Erzincan Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
2023, 16(3)- 782-799
ISSN: 1307-9085, e-ISSN: 2149-4584
Araştırma Makalesi

Erzincan University
Journal of Science and Technology
2023, 16(3)- 782-799
DOI: 10.18185/erzifbed.1294815
Research Article

Analyzing of Total Number of Railway Accidents in Türkiye via Different Computational Models

Ziya Cakici[1*], Ali Mortazavi[1], Oruc Altintasi[2]

[1]Izmir Democracy University, Department of Civil Engineering, 35140, Izmir, Türkiye
[2]Izmir Katip Celebi University, Department of Civil Engineering, 35640, Izmir, Türkiye

Received: 09/05/2023, **Revised**: 17/07/2023, **Accepted**: 04/08/2023**, Published**: 31/12/2023

Abstract

Accurate prediction of transport-related accidents is considered an important step in assessing the magnitude of the transport-related problems and accelerating decision-making to mitigate them. Therefore, such studies are of great importance for decision makers. In this study, it is aimed to accurately determine (estimate) the annual total number of railway accidents in Türkiye, considering the track length, train-km and Gross National Product (GNP) variables obtained from Türkiye Statistical Institute. In this context, firstly, four different computational models, three of which are optimization-based (one linear, the others nonlinear) and one based on Artificial Neural Network (ANN), are created. Subsequently, the goal was to minimize the Mean Square Error (MSE) between the observed and modeled data for each computational model developed. In the optimization-based models, the selection of the most suitable internal weighting coefficients was accomplished by utilizing the Differential Evolution Algorithm. Finally, within the scope of the study, all statistical results (mean square error, coefficient of determination) obtained for four different calculation models are compared with each other. Consequently, the analysis of the total number of railway accidents in Türkiye reveals that the quadratic model yields more realistic results compared to the other models.

Keywords: railway accident, differential evolution algorithm, artificial neural networks, linear model, non-linear model

Türkiye’ deki Toplam Demiryolu Kaza Sayılarının Farklı Hesaplama Modelleri ile Analizi

Öz

Ulaştırma ile ilgili kazaların doğru tahmini, ulaştırma kaynaklı sorunların büyüklüğünü değerlendirmede ve hafifletmeye yönelik karar vermeyi hızlandırmada önemli bir adım olarak kabul edilmektedir. Bu nedenle, bu tür çalışmalar karar vericiler tarafından büyük bir önem teşkil etmektedir. Bu çalışmada, Türkiye İstatistik Kurumu’ ndan temin edilen demiryolu hat uzunluğu, tren-km ve Gayri Safi Milli Hasıla (GSMH) değişkenleri göz önünde bulundurularak Türkiye’ deki yıllık toplam demiryolu kaza sayısının doğru bir şekilde belirlenmesi (tahmin edilmesi) amaçlanmıştır. Bu bağlamda, öncelikli olarak, üçü optimizasyon tabanlı (birisi lineer, diğerleri non-lineer) birisi de yapay sinir ağı tabanlı olmak üzere dört farklı hesap modeli oluşturulmuştur. Daha sonra, oluşturulan hesaplama modellerinin her biri için, gözlemlenen ve modellenen veriler arasındaki Ortalama Karesel Hata (OKH) minimize edilmeye çalışılmıştır. Optimizasyon tabanlı modellerde, en uygun koşulu ifade eden dâhili ağırlık katsayıları Diferansiyel Gelişim Algoritması kullanılarak belirlenmiştir. Son olarak da, dört farklı hesaplama modeli için elde edilen tüm istatistiksel sonuçlar (ortalama karesel hata, belirleme katsayısı) birbirleriyle karşılaştırılmıştır. Sonuç olarak, Türkiye’ deki demiryolu kazalarının toplam sayısının analizinde, karesel model ile diğer modellere kıyasla daha gerçekçi sonuçlar elde edilebileceği görülmüştür.

Anahtar Kelimeler: demiryolu kazası, diferansiyel gelişim algoritması, yapay sinir ağları, doğrusal model, doğrusal olmayan modeller

*Corresponding Author: ziya.cakici@idu.edu.tr
Ziya CAKICI, https://orcid.org/0000-0001-7003-815X
Ali MORTAZAVI, https://orcid.org/0000-0002-6089-7046
Oruc ALTINTASI, https://orcid.org/0000-0002-4217-1890

1. Introduction

Railway systems can be considered as one of the safest means of transportation and play a vital role in the development of a country [1]. While the railway system generally ensures safer travel, accidents can still occur due to train derailments, fires, or collisions, posing threats to transportation safety and resulting in loss of life [2, 3]. Kyriakidis et al. reported accident-causing factors as infrastructure, environment, human operators, and management [4]. San Kim and Yoon evaluated the reasons for railway accidents in two dimensions as “system” and “human,” in which the latter constituted 68.4% of the total accidents, while “system” accounted for 27.7% [5]. However, Gibson et al. reported that human factors accounted for 80% of the major railway accidents in London [6]. Therefore, Ghofrani et al. focused on establishing cost-effective risk management strategies that required understanding the root causes of railway accidents over historical data analysis, drawing insights, formulating accident prevention strategies, and ultimately ensuring the safety of railway operations. [7].

Numerous studies have concentrated on predicting railway accidents and identifying their causes. Traditional models, such as the Multinomial logit model, have often been employed to establish associations between exploratory variables and accident frequency. [8-10], ordered regression model [11], spatial regression model [12], linear regression model [12], negative binomial regression model [13]. Akalın explored the factors affecting the tram accident severity via Multinomial Logit Model (MNL) for the cities of Eskisehir (in Türkiye), Blackpool, London, Manchester, Nottingham and Sheffield (in England). The model results were almost the same for all cities; rail gauge width and dividedness of the roads were found to be significant parameters [10]. Iranitalab and Khattak investigated the factors affecting highway-rail grade crossing accidents; the location where the vehicular speeds high significantly impact the accident frequency [9]. However, Liu and Khattak stated that gate violation was the main reason for highway-rail grade crossing using the 10-year crash data for the USA [12]. Liu et al. proposed a linear regression model to predict freight train derailments in which the traffic volume and the weather conditions were found to be crucial parameters for the USA [13].

