Earthquake Prediction by Using Time Series Analysis

Sultan LÖK
Department of Software Engineering
Fırat Üniversitesi
Elazığ, Turkey
loksultan31@gmail.com

Murat KARABATAK
Department of Software Engineering
Fırat Üniversitesi
Elazığ, Turkey
mkarabatak@firat.edu.tr

2021 9th International Symposium on Digital Forensics and Security (ISDFS) | 978-1-6654-4481-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISDFS52919.2021.9486358

***Abstract*— Nowadays, with the developing technologies, big data stored has started to be formed. Data mining methods have been developed due to the need to obtain information from these stored data. These methods include clustering, classification, association rule, and time series. In this study, time series analysis, one of the data mining methods, was emphasized. Time series provide predictions about future time by time data. Time series are divided into two: linear and non-linear methods. Linear time series methods estimate by assuming that the series is stationary. Non-linear time series methods predict based on the raw version of the series in the real world. In this study, artificial neural networks (ANNs), one of the non-linear methods of Time Series Analysis, are used. Earthquake data were discussed with ANNs. Estimates were made by analyzing earthquake data.**

Keywords*— *Data Mining, Time Series, Artificial Neural Networks, Earthquake Data

I. INTRODUCTION

Today, the amount of digital data is increasing rapidly. With the increasing data, obtaining valuable information by using these data has come to the fore. Data mining methods have been developed to transform data into information. These methods are specified as classification, clustering, association rules, and time series analysis. In this study, time series analysis, which allows predictions about future data based on-time data, is discussed. There is a large number of series that contain time data in almost every field. Time series are often defined as a series of events or observations described numerically in nature, recorded on a regular or irregular basis of time [1].

Time series data are generated in a wide variety of applications in different fields. For example, daily fluctuations in the stock market trace produced by a computer cluster, medical and biological experimental observations, readings from sensor networks, location updates of moving objects in location-based services, etc. With the proliferation of time series data, new methodologies and technologies were created to manage, analyze, transform and visualize this data. One of these methodologies is artificial neural network (ANN).

Artificial neural networks were developed with the need arising from the fact that Real-world systems are generally not linear. ANNs represent a powerful biological tool that reproduces very complex non-linear dependencies [2]. ANN provides learning from existing data and creating a prediction model. ANN model is among the most widely used prediction models in the data estimation method. There are many examples of these studies in the literature.

Narayanakumar and Raja evaluated the performance of Back Propagation (BP) and Multilayer Perceptron (MLP) neural network techniques in predicting earthquakes. Event time, latitude, longitude, depth, and magnitude data were taken to transform them into inputs for neural networks. As a result of the applications, they showed that the BP neural network method can better predict earthquakes of 3 to 5 magnitude than previous studies but could not provide good results for 5 to 8 magnitude earthquakes due to the lack of sufficient data [3].

Çelik, Atalay, and Bayer worked on earthquake prediction from seismic impacts using ANNs and SVM (Support Vector Machines) with intense seismic activity in the southwest of Turkey. They have trained ANN using four different regions of seismic data. Early detected with a rate of 83% with ANNs. 91% early detected with Support Vector Machines [4].

In this study, earthquake data of Elazığ province were used with ANN method for earthquake prediction. Estimates were made by analyzing the earthquake data of Elazığ province. In line with the predictions, performance adequacy was discussed, and recommendations were made.

II. TIMES SERIES

Time series is the series that observe the movement of a variable over time and display the observation results chronologically [5]. Changes in the amount of precipitation in a year, changes in currency units according to time, instant changes in heart rhythm measurements, the number of product sales of a company between 1965-1970, etc. are some of the time series examples.

Time series are classified according to the number of variables and the method of measurement. Univariate time series is time series containing variable of a single. Multivariate time series is time series containing variables of more than one. A time series is also classified as continuous or discrete. A continuous time series, includes observations measured at always time. A discrete-time series includes observations measured at discrete time [6, 7]. These

observations are recorded and enable the application of time series analysis.

