitive to changes in data size.

When analyzing statistical data sizes ranging from 7 to 10, the occurrences of different variables within a relative error of less than 5% were examined. It was observed that the variable  $X_5$ , representing the number of construction

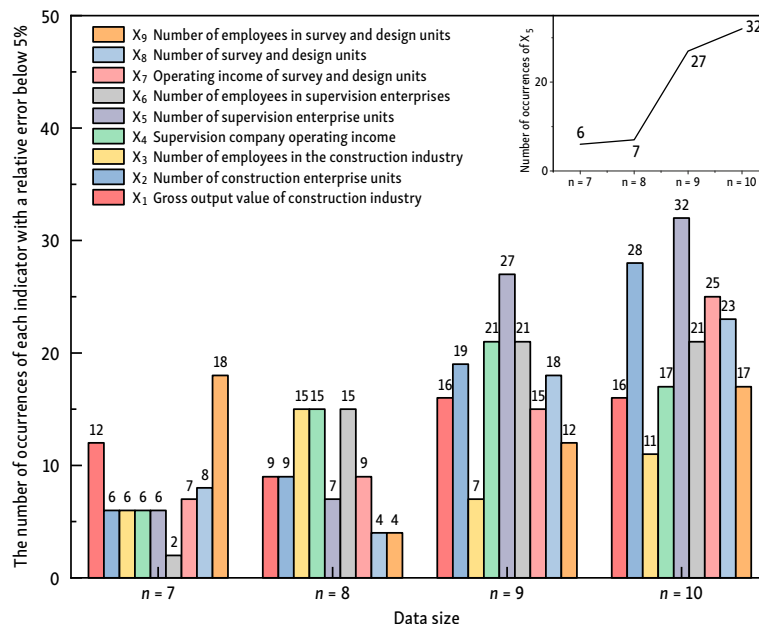


Figure 8. Occurrence count of different variables with relative errors below 5%

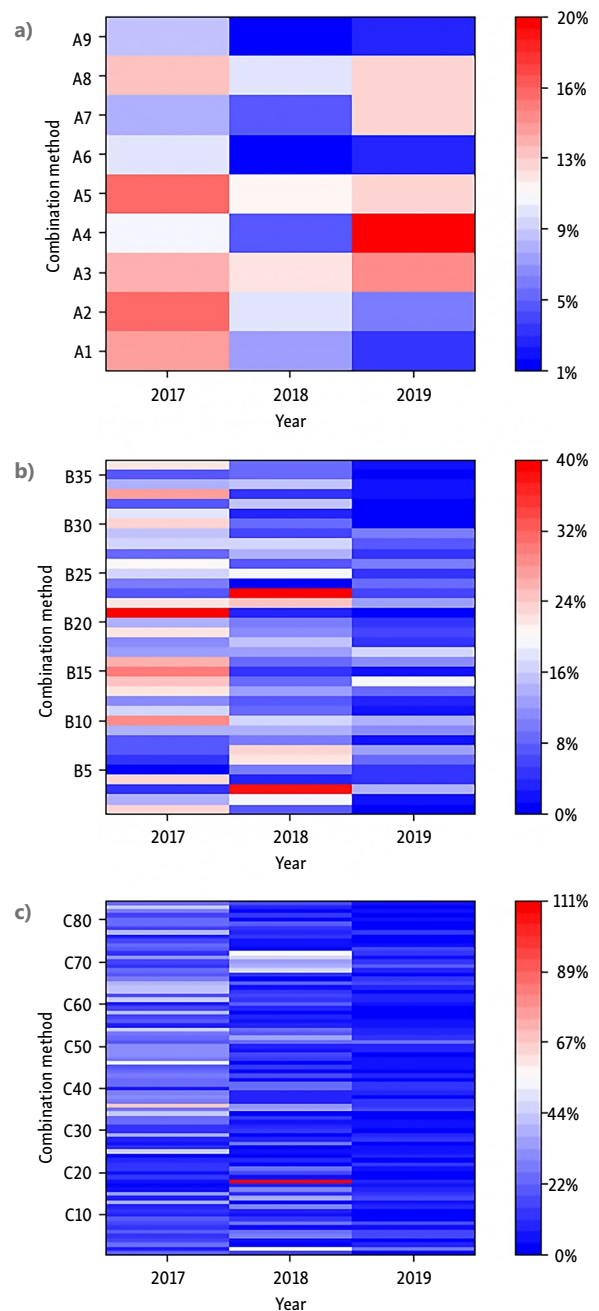
engineering supervision companies, appeared most frequently in the prediction results with an error of less than 5%, occurring 72 times. This statistical result highlights the importance of the  $X_5$  variable in improving the prediction accuracy of the model, suggesting a pivotal role of construction engineering supervision companies in mitigating the mortality rate of construction accidents. The responsibilities of supervision companies typically include monitoring the quality, progress, and cost of construction projects to ensure compliance with established standards and safety protocols. Therefore, the effective operation of supervision companies is closely related to accident reduction and safety improvement.