Recently, machine learning (ML) algorithms have been preferred for prediction purposes due to their significant strength over traditional ones, especially when the size of accident data is extensive. Hence ML algorithms can be used to determine the hidden relationship between the accident-causing factors and accident frequency [1, 9, 14-17]. Iranitalab and Khattak compared the strength of MNL, k-Nearest Neighbor (kNN), Support Vector Machine (SVM) and Random Forest (RF) for predicting railway accidents. The results indicated that RF produced more reliable estimation results than the others [15]. Bridgelall and Tolliver explored the factors associated with derailment accidents in the USA. Two years of crash data were used, including over 8000 records. Eleven ML techniques were utilized; the extreme gradient boosting method outperformed, producing 89% prediction accuracy. Excess speed and signalization parameters were found to be significant affecting the derailment accidents [18]. Similarly, Meng et al. used a historical dataset taken from Federal Railroad Administration (FRA) in USA. They investigated the prediction power of Artificial Neural Network (ANN), XGBoost, GBDT, Stacking and AdaBoost methods for predicting railway accidents [1].

In contrast to Bridgelall and Tolliver, the authors identified the significant factors specific to each accident type, and they found that the AdaBoost-Bagging method yielded lower prediction errors across all cases. Li et al., on the other hand, employed SVM for crash severity analysis and examined the superior performance of SVM in comparison to the ordered probit model. Data regarding crashes were gathered from 326 locations in the state of Florida, USA. The results of the sensitivity analysis revealed that SVM exhibited lower prediction errors when compared to the ordered probit model. [19]. Wujie et al. used the Bayesian method to predict railway accidents using 8440 samples from 2017 to 2018 in China. The results showed that the seasons, location, and human factors were found to be significant parameters predicting accidents [17].

Different from the studies mentioned above, a comprehensive study was conducted by Evans [20], investigating fatal train accidents and trends in Europe during 1990-2019. The author considered collisions and derailment accidents and associated them with train kilometers (train-km). Fatal train accident rates were evaluated per year and per train-km as well. Instead of more complicated ML models, the curve fitting process was employed to examine the trend between fatal train accidents and train-km. The descriptive evaluation results indicated that signal passed at danger and overspeeding were the most influential parameters affecting fatal railway accidents.

Accurate accident prediction models provide transportation planners and engineers with ideas for determining new policies, plans and strategies about safety and taking the necessary measures. However, accurate prediction of accidents is considered an important step in assessing the magnitude of the problems and accelerating decision-making towards mitigation. Therefore, in this study, it is aimed to determine in advance the dimensions that the problems related to railway safety in Türkiye can reach. Existing literature demonstrates that railway accident prediction and the underlying causes have been extensively explored through traditional and machine learning-based methods, primarily focusing on specific accident types. However, only a limited number of studies have investigated the macro-level factors contributing to accidents. Total track length, train-km, and gross national product per capita are also very crucial parameters affecting the railway accidents which were not handled together in the existing literature. Based on these facts, the current work proposes a novel model to predict railway accidents by considering these parameters nationwide. As a case study, the accident data and other parameters were taken from Turkish Standard Institute for the years of 2004-2021. For this aim, the Differential Evolution (DE) algorithm as a population-based method is applied to solve the corresponding optimization model. This method does not demand any gradient information of the objective function of optimization problem. This task makes this method proper alternative for solving complex engineering problems on which defining a continuous objective function is difficult or impractical. To verify and assess the acquired explicit formulations, an extra model applying Artificial Neural Network (ANN) is, also, developed. The results are announced and interpreted using comparative tables and diagrams.

The rest of this work is arranged as follows. In the next section, the DE method is described in detail. The proposed model and methodology is described in Section 3. The developed models and their specification are given and compared in Section 4. In Section 5, a brief conclusion is given for this work.

2. Differential Evolution Algorithm

Differential Evolution (DE), one of the population-based meta-heuristic optimization algorithms, was introduced by R. M. Storn and K. Price in 1995. DE can give effective results especially in optimization problems that have continuous variables [21]. Since DE is simple, fast, and easy, it has been used to solve many engineering problems for nearly 30 years [22]. The operation steps of DE algorithm are presented in Figure 1 [23].

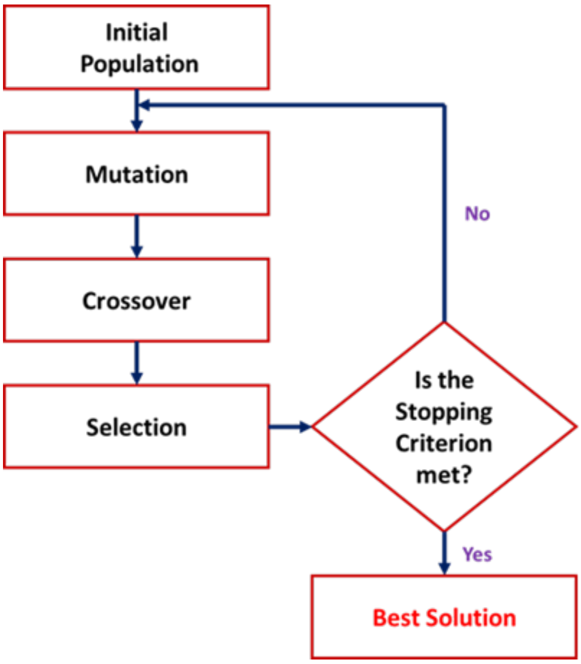

Figure 1. The operation steps of DE algorithm

As seen in Figure 1, firstly, initial population for the algorithm is created. Then, the initial population is improved by applying mutation, crossover, and selection operators, respectively, until the stopping criterion is met (throughout the iterations). Thus, the optimal or near optimal solution of the problem can be obtained. The procedure of DE is composed of eight parts. These parts can be explained as follows:

Determination of the control parameters: In DE, population size (p) can be fixed or variable. Population size must be greater than or equal to 4. A value between 0 and 2 is recommended for the mutation-scaling factor (F) in the literature. Besides, crossover rate (CR) changes between 0.5 and 1, in generally [24]. The maximum number of generations (G_{max}) is determined by the users and can be changed according to the difficulty of the problem.

Creation of the initial population: In DE, the initial population consists of solution vectors (in the number of p). Each of these solution vectors is called “chromosome”. In addition, each solution vector consists of “genes”, that is, “decision variables”.