It is stated that the time series is affected by one or more components according to their structure. These components are [8-10];

- Trend: It is a state of stability after a rise or fall in time series observed for a long time.
- Seasonal fluctuation: It is the upward or downward mobility observed in seasonal times in time series [8].
- Cyclical fluctuation: It refers to the upward or downward mobility in the time series, which is not seasonal but occurs in the direction of important factors [9].
- Irregular random movement: It is the unpredictable rise and fall movements that occur irregularly in time series, except for seasonal or cyclical fluctuation. [8-10].

A. The Concepts of Stationary

Time series values approaching a specific value or fluctuating around the expected value is called stationarity. If statistics such as mean, variance, and covariance in time series are constant over time, time series is stationary. In stationary time series observations; Other structures depending on seasonal effects, trends, and time index are not seen [12, 15].

Series with variable mean, variance, and covariance in time series is not stationary. In non-stationary time-series observations; Seasonal effects, trends, and other structures depending on time index are seen. Two methods can analyze non-stationary time series:

1. Methods developed to stabilize non-stationary time series are used. These methods are stabilized the time series. Stationarized time series are analyzed by methods analyzing stationary time series. Among these methods; AR, MA, ARMA, Box-Jenkins etc. has [7].

2. Analysis methods are used that can analyze stationary and non-stationary series. These methods provide the most accurate estimates by analyzing the time series. Among these methods; ANNs, support vector machines etc. has [7].

B. Artificial Neural Network (ANN)

ANN was proposed as an alternative technique to time series estimation and has gained immense popularity in recent years. The primary purpose of ANNs is a model developed to imitate the human brain by machines [2, 10]. Similar to the study of the human brain, ANNs try to recognize patterns in input data, learn from the data, and then obtain predictions from the data with generalized analysis results.

ANNs have distinct features that make them an alternative technique for time series analysis and prediction [11]. These features;

- ANNs are data-driven and self-adaptable [5, 11]. Thanks to this feature of ANN, there is no need to specify a model form or make pre-assumptions about the statistical distribution of the data; the desired model is created in a flexible manner according to the features presented from the data.
- ANNs are not linear, unlike traditional linear approaches, more practical and accurate results are obtained in modeling complex data models [13, 14]. Many cases showing that ANNs perform better analysis and predictions than various linear models have been observed in the literature reviews.
- ANNs are universal functional approaches. They have shown that a network can bring any continuous function closer to any desired precision [13, 14]. ANNs use parallel processing of information from data to approach a large class of functions with high accuracy. In addition, it can handle situations where input data is incorrect, incomplete, or blurry.

1) ANN Architecture

A neural network is software structures that contain multiple artificial neurons arranged in a series of layers.

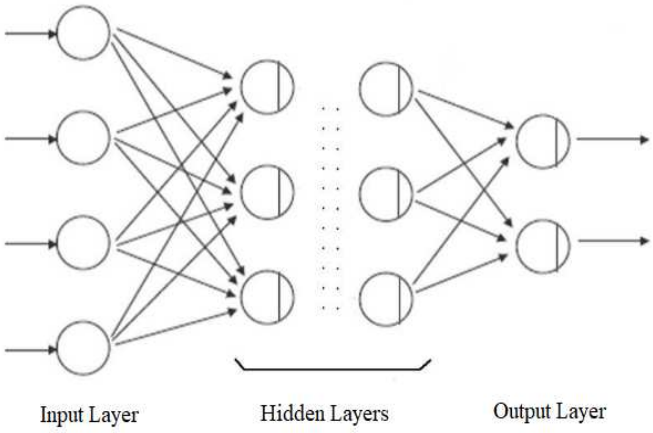

Fig. 1. ANN Layers

The neural network layers and functions of in Fig.1 are explained in the following items. These;

- Input layer: It is the layer where data/inputs from the outside world are received. No information processing is performed on this layer. The transmission of data to the middle layer is performed [10, 12].
- Intermediate (hidden) layers: Data from the input layer are processed in this layer. It is possible to solve many problems with one intermediate layer. If the relationship between input/output of the problem the network is desired to learn is not linear and the complexity increases, it is used in more than one middle layer. After the network's learning is completed, the information obtained is transmitted to the output layer [10, 12].
- Output layer: It produces output by processing the information coming from the middle layer to the output

produced by the network for the input provided to the network from the input layer. This output is transmitted to the outside world [12, 15, 16].