Overall, the occurrence number of the 9 variables demonstrated a steady increase, confirming the stability and reliability of the multivariate grey prediction model with a larger dataset. These results not only strengthen the basis of data-centric decision-making but also provide an empirical foundation for optimizing variable selection in predictive models for construction industry accident fatalities. Moving forward, according to the frequency of different variables, more attention should be given to variables of higher importance, while those with lesser predictive contributions ought to be either excluded or given reduced weight. Although the grey correlation threshold ( $\gamma > 0.5$ ) selected nine key variables, this strict criterion might overlook other potentially influential predictors with slightly lower correlations, which could further improve the model's accident prediction accuracy.

### 4.3. Prediction contributions of different historical years

In Section 4.1 of this study, an enhanced version of the multivariate grey prediction model was developed to predict the model's performance as the dataset expands. Based on this finding, the prediction accuracy of the model across different years was explored. For this purpose, a model using a decade-long dataset, spanning from the initial ten years, was employed to predict the number of construction accident deaths in various subsequent years. By employing grey prediction techniques from 2017 to 2019, the relative prediction errors were calculated for different combinations, as shown in Figure 9. The relative error for each combination in the prediction from 2017 to 2019 is detailed in the Appendix. The following figure illustrates the results of one variable (9 groups), two variables (36 groups), and three variables (84 groups) predicting the number of deaths resulting from construction accidents from 2017 to 2019. Through assessing the model's performance against recent year data, the impact of prediction accuracy from 2017 to 2019 on the accuracy of predictions for 2020 was explored.

The prediction outcomes reveal varying results across different variable combinations ( $\text{Var}_1^{\text{one}}$  to  $\text{Var}_1^{\text{nine}}$ ) for the years 2017, 2018 and 2019. selected combinations of the multivariate grey prediction model are shown to visualize the prediction results for the years 2017–2019, as illustrated in Figure 10.



**Figure 9.** Relative errors in predicted fatality numbers from 2017 to 2019 under different combination conditions: a – one variable (9 groups); b – two variables (36 groups); c – three variables (84 groups)

The performance of the same variable in different combinations varied significantly. For example, the variable  $X_7$ , operating income of survey and design units, appears in the combinations  $\text{Var}_5^{\text{two}}$ ,  $\text{Var}_{10}^{\text{three}}$ , and  $\text{Var}_6^{\text{five}}$ , but the relative errors of these combinations fluctuate greatly. This indicates that the influence of this variable on the modeling results is highly uncertain across different variable combinations. In a multivariate model, each variable influences the prediction results through a network of interactions rather than in isolation. The impact of a single variable is affected by the combination in which it is included. For example, when the operating income of survey

and design units is incorporated alongside other industry economic indicators, such as the total output value of the construction industry and the scale of the construction market, these variables may produce synergistic effects. For example, the expansion of the market scale can increase the demand for design services, thereby increasing the operating income of design units and influencing project capital allocation and quality control standards. When considered alongside the supervision industry, the predictive influence of the design unit may be diminished in the strict regulatory environment. Therefore, multivariate prediction enables a more comprehensive understanding of the interactions between variables.

When examining the relative errors of different combinations, varying trends over time can be observed. For example, the combination  $Var_6^{one}$  had a relative error of 9.42% in 2017, which decreased to 1.19% in 2018. Conversely, the combination  $Var_5^{two}$  had a relative error of 1.12% in 2017, but this increased significantly to 9.40% in 2018. These changes indicate that the same combination can have different effects on prediction accuracy in different periods. Therefore, this paper explores the mutual influence of forecasting across different periods by analyzing the relative size of errors.

The relative error obtained for  $n$  different combinations of predictions for each year are  $r_1, r_2, r_3, \dots, r_n$ .

$$\text{Let } R_{\max} = \max\{r_1, r_2, r_3, \dots, r_n\}, R_{\min} = \min\{r_1, r_2, r_3, \dots, r_n\}.$$

Then, the relative position size of the relative error for the  $k$ th combination can be defined as a consistent function:

$$RV_k = \frac{(r_k - R_{\min})}{(R_{\max} - R_{\min})}. \tag{17}$$

For each year from 2017 to 2020, there are  $n$  consistency functions  $RV_k$ :

- The consensus function defining the relative error in 2017 is the  $RV_{k1}$ ;
- The consensus function defining the relative error in 2018 is the  $RV_{k2}$ ;
- The consensus function defining the relative error in 2019 is the  $RV_{k3}$ ;
- The consensus function defining the relative error in 2020 is the  $RV_{k4}$ .

Previous predictions were often made by calculating the average relative error from 2017 to 2019 and selecting the prediction variable with the smaller average relative error. This approach means that the smaller the value of  $1/3 * RV_{k1} + 1/3 * RV_{k2} + 1/3 * RV_{k3}$ , the better the prediction effect for 2020. In order to explore the different prediction weights of 2017–2019 relative to 2020, this study defines:

$$RV_{kp} = c_1 * RV_{k1} + c_2 * RV_{k2} + c_3 * RV_{k3} + h, \tag{18}$$

where  $RV_{kp}$  is the predicted value of the consistent function for 2020,  $c_1, c_2$  and  $c_3$  are the coefficients representing the impact of 2017, 2018, and 2019 on the 2020 forecast, respectively, and  $h$  is a constant.