After the initial population is created, mutation, crossover and selection operators are applied G_{max} times. The best solution in the last generation is the solution to the problem.

Mutation: This operator improves the performance of the algorithm and strengthens the algorithm. With the mutation operator, random changes are made on the genes of the current chromosome. Thus, appropriate increases in current vectors can be achieved at the right times [25, 26]. For mutation, three different chromosomes are selected apart from the current chromosome. In mutation process, firstly, the selected second chromosome is taken out from the selected first chromosome. In the second step, obtained difference chromosome is multiplied by the mutation-scaling factor. In the third step, the scaled difference chromosome is summed up with the selected third chromosome. Finally, a new solution vector to be used for crossover is obtained.

Crossover: The purpose of crossover is to provide the solution of the problem by creating new solution vectors. At this stage, a trial vector for the generation G+1 is created by using current solution vector for the generation G. For the trial chromosome, each gene is selected from a new solution vector that is formed by the mutation in the ratio of CR. In addition, these genes are taken from the current solution vector in the ratio (1-CR).

Evaluation: In this part of the DE, the fitness value for the new solution vector is determined. At this stage, all variables belonging to the trial chromosome are placed in the relevant places in the optimization problem. Then, the objective function value is calculated.

Selection: In the selection part, the fitness value of the trial chromosome is compared with the fitness value of the current chromosome. In case of the fitness of the current chromosome is better, current chromosome continues in the population for at least one more generation. Otherwise, trial chromosome is passed on to the next generation as the new member of the population [27].

Stopping the algorithm: In DE, stopping the algorithm can be achieved in two different ways. When the number of iterations (G) reaches the maximum number of iterations (G_{max}), the algorithm can be terminated. Also, when the difference between the best and worst fitness value is quite low (10-5, 10-6 etc.), the algorithm can be terminated.

Finalization of the problem: After the algorithm is terminated, the vector that has best fitness is called as final solution vector. The numerical values for the genes in the final solution vector indicate the optimum values of the variables for the current optimization problem.

3. Methodology

As it is known, the total track length and the train-km are important parameters for evaluating railway operations. In addition, gross national product (GNP) per capita is one of the most important development indicators for countries [28]. GNP is the total value of all the goods and services produced by a country in a year including income from foreign investments, divided by the number of populations. As the gross national product per capita increases, the welfare level of countries also increases. Thus, higher quality and safer engineering investments (highways and railways, bridges, public transport facilities and etc.) can be made by decision-makers/governments. This helps to reduce transportation-related accidents [29]. When the literature is investigated carefully, it is seen that the GNP per capita is used for accident modelling in many studies [30, 31]. According to this, it can be said that total track length, train-km and gross national product per capita have negative or positive effects on railway accidents. Therefore, in this study, the total number of railway accidents in Türkiye has been tried to be modeled considering these parameters. Data containing the indicators for the years of 2004-2021 are obtained from the Turkish Statistical Institute (TÜİK) website and presented in Table 1 [32].

Table 1. Data used for modeling of the total number of railway accidents in Türkiye

Years	Total Track Length (km) (10^3)	Train-km (10^6)	Gross National Product Per Capita ($) ($10^3$)	Total Number of Railway Accidents (10^2)
2004	10.968	45.873	6.102	5.55
2005	10.973	45.395	7.456	5.22
2006	10.984	44.206	8.102	4.55
2007	10.991	43.102	9.792	3.94
2008	11.005	42.760	10.941	3.86
2009	11.405	41.788	9.103	2.99
2010	11.940	39.025	10.743	1.94
2011	12.000	44.559	11.421	1.77
2012	12.008	40.635	11.796	1.47
2013	12.097	33.755	12.615	0.89
2014	12.485	47.585	12.158	0.93
2015	12.532	46.761	11.006	1.01
2016	12.532	48.015	10.895	1.20
2017	12.608	49.190	10.590	0.53
2018	12.740	53.864	9.453	0.71
2019	12.803	57.705	9.127	0.83
2020	12.803	45.518	8.538	0.66
2021	13.022	44.181	9.587	0.73

* Train-km: Unit measure of transport service representing the movement of a train over one kilometer

Three different forms of mathematical models, of which one of these is linear and the others are non-linear (exponential and quadratic), were used to model the total number of accidents. These mathematical forms can be represented as follows:

- Linear form:
  $Y_l = w_1X_1 + w_2X_2 + w_3X_3 + w_4$
- Exponential form:
  $Y_e = w_1 + w_2{X_1}^{w_3} + w_4{X_2}^{w_5} + w_6{X_3}^{w_7}$
- Quadratic form:
  $Y_q = w_1X_1 + w_2X_2 + w_3X_3 + w_4X_1X_2 + w_5X_1X_3 + w_6X_2X_3 + w_7{X_1}^2 + w_8{X_2}^2 + w_9{X_3}^2 + w_{10}$

Where Y_l, Y_e and Y_q are the total number of railway accidents (modeled) in the forms of linear, exponential, and quadratic, respectively. X_1 is the total annual track length (km), X_2 is the annual train-km and X_3 is the annual gross national product per capita (\$). w_1, w_2, w_3, …, w_n are the corresponding weighting coefficients.

As seen in Table 1, the variables (total track length, train-km and gross national product per capita and the total number of railway accidents) have different orders of magnitudes. Therefore, the parameters are normalized in the modelling stage as shown in Equation 1-4 [33]. In these equations, min and max represent minimum and maximum values of variables from 2004 to 2021, respectively. It should be noted that in Equation 4, Y represents the observed total number of railway accidents.

$$\frac{X_1 - X_{1_{min}}}{X_{1_{max}} - X_{1_{min}}} = \frac{X_1 - \left(10.968\times10^3\right)}{\left(13.022\times10^3 - 10.968\times10^3\right)} \tag{1}$$

$$\frac{X_2 - X_{2_{min}}}{X_{2_{max}} - X_{2_{min}}} = \frac{X_2 - \left(33.755\times10^6\right)}{\left(57.705\times10^6 - 33.755\times10^6\right)} \tag{2}$$

$$\frac{X_3 - X_{3_{min}}}{X_{3_{max}} - X_{3_{min}}} = \frac{X_3 - \left(6.102\times10^3\right)}{\left(12.615\times10^3 - 6.102\times10^3\right)} \tag{3}$$

$$\frac{Y - Y_{min}}{Y_{maks} - Y_{min}} = \frac{Y - \left(0.53\times10^2\right)}{\left(5.55\times10^2 - 0.53\times10^2\right)} \tag{4}$$

The accurate modelling of the total number of railway accidents can be achieved by determining of most appropriate (optimum) weighting coefficients. This means that the difference between the observed and modelled total number of railway accidents must be minimized. Therefore, an algorithm is applied to curve fitting problem to determine optimum weighting coefficients for the considered mathematical forms.