Training is carried out on the data based on ANN layers and functions. The network learning performance of ANN can be measured according to the performance of the training. To measure the training performance of ANN, after the training is over, the network's decision about them is checked by showing the examples that the network did not see during the training. If the network can produce correct answers to instances that it has not seen, then it is said to have good performance and learned the network.

III. DATASET

In this study, earthquake data affecting the province of Elazig on the Eastern Anatolian fault line were examined. The data were obtained from Boğaziçi University Kandilli Observatory and Earthquake Research Institute (KOERI). Earthquake data includes earthquakes in the circular area with a radius of 70 kilometers, taking the start of 38.35 latitudes 39.15 longitudes. A dataset containing all earthquakes with a depth of 0-500 and a magnitude (Mag.) of 0-9.0 was obtained between 21.10.1900 and 21.10.2020. There are 8216 records in this dataset.

- Each record contained information such as date, time, latitude (Lat.), longitude (Long.), depth (km), magnitude (xM, MD, ML, Mw, Ms, and Mb), type, and city. Since there was no data for each record in this information, MD, ML, Mw, Ms, and Mb sizes and type information were deleted. Time information (T. Zone); divided into zones morning (8-18), evening (19-24), and night (0-7). Type of explosion 16 records were deleted. According to the location information; there were 5563 ELAZIG, 2085 MALATYA, 204 ADIYAMAN, 30 SANLIURFA, 229 DIYARBAKIR, 86 TUNCELİ. 3 records without location information were deleted. ELAZIG city's(C) records were arranged as 1 and other provinces as 0. Dataset1 was created by removing noisy data. Sample records of Dataset1 are given in the table (Table I).

- Aftershocks that occurred after the 6.8 magnitude Elazig earthquake on January 24, 2020 reduced the analysis performance of Dataset1. To increase this performance, Dataset2 was created by removing the data after January 24, 2020 from Dataset1.

- To increase the effects of dataset2 on earthquake prediction performance, Dataset3 was created by adding earthquake distances. Dataset3 was created by calculating the three-dimensional distance (Dis.) between two earthquake data (Table II).

In this study, prediction performance was evaluated by analyzing earthquake datasets with ANNs in time series analysis.

TABLE I. DATASET1 SAMPLE RECORD

Date	T. Zone	Lat.	Long.	Depth	Mag.	C
08.02.1930	1	38,52	39,4	100	5,3	1
09.01.1931	1	38	38,5	30	5,2	0
23.09.1940	3	38,96	39,32	80	5,2	0
18.08.1948	3	38,51	39,25	10	5,3	1
25.04.1949	3	38,27	38,99	80	5,5	0

TABLE II. DATASET3 SAMPLE RECORD

Date	T. Zone	Lat.	Long.	Depth	Mag.	Dis.
08.02.1930	1	38,52	39,4	100	5,3	0
09.01.1931	1	38	38,5	30	5,2	130,102
23.09.1940	3	38,96	39,32	80	5,2	133,267
18.08.1948	3	38,51	39,25	10	5,3	80,579
25.04.1949	3	38,27	38,99	80	5,5	78,542

IV. METHOD

For the analysis of earthquake data, ANNs, which are preferred of non-linear datasets in time series analysis methods, were used. It was used on the Matlab platform for data analysis of ANNs. Analysis steps are shown below.

- The Matlab platform proposes three types of problems for the analysis of linear and non-linear data.
  i. Predicting the future data using current data and historical data (Problem type 1)
  ii. Predict the future data using historical data (Problem type 2)
  iii. Predict the future data using current data (Problem type 3)

- After the problem type is selected, data set selection and pre-processing will be performed. After selecting the dataset, pre-operations will be carried out, and the application step will be started.

- In the implementation phase, there are three types of target time series analysis steps: Training (Tra.), verification (Ver.), and testing (Tes.) steps. After the training, validation, and test values are determined, the network model is created. In the network model, the number of hidden layers and the number of delays are valued. There is no certainty for giving these values. The trial and error method is used to find the best results.