The analysis examined the relative error over four years, with the consensus function depicted in Table 8 for the period 2017–2020. This table allows for evaluation and comparison of the stability and reliability of predictive performance over these four years. The consensus function serves as an indicator of the predictive ability of each combination within each year, calculated from the prediction's relative error for each year. A lower value of this function indicates a smaller prediction error for the combination across different years.

Linear regression was conducted using the consensus function of 2017–2019 as the independent variable and the consensus function of 2020 as the dependent variable. The regression yielded the predicted value of the consensus function for 2020. The results of the linear regression analysis, compared with the actual values, are shown in Figure 11. The coefficient of the linear expression in this regression model quantifies the predictive contribution of different historical years (2017–2019) to the prediction for

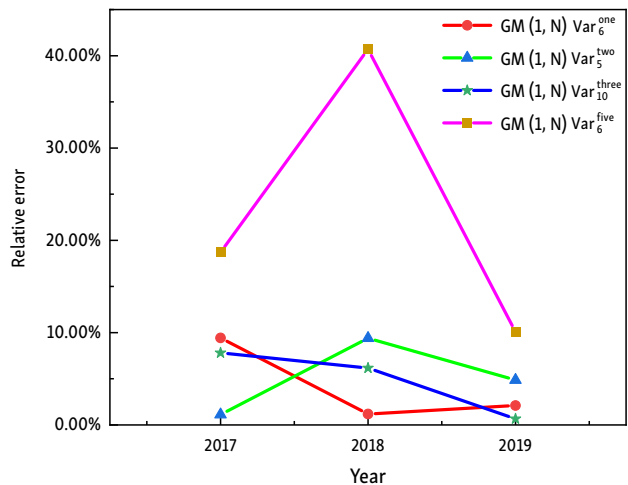


Figure 10. Differences in prediction accuracy under selected multivariate grey prediction combinations

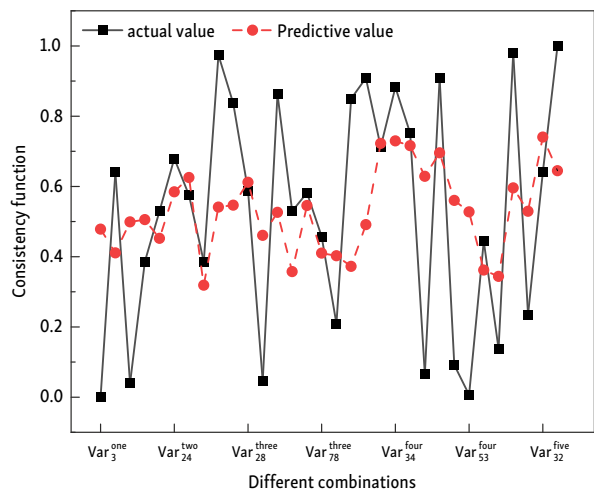


Figure 11. Comparison of predicted and actual values of the consistency function for predicting the number of fatalities in 2020 from 2017 to 2019

the year 2020. The linear expression is as follows:

$$RV_{k4} = -0.43*RV_{k1} - 0.032*RV_{k2} + 0.01*RV_{k3} + 0.765. \quad (19)$$

The results reveal that the relative error in 2017 holds the most significant influence on 2020, with a coefficient of  $-0.43$ . This indicates that when there is a large prediction error in 2017, the error in 2020 may be reduced,