In all mathematical forms, the algorithm firstly selects the weighting coefficients in randomly. Then, it is run to determine optimum weighting coefficients throughout the predetermined number of iterations. In this study, the DE algorithm was used for this purpose.

Since the curve fitting problems can be defined as minimizing the errors between observed and modeled (predicted) values, Mean Squared Error (MSE) was used as the objective function in optimization process. MSE is the average squared difference between the observed and predicted values. It can be considered as a risk function (cost function) for optimization problems [33, 34]. Therefore, MSE can be minimized to determine optimum weighting coefficients. Objective function for three mathematical forms (linear, exponential, and quadratic) is given in Equation 5.

$$min\, F\left(x\right)=\frac{1}{m}\sum_{i=1}^{m}\left(Y_{observed}-Y_{mod\,eled}\right)^2 \tag{5}$$

After minimizing the MSE, obtained results should be evaluated statistically. One of the most used measures for a statistical evaluation is the coefficient of determination (R^2) value. R^2 provides a measure of how well observed results are represented by the model, taking into account the proportion of total variation of results explained by the model [35]. This value normally ranges from 0 to 1. As it gets closer to 1, the reliability of the model increases. R^2 of 1 indicates that the results obtained with the model are exactly the same as the observed results. In short, if R^2 equals to 1, it can be said that the model data fits perfectly with the observation data. In statistics, R^2 can be seen as more informative than Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE). Therefore, R^2 values for the mathematical forms considered in this study are calculated as shown in Equation 6-8.

$$SS_{res}=\sum_{i=1}^{n}\left(Y_i-Y_{i_m}\right)^2 \tag{6}$$

$$SS_{tot}=\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2 \tag{7}$$

$$R^2=1-\frac{SS_{res}}{SS_{tot}} \tag{8}$$

In Equation 6-8, n is the number of total data. Y_i represents the observed value for observation i. Y_{i_m} shows the value (Y_l, Y_e and Y_q) obtained with the model for observation i. $\overline{Y}$ is the mean of observed data. While SS_{res} indicates the sum of squared estimate of errors, SS_{tot} expresses the total sum of squares.

4. Analyzes and Results

In this section, firstly, three distinct explicit mathematical model based on the direct optimization approach is given. Then, an alternative ANN-based black-box model is addressed, and results are compared.

4.1. Explicated Mathematical Model

In this part of the study, the total number of railway accidents in Türkiye has been tried to be modeled by considering three types of mathematical forms. For this purpose, in the first step, three different scripts for all mathematical forms were written in MATLAB. In these scripts, MSE for observed and model data was minimized using the Differential Evolution Algorithm. Control parameters of the algorithm were selected considering the previous studies in the literature and were presented in Table 2 [22, 36, 37].

Table 2. Selected values for the control parameters of the DE

Control Parameters	Selected Value
Population size (p)	50
Mutation-scaling Factor (F)	0.8
Crossover rate (CR)	0.8
Maximum number of generations (G_{max})	10000

In the second step, each script was run ten times. The corresponding weighting coefficients, MSE and R^2 values were obtained for each run. In the third step, the best, worst and mean values of MSE and the best and worst values of R^2 were determined for the mathematical forms discussed. The obtained results were summarized in Table 3.

Table 3. MSE and R^2 values for all mathematical forms

	MSE			**R^2**	
	Best	**Worst**	**Mean**	**Best**	**Worst**
Linear Form	0.0029	0.0029	0.0029	0.9774	0.9774
Exponential Form	0.0015	0.0015	0.0015	0.9891	0.9891
Quadratic Form	**9.5558E-04**	0.0014	0.0012	**0.9929**	0.9892

As seen in Table 3, the best and worst MSE values for the linear form were obtained as 0.0029. The best and worst R^2 value was also determined as 0.9774. For the exponential form, the MSE value decreased by almost half, and R^2 values for both the best and the worst run was determined as 0.9891. For the quadratic form, the best and the worst values of MSE was obtained as approximately 0.0009556 and 0.0014, respectively. Since the quadratic form has many weighting coefficients (ten weighting coefficients), different possible solutions were obtained for each run. The differences between best and worst MSEs resulted from this situation. Similarly, the best and the worst values of R^2 for quadratic form were obtained as 0.9929 and 0.9892, respectively. When the Table 3 were examined in detail, it can be said that quadratic form gives better results (lowest MSE and highest R^2) than the other (linear and exponential) forms. In the next step, weighting coefficients that provides minimum MSE and maximum of R^2 for all mathematical forms were determined. The corresponding weighting coefficients that give the best results were shown in Table 4.

Table 4. The corresponding weighting coefficients that provide best models for all mathematical forms

Linear Form	**Exponential Form**	**Quadratic Form**
w_1 = -0.79794 w_2 = 0.06773 w_3 = -0.37379 w_4 = 0.91172	w_1 = 0.92032 w_2 = -0.80241 w_3 = 0.59827 w_4 = 0.08690 w_5 = 2.37429E-07 w_6 = -0.28772 w_7 = 0.50852	w_1 = -0.09949 w_2 = 0.11767 w_3 = -0.65294 w_4 = -2.29909 w_5 = -0.10646 w_6 = 0.88518 w_7 = 0.37296 w_8 = 1.02775 w_9 = 0.09317 w_{10} = 0.66184

After the best corresponding weighting coefficients were determined, modeled total railway accident numbers were calculated by using normalized total railway accident numbers of data from 2004 to 2021. Comparison of observed and model data (the total number of railway accidents) for all considered mathematical forms is presented graphically in Figure 2.