- With the determination of the network model for three non-linear problems, the network model is displayed in the Matlab application.

- The network model will be constructed and trained in an open-loop fashion. Open-loop (single step) is more efficient than closed-loop (multi-step) training. The open-loop provides the correct historical outputs to the network while training it to produce direct current outputs. After

training, the network can be converted to a closed-loop form or any other application needs [16].

- After the ANN model is selected, the algorithms used to train the network are selected on the training screen of the Matlab application. These algorithms (Alg); Levenberg-Marquardt (L), Bayesian Regularization (B), Scaled Conjugate Gradient(S).
- After the ANN model training process, a comparison is made using performance criteria. Mean square error (MSE: Values close to zero indicate better performance) and Regression (R: Values close to one indicate better performance.) measures are used as performance measures [16]. According to these criteria, the performance value of training, validation, and test data is evaluated. The performance of education can be examined both in terms of values and graphics.

According to the result values and/or result in graphics of the ANN training, it can be decided to continue or complete the training.

V. Results

Within the scope of the study, earthquakes on the Eastern Anatolian fault line affecting the province of Elazig were examined. These records contain the date, time, latitude, longitude, depth (km), and magnitude information of the earthquake. Noisy data of recordings and recording information are arranged. As a result of the regulations, Dataset1 was created.

ANN training of Dataset1 was carried out on Matlab platform with problem types. Training performance in raw data has been tested all problem types and training algorithms (constant training/validation/tests and hidden neurons/delays) in the Dataset1. Educational performance results are given in Table III.

TABLE III. Dataset1 Training Performance

Dataset	Problem Type	H/D	Alg.	MSE	R
1	1	10/2	L	3.04e-1	6.98 e-1
1	1	10/2	B	2.98 e-1	6.93 e-1
1	1	10/2	S	3.32 e-1	6.52 e-1
1	2	10/2	L	3.00 e-1	6.86 e-1
1	2	10/2	B	3.07 e-1	6.84 e-1
1	2	10/2	S	3.15 e-1	6.76 e-1
1	3	10/2	L	3.07 e-1	6.89 e-1
1	3	10/2	B	2.96 e-1	6.95 e-1
1	3	10/2	S	3.22 e-1	6.56 e-1

In the Table III, as a result of the pieces of training, the L and B algorithms provide the highest performance in Dataset1 in problem 1 and 2 types. However, many records are detected incorrectly. The error rate is between -2 and 3. Almost all records have errors. For this reason, educational performance is not at the desired level. When the training graphs were examined, it was observed that the data training was disrupted due to the aftershocks that occurred after January 24, 2020. Since high-magnitude earthquakes are not frequently encountered in the region, the need to test the education level arose by removing them from the dataset. With this requirement, the Dataset2 was created.

By testing all problem types and training algorithms (regular training/validation/tests and hidden neurons/delays) in dataset2, training performance in raw data has been tested. As a result of these pieces of training, it is observed that the performance has increased according to dataset1. However, there is an error rate between -2 and 2 for earthquake records, which is unacceptable for earthquake data. For this reason, values are given to produce the most accurate results for hidden neurons and delay data. The results with these values are shown in Table IV.

TABLE IV. Dataset2 Traning Performance

Dataset	Problem Type	H/D	Alg.	MSE	R
2	2	10/2	L	1.72e-1	7.45 e-1
2	2	10/2	L	1.76e-1	7.31e-1
2	2	10/2	L	1.85e-1	7.29e-1
2	2	10/2	L	1.40e-1	7.94e-1
2	2	10/2	B	1.92e-1	7.03e-1
2	2	10/2	B	1.91e-1	7.05e-1
2	2	10/2	B	1.90e-1	7.19e-1
2	2	10/2	B	1.93e-1	7.09e-1

As seen in Table IV, MSE values and R values increased in all algorithms and problem types. However, despite the increase in performance, there is an error rate in all earthquake data. These error rates are between -2 and 2 values. The performance graph of the record with the best performance R-value of 7.05 in Table IV is given in Fig. 2.