**Table 8.** Consistency function for combinations with relative errors under 20% in fatality numbers from 2017 to 2020

Variable combinations	2017	2018	2019	2020
Var <sub>3</sub> <sup>one</sup>	0.64	0.56	0.86	0.00
Var <sub>5</sub> <sup>one</sup>	0.80	0.54	0.73	0.64
Var <sub>8</sub> <sup>one</sup>	0.60	0.46	0.71	0.04
Var <sub>17</sub> <sup>two</sup>	0.58	0.64	0.99	0.39
Var <sub>20</sub> <sup>two</sup>	0.70	0.51	0.27	0.53
Var <sub>24</sub> <sup>two</sup>	0.43	0.00	0.49	0.68
Var <sub>35</sub> <sup>two</sup>	0.30	0.40	0.04	0.58
Var <sub>9</sub> <sup>three</sup>	1.00	0.56	0.16	0.39
Var <sub>22</sub> <sup>three</sup>	0.53	0.03	0.67	0.97
Var <sub>26</sub> <sup>three</sup>	0.51	0.06	0.36	0.84
Var <sub>28</sub> <sup>three</sup>	0.33	0.53	0.73	0.59
Var <sub>30</sub> <sup>three</sup>	0.67	0.56	0.02	0.05
Var <sub>60</sub> <sup>three</sup>	0.49	1.00	0.31	0.86
Var <sub>73</sub> <sup>three</sup>	0.96	0.09	1.00	0.53
Var <sub>76</sub> <sup>three</sup>	0.46	0.73	0.12	0.58
Var <sub>78</sub> <sup>three</sup>	0.79	0.58	0.13	0.46
Var <sub>81</sub> <sup>three</sup>	0.80	0.58	0.00	0.21
Var <sub>24</sub> <sup>four</sup>	0.90	0.39	0.87	0.85
Var <sub>25</sub> <sup>four</sup>	0.64	0.10	0.33	0.91
Var <sub>32</sub> <sup>four</sup>	0.10	0.19	0.53	0.71
Var <sub>34</sub> <sup>four</sup>	0.08	0.17	0.53	0.88
Var <sub>39</sub> <sup>four</sup>	0.07	0.75	0.38	0.75
Var <sub>48</sub> <sup>four</sup>	0.31	0.21	0.45	0.07
Var <sub>50</sub> <sup>four</sup>	0.17	0.00	0.43	0.91
Var <sub>52</sub> <sup>four</sup>	0.45	0.55	0.72	0.09
Var <sub>53</sub> <sup>four</sup>	0.56	0.05	0.34	0.01
Var <sub>108</sub> <sup>four</sup>	0.90	0.71	0.82	0.44
Var <sub>122</sub> <sup>four</sup>	0.96	0.29	0.25	0.14
Var <sub>29</sub> <sup>five</sup>	0.33	1.00	0.28	0.98
Var <sub>31</sub> <sup>five</sup>	0.52	0.48	0.27	0.24
Var <sub>32</sub> <sup>five</sup>	0.00	0.92	0.51	0.64
Var <sub>46</sub> <sup>five</sup>	0.22	0.78	0.01	1.00

and vice versa, indicating an inverse correlation or reverse trend between these two years. Conversely, the relative error in 2018 has a lesser impact on 2020, indicated by a coefficient of  $-0.032$ , suggesting that the prediction accuracy in 2018 is not strongly correlated with accuracy in 2020. At the same time, the relative error in 2019 exhibits a weak impact on 2020 with a coefficient of  $0.01$ , indicating that an improvement in prediction accuracy in 2019 may slightly increase the accuracy of predictions for 2020.

In more distant years, a negative correlation is observed. This phenomenon is sometimes referred to as risk compensation theory (RCT) in the field of safety management (Wilde, 1982). When the safety of a system improves, individuals may engage in riskier behaviors, thereby offsetting the overall reduction in risk. Similarly, if forecasts indicate lower future accident rates, policymakers might become less vigilant and allocate fewer resources to safety measures, potentially resulting in an eventual increase in actual accident rates. The constant term in this prediction model is  $0.765$ , influenced by factors such as the decline in construction activity and the onset of the epidemic in 2020, as well as the rising number of deaths in 2017–2019. Therefore, a baseline prediction error value is established.

## 5. Conclusions

In multivariate grey prediction, selecting appropriate variables is essential to enhance prediction accuracy. The traditional multivariate grey prediction model, GM(1, N), typically relies on several variables with high correlation as predictive factors, often resulting in limited comprehensiveness. This study considered multiple potential influencing factors by exploring 511 combinations of nine variables, aiming to improve both the accuracy and interpretability of predictions. The specific conclusions are drawn as follows:

- (1) The study identified that using approximately four control variables yields optimal predictive accuracy. This finding underscores the importance of balancing model complexity with the number of variables included. By limiting the model to four variables, we were able to achieve a model that is not only accurate but also avoids overfitting, which could obscure the underlying relationships within the data.
- (2) The findings indicate that as the data size increased from 5 years to 10 years, the prediction accuracy improved significantly. This improvement was particularly pronounced with data sizes exceeding 7 years, where the model's predictive performance became more stable and accurate. Consequently, when using multivariate grey prediction, selecting a broader data range becomes crucial to obtain more comprehensive information.
- (3) An analysis of the model's predictive performance across different years revealed a correlation between past prediction accuracy and future predictions. The findings indicated an inverse relationship with the prediction accuracy of 2017 and a

positive relationship with that of 2018, highlighting the importance of considering historical prediction performance when forecasting future construction accidents.

The multivariate model has an error rate of less than 0.5% in prediction accuracy, which effectively improves the information capture ability and prediction accuracy. After a comprehensive evaluation of different combinations of variables, it was concluded that among all the combinations of variables, the prediction effect was best when the number of control variables was about four. The study further shows that in multivariate models, appropriate data sizes ( $n = 7$  to  $n = 10$ ) can minimize model complexity while maintaining prediction accuracy. At the same time, this study reveals the importance of variable selection, and quantifies the prediction contribution of historical years to subsequent years by defining a consensus function and assigning weights to the prediction effects of different years.

This study has provided initial research results by deeply analyzing the application of a multivariable grey prediction model in forecasting construction accidents, providing a new perspective and tool for accident prediction and risk management in construction projects. However, there is potential for further expansion and refinement in this area. One limitation is the shift from macro trends to micro level safety management practices, as the applicability of the model may vary due to differences in local regulations, cultural practices, and economic conditions. In addition, further investigation is needed into the selected variables and the mechanisms behind their impact on construction accidents. Future research can benefit from exploring the causal relationship between variables and accident outcomes to enhance the explanatory power and practical application of the model.