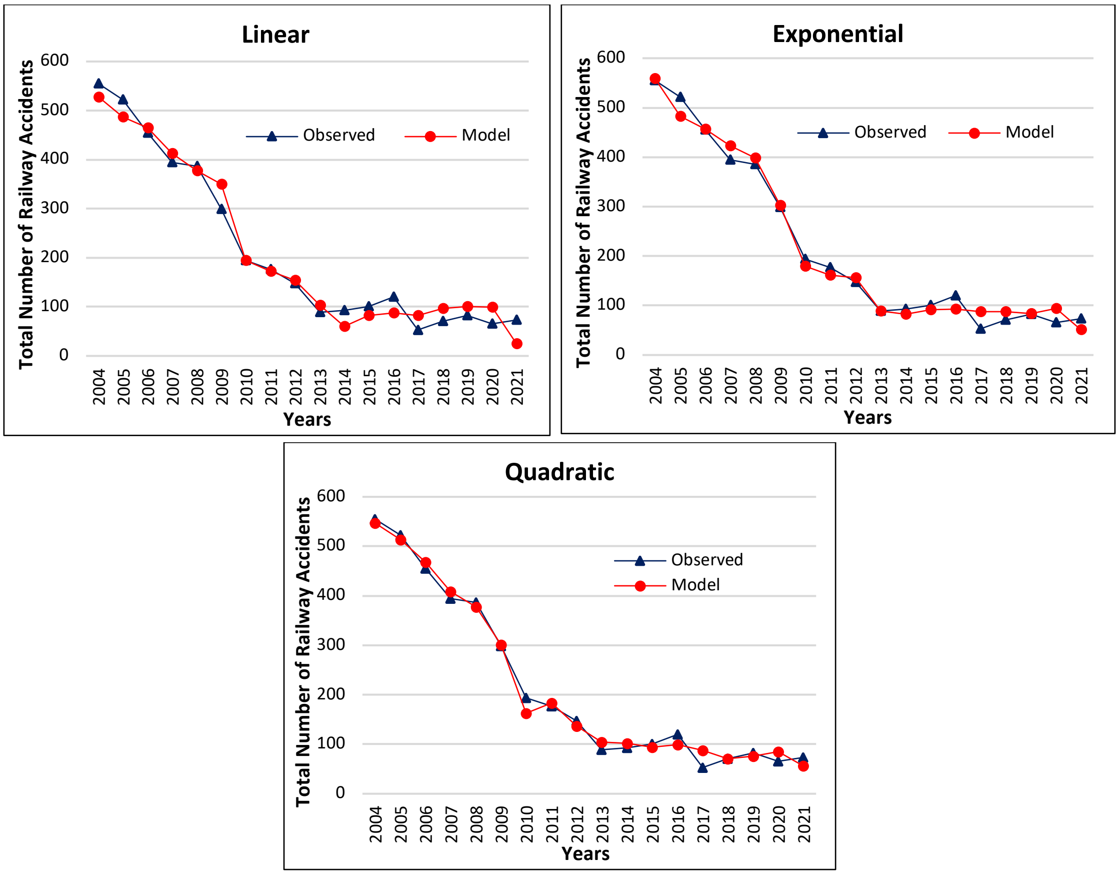

Figure 2. Comparison of observed and modeled data for considered mathematical forms

As can be seen in Figure 2, the differences between observed and modeled data for the linear form are quite large. For the exponential form, the differences are noticeably reduced compared to the linear form. Lowest differences between the observed and modeled data are seen in the quadratic form. This confirms the obtained results presented in Table 3. In the next step, it is aimed to compare the power of all mathematical forms. For this purpose, observed and modeled data are compared (total number of railway accidents) considering the 1:1 line. In a powerful model, the points representing observed and modeled values must be above the 1:1 line. This implies that R^2 is equal to 1. Figure 3 shows the scatterplot of modelled versus observed values for linear, exponential, and quadratic forms.