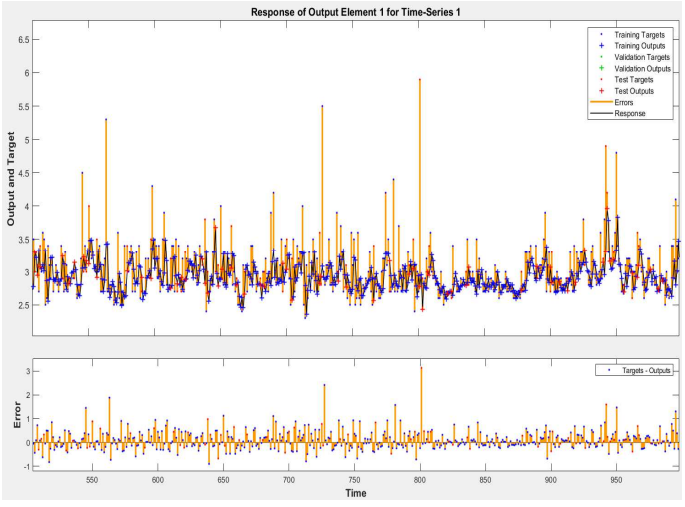

Fig. 2. Dataset2 training error rates graphic

The error rates are observed in the data between 500-1000 in Fig. 2. In almost all data, errors occurred in the prediction performance. These error rates are not acceptable for earthquake data. For this reason, the values of hidden neurons and delays data have been changed to produce the most accurate results (H:35 D:15, H:25 D:15, H:35 D:5, H:50 D:15) (fixed training/validation/ tests). The performance results in Dataset2 with these values are shown in Table V.

TABLE V. DATASET2 TRANING H/D VALUES PERFORMANCE

Dataset	Problem Type	H / D	Alg.	MSE	R
2	1	35/15	L	1.30e-1	8.24e-1
2	1	25/15	L	1.38e-1	8.00e-1
2	1	35/5	L	1.26e-1	8.20e-1
2	1	50/15	L	2.30e-1	6.50e-1
2	1	35/15	B	2.36e-1	9.96e-1
2	1	25/15	B	1.22e-1	9.83e-1
2	1	35/5	B	9.00e-2	8.78e-1
2	1	50/15	B	8.18e-4	9.98e-1
2	2	35/15	L	1.72e-1	7.45e-1
2	2	25/15	L	1.76e-1	7.31e-1
2	2	35/5	L	1.85e-1	7.29e-1
2	2	50/15	L	1.40e-1	7.94e-1
2	2	35/15	B	1.49e-1	7.81e-1
2	2	25/15	B	1.91e-1	7.05e-1
2	2	35/5	B	1.90e-1	7.19e-1
2	2	50/15	B	1.93e-1	7.09e-1

In Table V, higher performance was obtained with varying H and D values in the data of Dataset2. The highest performances were obtained in problem type1 with different H and D values. The results obtained by the Bayesian regularization algorithm with problemtype1 H:25 H:15 are given in Fig. 3.

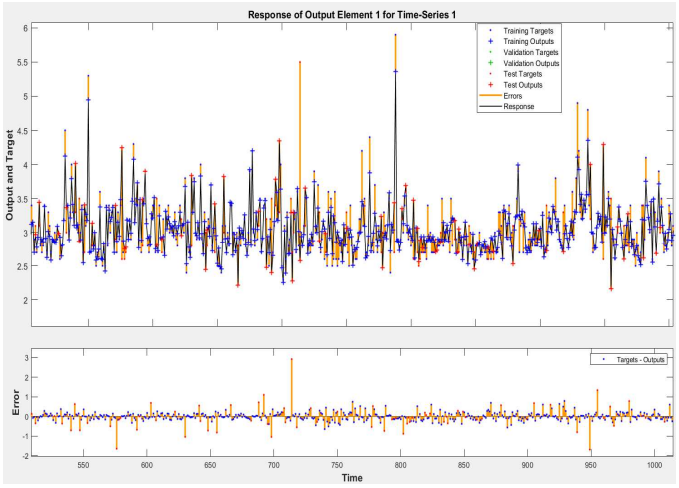

Fig. 3. Dataset2 training H/D values error rates graphic

As shown in Fig. 3, there are very high errors, although the error rates decrease compared to Fig.2. These error rates are seen in all data, albeit slightly. To reduce these errors in earthquake prediction, three-dimensional distances between earthquakes are added to Dataset2. Dataset3 was created with the added distance data. H and D values in Dataset3 and problemtype1 and training analyzes were made with B and L algorithms. The results of training are given in Table VI.