With the development and popularization of big data technology, the use of more extensive and multi-dimensional data resources, such as socio-economic data, project execution details, and environmental monitoring data, can further improve the comprehensiveness and accuracy of prediction models. By integrating this data, predictive models are able to more comprehensively capture the factors that influence the occurrence of safety incidents, enabling more accurate risk assessments. Therefore, future research should be committed to further improving the prediction effect of production safety accidents in housing municipal engineering through data-driven, so as to better serve the construction industry with China's housing municipal engineering as the core with the rapid development of urbanization, and provide scientific support and decision-making tools for the safety management of the construction industry.

## Acknowledgements

The authors gratefully acknowledge the support of the National Science Foundation of China (Grant No. 52004139) and the National Key R&D Program of China (No. 2017YFC0804901), and the Fundamental Research Funds for the Central Universities (No. FRF-TP-22-120A1).

## Author contributions

Authors Jian Liu and Ye He contributed equally to this work.

## References

- Alkaissy, M., Arashpour, M., Ashuri, B., Bai, Y., & Hosseini, R. (2020). Safety management in construction: 20 years of risk modeling. *Safety Science*, *129*, Article 104796. <https://doi.org/10.1016/j.ssci.2020.104796>
- Bing, W. (2022). Using an evidence-based safety approach to develop China's urban safety strategies for the improvement of urban safety: From an accident prevention perspective. *Process Safety and Environmental Protection*, *163*, 336–346. <https://doi.org/10.1016/j.psep.2022.05.018>
- Chang, T.-C., Chang, H. T., & You, M.-L. (1999). Inverse approach to find an optimum  $\alpha$  for grey prediction model. In *IEEE SMC'99 Conference Proceedings* (pp. 971–976). IEEE. <https://doi.org/10.1109/ICSMC.1999.814159>
- Chen, C.-K., & Tien, T.-L. (1996). A new transfer function model: The grey dynamic model GDM(2,2,1). *International Journal of Systems Science*, *27*(12), 1371–1379. <https://doi.org/10.1080/00207729608929343>
- Chen, F., Wang, H., Xu, G., Ji, H., Ding, S., & Wei, Y. (2020). Data-driven safety enhancing strategies for risk networks in construction engineering. *Reliability Engineering & System Safety*, *197*, Article 106803. <https://doi.org/10.1016/j.res.2020.106803>
- Chen, M. C., Chen, F. L., & Zhang, H. (2011). Multi-dimensional grey construction deformation prediction model research. *Advanced Materials Research*, *287–290*, 3116–3119. <https://doi.org/10.4028/www.scientific.net/AMR.287-290.3116>
- Cheng, M., Li, J., Liu, Y., & Liu, B. (2020). Forecasting clean energy consumption in China by 2025: Using improved grey model GM(1, N). *Sustainability*, *12*(2), Article 698. <https://doi.org/10.3390/su12020698>
- Cheng, M., Liu, Y., & Li, J. (2023). A new modeling method of gray GM(1, N) model and its application to predicting China's clean energy consumption. *Communications in Statistics – Simulation and Computation*, *52*(8), 3712–3723. <https://doi.org/10.1080/03610918.2021.1944641>
- Deng, J. L. (1989). Introduction to grey system theory. *The Journal of Grey System*, *1*(1), 1–24.
- Du, X., Wu, D., & Yan, Y. (2023). Prediction of electricity consumption based on GM(1,Nr) model in Jiangsu province, China. *Energy*, *262*, Article 125439. <https://doi.org/10.1016/j.energy.2022.125439>
- Elmessery, W. M., Habib, A., Shams, M. Y., El-Hafeez, T. A., El-Messery, T. M., Elsayed, S., Fodah, A. E. M., Abdelwahab, T. A. M., Ali, K. A. M., Osman, Y. K. O. T., Abdelshafie, M. F., Abd El-wahhab G. G., & Elwakeel, A. E. (2024). Deep regression analysis for enhanced thermal control in photovoltaic energy systems. *Scientific Reports*, *14*(1), Article 30600. <https://doi.org/10.1038/s41598-024-81101-x>
- Farghaly, H. M., Ali, A. A., & Abd El-Hafeez, T. (2020a). Building an effective and accurate associative classifier based on support vector machine. *Sylwan*, *164*(3), 39–56.
- Farghaly, H. M., Ali, A. A., & El-Hafeez, T. A. (2020b). Developing an efficient method for automatic threshold detection based on hybrid feature selection approach. In R. Silhavy (Ed.), *Advances in intelligent systems and computing: Vol. 1225. Artificial intelligence and bioinspired computational methods* (pp. 56–72). Springer, Cham. [https://doi.org/10.1007/978-3-030-51971-1\\_5](https://doi.org/10.1007/978-3-030-51971-1_5)
- Fatemeh, M., & Vedat, T. (2023). A data-driven recommendation system for construction safety risk assessment. *Journal*