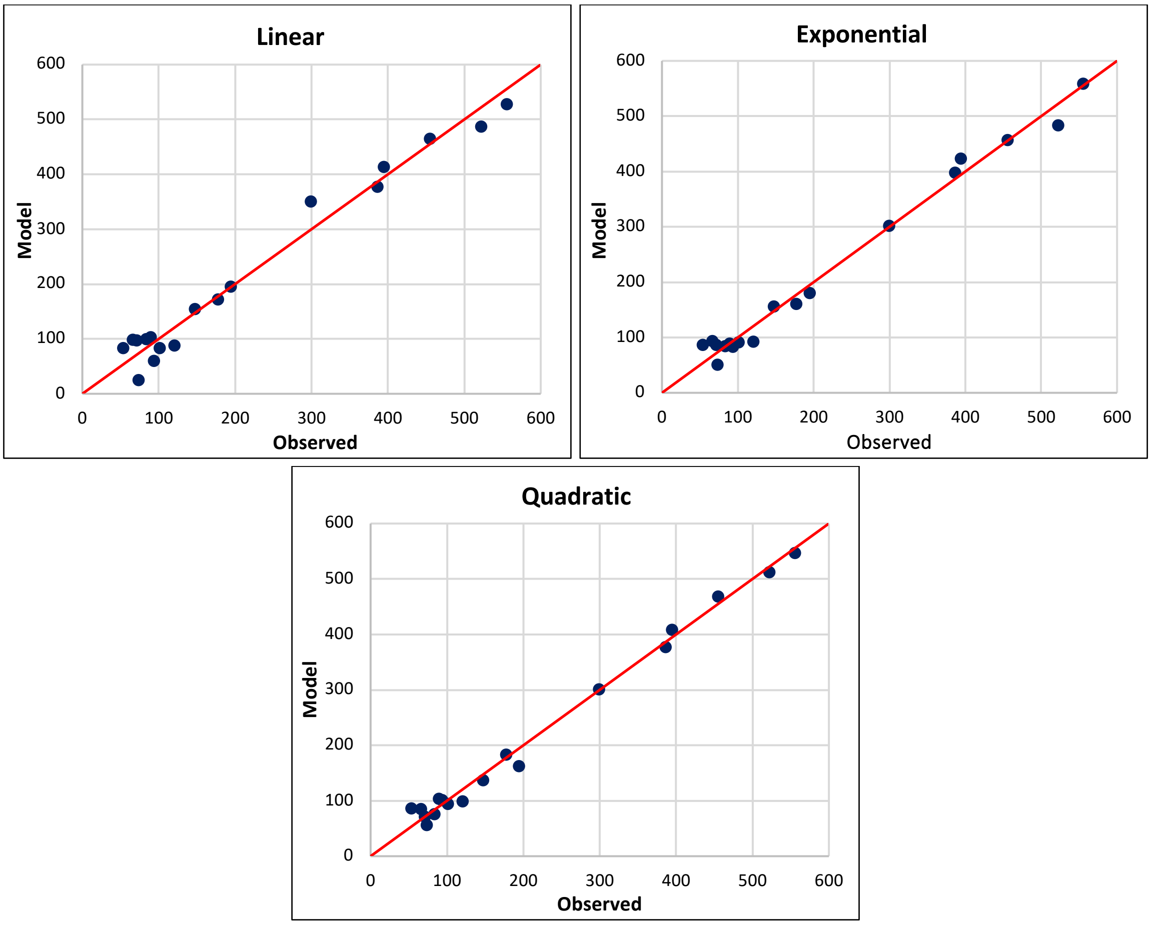

Figure 3. The scatterplot of modelled versus observed values for considered mathematical forms

Figure 3 illustrates that the distribution of blue points for the quadratic form appears more uniform compared to the linear and exponential forms. In the quadratic form, the majority of blue points are concentrated near the red (1:1) line. This observation highlights the effectiveness and strength of the quadratic form.

4.2. The Model Based on the Artificial Neural Network

To evaluate the quality of the models generated in the previous section (e.g., the linear, quadratic, and exponential formulations), an additional model utilizing Artificial Neural Networks (ANN) is developed in this section for comparative analysis. Artificial Neural Networks (ANNs) are computational procedures designed to emulate the capabilities of the human brain, including the ability to learn, derive new information, and create or discover new knowledge without human intervention. An ANN is trained using a dataset to learn patterns and relationships between the input and output variables. During training, the weights and biases of the neurons are adjusted to minimize the error between the predicted and actual outputs. Once trained, the ANN can be used to make predictions on new data [38]. Artificial neural networks emerged because of mathematical modelling of the human brain learning process. It mimics the structure of biological neural networks in the brain and their ability to learn, remember and generalize [39]. In all ANN-based models, a mathematical structure is considered, which of course can be displayed graphically and has a series of parameters and adjustment screws. This general structure is adjusted and optimized by a training algorithm so that it can show proper behavior.

One of the privileged neural systems is the Multi-Layer Perceptron (MLP) model, which simulates the transmission function of the human brain. In this type of neural network, the behavior of the human brain and signal propagation have been considered, and hence, they are called feedforward networks. Based on this information, In the current section, an alternative implicit model based on the Artificial Neural Network (ANN) approach is developed. The inputs (Total Track Length, Train-km, Gross National Product Per Capita) and output (Total Number of Railway Accidents) of this model are the same as those used in the previous section. As depicted in Figure 4, the ANN-based model comprises a single hidden layer consisting of ten perceptrons. After conducting several tests on different configurations of the ANN model, the presented architecture was selected as it provides the minimum level of complexity (i.e., the minimum number of layers) required for effective performance.

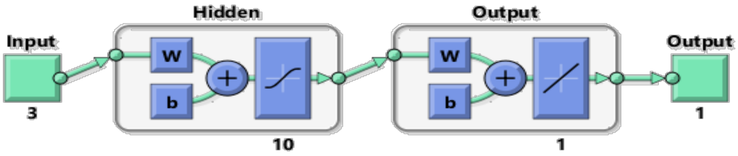

Figure 4. Architecture of ANN-based model

In the created ANN model, while 70% of the data was used for training, 15% of the data was used for validation. The remaining data (15%) was also applied for testing. The correlation coefficients (R) for the training, validation, testing, and entire model can be found in Figure 5.

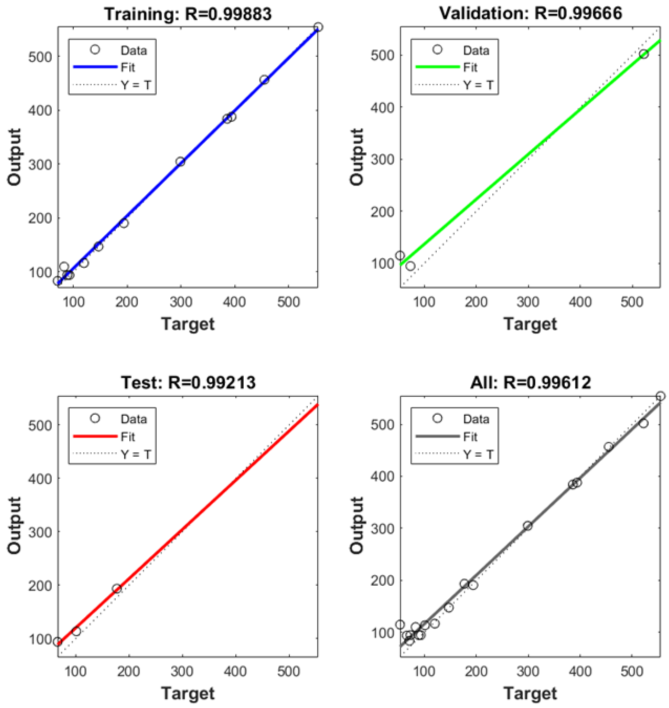

Figure 5. The correlation coefficient (R) for training, validation, test, and entire model

Table 5. The results for ANN-based approximation model and observed values

Years	Total Track Length (km) (10^3)	Train-km (10^6)	Gross National Product Per Capita ($) ($10^3$)	Total Number of Railway Accidents (10^2)	
				Observed values	Model values
2004	10.968	45.873	6.102	5.55	5.72
2005	10.973	45.395	7.456	5.22	5.22
2006	10.984	44.206	8.102	4.55	4.55
2007	10.991	43.102	9.792	3.94	4.09
2008	11.005	42.760	10.941	3.86	3.86
2009	11.405	41.788	9.103	2.99	2.99
2010	11.940	39.025	10.743	1.94	1.94
2011	12.000	44.559	11.421	1.77	2.15
2012	12.008	40.635	11.796	1.47	1.47
2013	12.097	33.755	12.615	0.89	0.89
2014	12.485	47.585	12.158	0.93	0.93
2015	12.532	46.761	11.006	1.01	1.36
2016	12.532	48.015	10.895	1.20	1.20
2017	12.608	49.190	10.590	0.53	0.83
2018	12.740	53.864	9.453	0.71	0.71
2019	12.803	57.705	9.127	0.83	0.83
2020	12.803	45.518	8.538	0.66	1.19
2021	13.022	44.181	9.587	0.73	0.73