TABLE VI. DATASET3 TRANING H/D VALUES PERFORMANCE

Dataset	Problem Type	H / D	Alg.	MSE	R
3	1	35/15	L	1.11e-1	8.58e-1
3	1	25/15	L	1.98e-1	7.21e-1
3	1	35/5	L	1.44e-1	8.10e-1
3	1	50/15	L	7.05e-2	9.07e-1
3	1	35/15	B	1.13e-3	9.98e-1
3	1	25/15	B	1.66e-2	9.78e-1
3	1	35/5	B	8.75e-2	8.82e-1
3	1	50/15	B	7.32e-1	9.99e-1

In Table VI, it is seen that training with Bayesian regularization algorithm is more successful. As the most successful training, the training result with good MSE and R criteria was selected. According to this result, the R criterion was analyzed as 9.77e-1 with the Bayesian regularization algorithm, which was H:25 D:15 according to problem type 1 in Dataset3. The graph of this analysis result is given in Fig. 4.

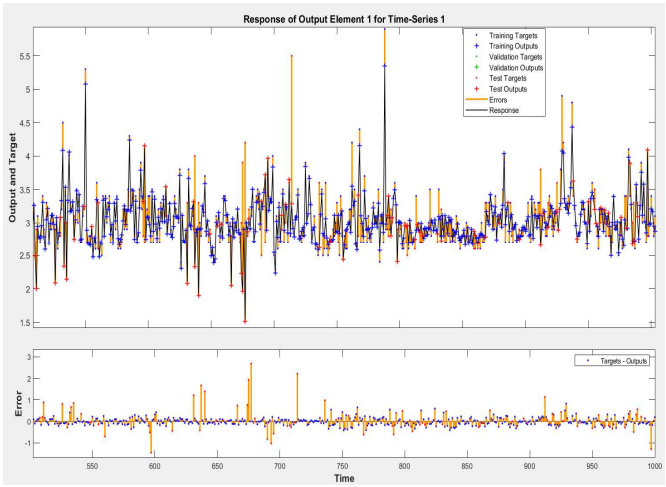

Fig. 4. Dataset3 training H/D values error rates graphic

According to Fig. 4, small-magnitude earthquakes are estimated with 0.0/0.1/0.2 difference in the training of earthquake data. ANN training performance is high in the training of small earthquakes. However, in ANN training, the error rate is high in large earthquakes. The prediction performance of the ANN for large earthquakes is low. The reason for this situation is due to the insufficient number of large earthquakes in the dataset..

VI. Conclusion and Evaluation

In recent years, increasing technologies and digital records have created the need to obtain information. Based on this need, methods were developed to obtain information from data. One of these methods is time series analysis. Time series analysis provides predictions about future data by analyzing time-dependent data. There are many estimation methods in time series analysis. These estimation methods were divided into linear and non-linear methods. The need for non-linear methods has increased for real-world data that linear methods cannot solve. In line with these needs, non-linear methods such as ANNs are used in time series.

The performance of ANNs in training non-linear data is increasing day by day. For this reason, earthquake prediction performance was measured from non-linear earthquake data. Earthquake data on the eastern Anatolian fault line affecting the province of Elazig were obtained from the KOERI. Missing data in earthquake data were deleted, and arrangements were made. ANN training was applied to the datasets (Dataset1, Dataset2, and Dataset3) obtained as a result of the regulations on the Matlab platform. Since there is no rule for determining the ANN success variables in the pieces of training, values were determined by the trial and error method (for problem type, hidden neurons, delays, and training algorithms). The performances were evaluated as a result of the pieces of training made on the determined values and datasets.