- of *Construction Engineering and Management*, 149(12), Article 04023157. <https://doi.org/10.1061/JCEMD4.COENG-13437>
- Fonseca, E. D., Lima, F. P. A., & Duarte, F. (2014). From construction site to design: The different accident prevention levels in the building industry. *Safety Science*, 70, 406–418. <https://doi.org/10.1016/j.ssci.2014.07.006>
- Gregersen, N. P., Nyberg, A., & Berg, H.-Y. (2003). Accident involvement among learner drivers—An analysis of the consequences of supervised practice. *Accident Analysis & Prevention*, 35(5), 725–730. [https://doi.org/10.1016/S0001-4575\(02\)00051-9](https://doi.org/10.1016/S0001-4575(02)00051-9)
- Hsiao, S.-W., & Liu, M. C. (2002). A morphing method for shape generation and image prediction in product design. *Design Studies*, 23(6), 533–556. [https://doi.org/10.1016/S0142-694X\(01\)00028-X](https://doi.org/10.1016/S0142-694X(01)00028-X)
- Hsu, L.-C. (2003). Applying the Grey prediction model to the global integrated circuit industry. *Technological Forecasting and Social Change*, 70(6), 563–574. [https://doi.org/10.1016/S0040-1625\(02\)00195-6](https://doi.org/10.1016/S0040-1625(02)00195-6)
- Hsu, C., & Wen, Y. (1998). Improved grey prediction models for the trans-pacific air passenger market. *Transportation Planning and Technology*, 22(2), 87–107. <https://doi.org/10.1080/03081069808717622>
- Ikpe, E., Hammon, F., & Oloke, D. (2012). Cost-benefit analysis for accident prevention in construction projects. *Journal of Construction Engineering and Management*, 138(8), 991–998. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0000496](https://doi.org/10.1061/(ASCE)CO.1943-7862.0000496)
- Jianhong, S., Shupeng, L., & Jing, Z. (2024). Using text mining and Bayesian network to identify key risk factors for safety accidents in metro construction. *Journal of Construction Engineering and Management*, 150(6), Article 04024052. <https://doi.org/10.1061/JCEMD4.COENG-14114>
- Lao, T., Chen, X., & Zhu, J. (2021). The optimized multivariate Grey prediction model based on dynamic background value and its application. *Complexity*, 2021, Article 6663773. <https://doi.org/10.1155/2021/6663773>
- Lei, M., & Wang, G. (2022). Analysis of regional financial risk identification and prediction under CVM-GM(1, N) algorithm. In *Proceedings of the 2022 International Conference on Artificial Intelligence and Smart Systems* (pp. 213–219). Atlantis Press. <https://doi.org/10.2991/aebmr.k.220502.039>
- Li, Y., & Li, M. (2015). Prediction research of death number in construction accident based on unbiased Grey-Fuzzy-Markov chain method. In *Proceedings of the 3rd International Conference on Advances in Energy and Environmental Science* (pp. 560–566). Atlantis Press. <https://doi.org/10.2991/ic3me-15.2015.111>
- Li, Q., & Zhang, X. (2024). Neural multivariate grey model and its applications. *Applied Sciences*, 14(3), Article 1219. <https://doi.org/10.3390/app14031219>
- Li, X., Li, N., Ding, S., Cao, Y., & Li, Y. (2023). A novel data-driven seasonal multivariable Grey model for seasonal time series forecasting. *Information Sciences*, 642, Article 119165. <https://doi.org/10.1016/j.ins.2023.119165>
- Lin, C.-T., & Yang, S.-Y. (2003). Forecast of the output value of Taiwan's opto-electronics industry using the Grey forecasting model. *Technological Forecasting and Social Change*, 70(2), 177–186. [https://doi.org/10.1016/S0040-1625\(01\)00191-3](https://doi.org/10.1016/S0040-1625(01)00191-3)
- Mao, M., & Chirwa, E. C. (2006). Application of Grey model GM(1,1) to vehicle fatality risk estimation. *Technological Forecasting and Social Change*, 73(5), 588–605. <https://doi.org/10.1016/j.techfore.2004.08.004>
- Ministry of Housing and Urban-Rural Development of the People's Republic of China. (2022). *Announcement on the production safety accidents of housing and municipal engineering in 2020*. <https://m.lubanlebiao.com/newsinfo/60645.html>
- Nini, X., Qiuhaio, X., G. M. A., Gui, Y., & Jingfeng, Y. (2020). Antecedents of safety behavior in construction: A literature review and an integrated conceptual framework. *Accident Analysis and Prevention*, 148, Article 105834. <https://doi.org/10.1016/j.aap.2020.105834>
- Penghui, L., Limao, Z., & T. R. L. K. (2023). Multi-objective robust optimization for enhanced safety in large-diameter tunnel construction with interactive and explainable AI. *Reliability Engineering and System Safety*, 234, Article 109138. <https://doi.org/10.1016/j.ress.2023.109172>
- Pinto, A., Nunes, I. L., & Ribeiro, R. A. (2011). Occupational risk assessment in construction industry – Overview and reflection. *Safety Science*, 49(5), 616–624. <https://doi.org/10.1016/j.ssci.2011.01.003>
- Shanshan, Z., & Hazem, E. (2022). A hybrid Grey system theory-based subcontractor selection model for high-stakes construction projects. *Journal of Construction Engineering and Management*, 148(6), Article 04022055. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0002302](https://doi.org/10.1061/(ASCE)CO.1943-7862.0002302)
- Simard, M., & Marchand, A. (1994). The behaviour of first-line supervisors in accident prevention and effectiveness in occupational safety. *Safety Science*, 17(3), 169–185. [https://doi.org/10.1016/0925-7535\(94\)90010-8](https://doi.org/10.1016/0925-7535(94)90010-8)
- Song, S. (1992). The application of Grey system theory to earthquake prediction in Jiangsu area. *The Journal of Grey System*, 4, 359–367.
- Song, L., He, X., & Li, C. (2011). Longitudinal relationship between economic development and occupational accidents in China. *Accident Analysis & Prevention*, 43(1), 82–86. <https://doi.org/10.1016/j.aap.2010.07.014>
- Sun, L., & Liu, G. (2011). Apply Gray-Markov Model to predict construction disaster death toll. In *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)* (pp. 4969–4972). IEEE. <https://doi.org/10.1109/AIMSEC.2011.6011412>
- Tam, C. M., Zeng, S. X., & Deng, Z. M. (2004). Identifying elements of poor construction safety management in China. *Safety Science*, 42(7), 569–586. <https://doi.org/10.1016/j.ssci.2003.09.001>
- Tien, T.-L. (2012). A research on the grey prediction model GM(1, n). *Applied Mathematics and Computation*, 218(9), 4903–4916. <https://doi.org/10.1016/j.amc.2011.10.055>
- Toole, T. M., & Gambatese, J. (2008). The trajectories of prevention through design in construction. *Journal of Safety Research*, 39(2), 225–230. <https://doi.org/10.1016/j.jsr.2008.02.026>
- Trivedi, H. V., & Singh, J. K. (2005). Application of grey system theory in the development of a runoff prediction model. *Bio-systems Engineering*, 92(4), 521–526. <https://doi.org/10.1016/j.biosystemseng.2005.09.005>
- Wang, H., & Wang, L. (2020). Logistics forecast of Malacca Strait Port using Grey GM (1, N) model. *Journal of Coastal Research*, 103(Supplement 1), 634–638. <https://doi.org/10.2112/SI103-129.1>
- Wang, Q., & Song, X. (2019). Forecasting China's oil consumption: A comparison of novel nonlinear-dynamic grey model (GM), linear GM, nonlinear GM and metabolism GM. *Energy*, 183, 160–171. <https://doi.org/10.1016/j.energy.2019.06.139>
- Wenli, L., Yixiao, S., Chen, L., Chengqian, L., & Zehao, J. (2023). Development of a non-Gaussian copula Bayesian network for safety assessment of metro tunnel maintenance. *Reliability Engineering & System Safety*, 238, Article 109316. <https://doi.org/10.1016/j.ress.2023.109316>
- Wilde, G. J. S. (1982). The theory of risk homeostasis: Implications for safety and health. *Risk Analysis*, 2(4), 209–225. <https://doi.org/10.1111/j.1539-6924.1982.tb01384.x>