* Train-km: Unit measure of transport service representing the movement of a train over one kilometer

Based on the given results, R value is acquired as 0.9961 for the model, it shows the proper performance of the attained model. To compare the performance of the model more precisely, the observed and attained approximate values are given in Table 5. It should be noted that the root-mean-square (RMS) of the total system is achieved as RMS=0.83. In the next section, the attained results from all formulation are statistically compared.

4.3. Comparing to Results

In this section, firstly, the normalized values for the models are remapped to their original state. Then, as shown in Table 6, MSE, RMSE and R^2 values for each model are calculated and compared.

Table 6. Real MSE, RMSE and R^2 values for each model

	Linear Model	**Exponential Model**	**Quadratic Model**	**ANN based Model**
MSE	725.72	371.00	242.89	382.89
RMSE	26.94	19.26	15.58	19.57
R^2	0.974	0.987	0.991	0.986

According to Table 8, the linear model stands in the last place among all other. The Exponential and ANN-based models, providing very close outcomes, stand in the next places. The Quadratic model with the lowest MSE and RMSE (242.89 and 15.58) and highest R^2 (0.991) values is ranked in the first place among all others. In addition, it should be noted that the number of iterations for linear, exponential, and quadratic models is 4250, 5450 and 5950, respectively. Based on the addressed results, the Quadratic model, as an explicit formulation, can be used for assessing/evaluating the total number of accidents in the railway system in Türkiye.

5. Conclusions

In the current study, applying the data presented by the Turkish Statistical Institution, four different models were developed to model the total number of accidents in the railway systems in Türkiye. In the modelling, track length, train-km and Gross National Product are considered as independent variables. While three of the developed models include linear and non-linear mathematical forms, the other is based on Artificial Neural Network. For this reason, the analyses made within the scope of this study were carried out in two stages.

In the initial stage, three distinct mathematical models were developed, one of which is linear while the other two are non-linear (exponential and quadratic). These models were developed to capture the relationship between dependent and independent variables. Due to the varying orders of magnitude among the variables used in the modeling process, a normalization process was employed to bring all variables within a consistent search domain. Afterwards, the most appropriate weighting coefficients for each model were determined by minimizing the mean square errors between the observed data and the model data. In the optimization process, a metaheuristic method, so-called Differential Evolution (DE), is employed. After spotting the most suitable models, the outputs are remapped to their original state. Then, MSE and R^2

values for each mathematical form were calculated. MSE values for the linear, exponential, and quadratic forms were obtained as 725.72, 371.00 and 242.89, respectively. Besides R^2 values for linear, exponential and quadratic forms were obtained as 0.974, 0.987 and 0.991, respectively. According to obtained statistical results, it can be said that the quadratic form is the most powerful model within these three models.

In the second stage, a multilayer ANN-based model was created, and the same independent variables were used for modelling. After denormalization of the outputs, the MSE and R^2 values were calculated. For the ANN-based model, these values were obtained as 382.89 and 0.986, respectively.

It should be noted that, in this study, since there are 18 data from 2004 to 2021 on TÜİK website, only these data were used in the development of the models. The more realistic results can be achieved by increasing the amount of data. In the light of obtained numerical results, it is seen that ANN-based model gives almost similar results with exponential model. Besides, compared to the other models (linear, exponential, and ANN-based models), it is concluded that quadratic model can provide the most realistic outcomes. The limitations of the current works can be outlined as follows.

The study focuses on modeling the number of railway accidents in Türkiye using three independent variables: total track length, train-km, and GNP per capita. The authors suggest that incorporating more independent variables can lead to more realistic outcomes. The study utilizes 18 data points from 2004 to 2021, acknowledging the need for additional data to enhance realism. However, reliable values for variables prior to 2004 are unavailable. The study includes a linear model and two non-linear models, highlighting the potential for different non-linear models to improve realism. Employing powerful computational approaches like deep learning and reinforcement learning, along with increased data, can help achieve highly realistic results. In the future, it is planned to modelling of total number of railway accidents in Türkiye, considering above-mentioned situations.

Ethics in Publishing

There are no ethical issues regarding the publication of this study.

Author Contributions

Z.C.: Conceptualization and methodology, Software, Validation, Writing original draft; A.M.: Conceptualization and methodology, Software, Validation, Writing original draft; O.A.: Conceptualization and methodology, Writing original draft.

References

[1] Meng, H., Tong, X., Zheng, Y., Xie, G., Ji, W., & Hei, X. (2022). Railway accident prediction strategy based on ensemble learning. Accident Analysis & Prevention, 176, 106817.

[2] Liu, J., Schmid, F., Li, K., & Zheng, W. (2021). A knowledge graph-based approach for exploring railway operational accidents. Reliability Engineering & System Safety, 207, 107352.

[3] Hadj-Mabrouk, H. (2020). Analysis and prediction of railway accident risks using machine learning. AIMS Electronics and Electrical Engineering, 4(1), 19-46.

[4] Kyriakidis, M., Majumdar, A., & Ochieng, W. Y. (2015). Data based framework to identify the most significant performance shaping factors in railway operations. Safety science, 78, 60-76.

[5] San Kim, D., & Yoon, W. C. (2013). An accident causation model for the railway industry: Application of the model to 80 rail accident investigation reports from the UK. Safety science, 60, 57-68.

[6] Gibson, W. H., Mills, A. M., Smith, S., & Kirwan, B. K. (2012). Railway action reliability assessment, a railway specific approach to human error quantification. Rail Human Factors. Supporting reliability, safety and cost reduction.

[7] Ghofrani, F., He, Q., Goverde, R. M., & Liu, X. (2018). Recent applications of big data analytics in railway transportation systems: A survey. Transportation Research Part C: Emerging Technologies, 90, 226-246.