It was observed that Dataset1 achieved the highest performance in the Levenburg and Bayesian algorithms in problem1 and problem2 types. However, due to the high error rates, it was removed from the dataset after the January 24, 2020 Elazig earthquake, and Dataset2 was created. With the addition of hidden neurons and delay values to Dataset2, the success rate has increased, but it is insufficient due to the high error rates. Dataset3 was created by adding the three-dimensional distance between the two earthquakes to the data of Dataset2 earthquakes. It was been observed that ANN training in Dataset3 show successfull performance in small-magnitude earthquakes, but not in high-magnitude earthquakes.

As a result of the ANN training conducted with earthquake data, it was determined that the data were insufficient to predict large earthquakes. By expanding the dataset, success can be achieved in the prediction of large earthquakes. The success rate can be increased if the scientific and natural conditions that occur during the earthquake are included in the data set. Determining the damages caused by the energy effects resulting from the earthquake and adding them to the dataset can increase the earthquake prediction performance.

References

[1] Z. Ozpolat and M. Karabatak, "Temperature Estimation with Time Series Analysis from Air Quality Data Set," 2019 7th International Symposium on Digital Forensics and Security (ISDFS), 2019, pp. 1-5.

[2] I. Kırbaş, "İstatistiksel Metotlar ve Yapay Sinir Ağları Kullanılarak Kısa Dönem Çok Adımlı Rüzgâr Hızı Tahmini", Sakarya University Journal of Science Institute, 2018, 22(1), pp 24-38.

[3] S. Narayanakumar and K. Raja. "A BP artificial neural network model for earthquake magnitude prediction in himalayas, india", Circuits and Systems, 7(11):3456, 2016.

[4] E. Çelik, M. Atalay and H. Bayer, "Yapay Sinir Ağları ve Destek Vektör Makineleri ile Deprem Tahmininde Sismik Darbelerin Kullanılması", IEEE 22nd Signal Processing And Communications Applications Conference, 23-25 April 2014, Trabzon, pp. 730-733.

[5] D. Kwiatkowski, P. Phillips, P. Scmidt, and Y. Shin "Testing the null hypothesis of stationarity against the alternative of a unit root." Journal of econometrics, The Netherlands, 1992, pp 1-3: 159-178.

[6] S. Oğhan, "Zaman Serisi Analiz Yöntemlerinin Karşılaştırılması", Master Thesis, 15 February 2010, İzmir, pp. 1-7.

[7] A. Lasfer, "Performance analysis of artificial neural networks in forecasting financial time series", Yüksek Lisans Tezi, American University of Sharjah, 2013.

[8] H. Yıldız Bozkurt, "Zaman Serisi Analizi", 2nd Enlarged Edition, pp. 6-26.

[9] A.W. Van der Vaart, "Time Series", VU University, Amsterdam, Lecture Notes, 1995-2010, pp 2-21.

[10] E. Öztemel, "Yapay Sinir Ağları", Papatya Publishing, Istanbul, 2013.

[11] TZ. Khalaf, "The Approach of Hybrid PSO-ANN AND PSO Models Based on Estimation of Cost and Duration of Construction Projects", Master Thesis, Kastamonu University, 2020.

[12] D. Kwiatkowski, P. Phillips, P. Scmidt, and Y. Shin "Testing the null hypothesis of stationarity against the alternative of a unit root." Journal of econometrics, The Netherlands, 1992, pp 1-3: 159-178.

[13] J. Mahmoudi, M. A. Arjomand, M. Rezaei, and M. H. Mohammadi, "Predicting the earthquake magnitude using the multilayer perceptron neural network with two hidden layers", Civil Engineering Journal, 2016, 2(1):1–12.

[14] C. Li and X. Liu. "An improved pso-bp neural network and its application to earthquake prediction", In Control and Decision Conference (CCDC), Chinese, 2016, ss 3434–3438.

[15] J. Sheng, D. Mu, H. Zhang ve H. Lv, "Seismotectonics Considered Artificial Neural Network Earthquake Prediction in Northeast Seismic Region of China", The Open Civil Engineering Journal, 2015, 9, ss 522-528.

[16] F.S. Çakır, "Yapay Sinir Ağları: MATLAB kodları ve MATLAB Toolbox Çözümleri", Nobel Academic Publishing, 2019.