- Wu, W.-Y., & Chen, S.-P. (2005). A prediction method using the grey model GMC(1,n) combined with the grey relational analysis: A case study on Internet access population forecast. *Applied Mathematics and Computation*, 169(1), 198–217.  
<https://doi.org/10.1016/j.amc.2004.10.087>
- Xiong, P., Xiao, L., Liu, Y., Yang, Z., Zhou, Y., & Cao, S. (2021). Forecasting carbon emissions using a multi-variable GM (1,N) model based on linear time-varying parameters. *Journal of Intelligent & Fuzzy Systems*, 41(6), 6137–6148.  
<https://doi.org/10.3233/JIFS-202711>
- Ye, J., Li, Y., Meng, F., & Geng, S. (2024). A novel multivariate time-lag discrete grey model based on action time and intensities for predicting the productions in food industry. *Expert Systems with Applications*, 238, Article 121627.  
<https://doi.org/10.1016/j.eswa.2023.121627>
- Yi, J., Kim, Y., Kim, K., & Koo, B. (2012). A suggested color scheme for reducing perception-related accidents on construction work sites. *Accident Analysis & Prevention*, 48, 185–192.  
<https://doi.org/10.1016/j.aap.2011.04.022>
- Zeng, L. (2018). Analysing the high-tech industry with a multivariable grey forecasting model based on fractional order accumulation. *Kybernetes*, 48(6), 1158–1174.  
<https://doi.org/10.1108/K-02-2018-0078>
- Zhou, Z., Goh, Y. M., & Li, Q. (2015). Overview and analysis of safety management studies in the construction industry. *Safety Science*, 72, 337–350. <https://doi.org/10.1016/j.ssci.2014.10.006>
- Zhu, R., Hu, X., Hou, J., & Li, X. (2021). Application of machine learning techniques for predicting the consequences of construction accidents in China. *Process Safety and Environmental Protection*, 145, 293–302.  
<https://doi.org/10.1016/j.psep.2020.08.006>