[8] Ye, F., & Lord, D. (2014). Comparing three commonly used crash severity models on sample size requirements: Multinomial logit, ordered probit and mixed logit models. Analytic methods in accident research, 1, 72-85.

[9] Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. Accident Analysis & Prevention, 108, 27-36.

[10] Akalın, K. B. (2016). Investigation of factors affecting tram accident severity with multinomial logit model". MSc thesis, Eskisehir, Turkey: Eskisehir Osmangazi University.

[11] Dabbour, E., Easa, S., & Haider, M. (2017). Using fixed-parameter and random-parameter ordered regression models to identify significant factors that affect the severity of drivers' injuries in vehicle-train collisions. Accident Analysis & Prevention, 107, 20-30.

[12] Liu, J., & Khattak, A. J. (2017). Gate-violation behavior at highway-rail grade crossings and the consequences: using geo-spatial modeling integrated with path analysis. Accident Analysis & Prevention, 109, 99-112.

[13] Liu, X., Saat, M. R., & Barkan, C. P. (2017). Freight-train derailment rates for railroad safety and risk analysis. Accident Analysis & Prevention, 98, 1-9.

[14] Baysari, M. T., McIntosh, A. S., & Wilson, J. R. (2008). Understanding the human factors contribution to railway accidents and incidents in Australia. Accident Analysis & Prevention, 40(5), 1750-1757.

[15] Iranitalab, A., & Khattak, A. (2020). Probabilistic classification of hazardous materials release events in train incidents and cargo tank truck crashes. Reliability Engineering & System Safety, 199, 106914.

[16] Mirabadi, A., & Sharifian, S. (2010). Application of association rules in Iranian Railways (RAI) accident data analysis. Safety Science, 48(10), 1427-1435.

[17] Wujie, J., Le, J., & Cheng, Z. (2022). Analyzing and predicting railway operational accidents based on fishbone diagram and Bayesian networks. Tehnički vjesnik, 29(2), 542-552.

[18] Bridgelall, R., & Tolliver, D. D. (2021). Railroad accident analysis using extreme gradient boosting. Accident Analysis & Prevention, 156, 106126.

[19] Li, Z., Liu, P., Wang, W., & Xu, C. (2012). Using support vector machine models for crash injury severity analysis. Accident Analysis & Prevention, 45, 478-486.

[20] Evans, A. W. (2021). Fatal train accidents on Europe's railways: An update to 2019. Accident Analysis & Prevention, 158, 106182.

[21] Tan, E., Sadak, D., & Ayvaz, M. T. (2020). Optimum design of sewer systems by using differential evolution algorithm. Turkish Journal of Civil Engineering, 31(5), 10229–10250.

[22] Kamal, M., & Inel, M. (2019). Optimum design of reinforced concrete continuous foundation using differential evolution algorithm. Arabian Journal for Science and Engineering, 44, 8401-8415.

[23] Sriboonchandr, P., Kriengkorakot, N., & Kriengkorakot, P. (2019). Improved differential evolution algorithm for flexible job shop scheduling problems. Mathematical and Computational Applications, 24(3), 80.

[24] Baskan, O. (2019). A multiobjective bilevel programming model for environmentally friendly traffic signal timings. Advances in Civil Engineering, 2019, 1-13.

[25] Abdelkader, E. M., Al-Sakkaf, A., Alfalah, G., & Elshaboury, N. (2022). Hybrid Differential Evolution-Based Regression Tree Model for Predicting Downstream Dam Hazard Potential. Sustainability, 14(5), 3013.

[26] Yu, X., Jiang, N., Wang, X., & Li, M. (2023). A hybrid algorithm based on grey wolf optimizer and differential evolution for UAV path planning. Expert Systems with Applications, 215, 119327.

[27] Elçi, A., & Ayvaz, M. T. (2014). Differential-evolution algorithm based optimization for the site selection of groundwater production wells with the consideration of the vulnerability concept. Journal of Hydrology, 511, 736-749.

[28] Murat, Y. S., & Ceylan, H. (2006). Use of artificial neural networks for transport energy demand modeling. Energy policy, 34(17), 3165-3172.

[29] Njoh, A. J. (2000). Transportation infrastructure and economic development in sub-Saharan Africa. Public Works Management & Policy, 4(4), 286-296.

[30] Ali, G. A., Al-Alawi, S. M., & Bakheit, C. S. (1998). A comparative analysis and prediction of traffic accident causalities in the Sultanate of Oman using artificial neural networks and statistical methods. Sultan Qaboos University Journal for Science, 3, 11-20.

[31] Shaik, M. E., Islam, M. M., & Hossain, Q. S. (2021). A review on neural network techniques for the prediction of road traffic accident severity. Asian Transport Studies, 7, 100040.

[32] https://www.tuik.gov.tr/

[33] Sonmez, M., Akgüngör, A. P., & Bektaş, S. (2017). Estimating transportation energy demand in Turkey using the artificial bee colony algorithm. Energy, 122, 301-310.

[34] Ceylan, H., Ceylan, H., Haldenbilen, S., & Baskan, O. (2008). Transport energy modeling with meta-heuristic harmony search algorithm, an application to Turkey. Energy policy, 36(7), 2527-2535.

[35] Korkmaz, E., & Akgüngör, A. P. (2018). Flower pollination algorithm approach for the transportation energy demand estimation in Turkey: model development and application. Energy Sources, Part B: Economics, Planning, and Policy, 13(11-12), 429-447.

[36] Baskan, O., & Ceylan, H. (2014). Differential evolution algorithm based solution approaches for solving transportation network design problems. Pamukkale University Journal of Engineering Sciences, 20(9), 324-331.

[37] Baskan, O., Ceylan, H., & Ozan, C. (2020). Investigating acceptable level of travel demand before capacity enhancement for signalized urban road networks. Turkish Journal of Civil Engineering, 31(2), 9897-9917.

[38] Cheng, C. J. (2008). Robust control of a class of neural networks with bounded uncertainties and time-varying delays. Computers & Mathematics with Applications, 56(5), 1245-1254.

[39] Sum, J., & Leung, A. C. S. (2008). Prediction error of a fault tolerant neural network. Neurocomputing, 72(1-3), 653-658.