## APPENDIX

```
% DGM(1,N) Model with 4 Independent Variables
% Purpose: Predictive modeling using cumulative data sequences
% -----
for n1 = 2:11
    for n2 = n1+1:11
        for n3 = n2+1:11
            for n4 = n3+1:11

                % --- Section 1: Data Initialization ---
                disp('----- Dividing Line -----');

                % Extract and preprocess input data (x0: original matrix)
                x1 = x0(1,:); x2 = x0(2,:); x3 = x0(3,:); x4 = x0(4,:);
                x5 = x0(5,:); x6 = x0(6,:); x7 = x0(7,:); x8 = x0(8,:);
                x9 = x0(9,:); x10 = x0(10,:); x11 = x0(11,:);

                q = length(x1);
                xf = []; xfd = xf';
                xf2 = []; xf3 = []; xf4 = []; xf5 = []; xf6 = [];
                xf7 = []; xf8 = []; xf9 = []; xf10 = []; xf11 = [];

                % --- Section 2: Cumulative Sum Calculation ---
                % Cumulative sum for each variable
                xf2 = [xf2, x2]; xf3 = [xf3, x3]; xf4 = [xf4, x4];
                xf5 = [xf5, x5]; xf6 = [xf6, x6]; xf7 = [xf7, x7];
                xf8 = [xf8, x8]; xf9 = [xf9, x9]; xf10 = [xf10, x10];
                xf11 = [xf11, x11];

                % Append zeros for matrix alignment
                xf2 = [xf2, xfd(1)]; xf3 = [xf3, xfd(2)];
                xf4 = [xf4, xfd(3)]; xf5 = [xf5, xfd(4)];
                % ... (similar lines for xf6 to xf11)

                xf21 = cumsum(xf2); xf31 = cumsum(xf3);
                xf41 = cumsum(xf4); xf51 = cumsum(xf5);
                % ... (similar lines for xf61 to xf111)

                % --- Section 3: Matrix Construction for Least Squares ---
                n = length(x1);
                m = length(xf(1,:));

                % Initialize zero vectors
                x00 = zeros(1, n-1);
                x01 = zeros(1, n+m);

                % Construct matrices F and Ff
                F = [x00', cumsum(x2)(2:n)', cumsum(x3)(2:n)', ...
                    cumsum(x4)(2:n)', cumsum(x5)(2:n)'];
                Ff = [x01', xf21(1:n+m)', xf31(1:n+m)', ...
                    xf41(1:n+m)', xf51(1:n+m)'];

                % --- Section 4: Parameter Estimation ---
                B = [-cumsum(x1)(1:n-1)', F(:,n1), F(:,n2), ...
                    F(:,n3), F(:,n4), ones(n-1,1)];
                Y = cumsum(x1)(2:n)';
                P = inv(B*B) * B'*Y; % Least squares solution

                % Extract coefficients
                b1 = P(1); b2 = P(2); b3 = P(3);
                b4 = P(4); b5 = P(5); b6 = P(6);
```

```

% --- Section 5: Model Simulation ---
% Simulate values (y1) and compute residuals
y1(1) = x1(1);
for k = 1:n-1
    y1(k+1) = -b1*y1(k) + b2*F(k,n1) + b3*F(k,n2) + ...
              b4*F(k,n3) + b5*F(k,n4) + b6;
end

% --- Section 6: Forecasting ---
% Predict future values (Fore0)
f1 = [x11(1)];
for k = 2:m+n
    f1(k) = -b1*f1(k-1) + b2*Ff(k,n1) + b3*Ff(k,n2) + ...
            b4*Ff(k,n3) + b5*Ff(k,n4) + b6;
end

% --- Section 7: Output Results ---
% Display selected variable combination
disp(['(1) Independent Variables: (', num2str(n1), ',', ...
      num2str(n2), ',', num2str(n3), ',', num2str(n4), ')']);

% Display model parameters
disp('(2) Model Parameters [b1, b2, ..., b6]:');
disp(P);

% Calculate and display errors
A = zeros(length(x1), 5);
mp = 0; % Mean percentage error
for k = 1:length(x1)
    A(k,1) = k;
    A(k,2) = x1(k);
    A(k,3) = y0(k);
    A(k,4) = A(k,3) - A(k,2);
    A(k,5) = 100 * abs(A(k,4)) / A(k,2);
    mp = mp + A(k,5);
end
mp = mp / (length(x1)-1);

disp('(3) Error Test Table:');
disp(' No. Actual Simulated Residual Error(%);');
disp(A);

disp('(4) Mean Relative Error (%):');
disp(mp);

disp('(5) Next-Year Forecast:');
disp(Fore0(n+m));

% Export results to Excel
t = table([n1,n2,n3,n4], Fore0_1, Fore0_2, mp, ...
          'VariableNames', {'Independent_Variables', ...
                            'Fitted_Value', 'Forecast_data', 'Average_Error'});
writetable(t, 'Four_Variable_Case.xlsx', ...
           'WriteMode', 'append');
end
end
end
end
end

```