Hindawi Publishing Corporation
Applied Computational Intelligence and Soft Computing
Volume 2012, Article ID 698071, 6 pages
doi:10.1155/2012/698071

Research Article

Monthly Rainfall Estimation Using Data-Mining Process

Özlem Terzi

Faculty of Technical Education, Suleyman Demirel University, 32260 Isparta, Turkey

Correspondence should be addressed to Özlem Terzi, ozlemterzi@sdu.edu.tr

Received 16 April 2012; Revised 13 July 2012; Accepted 18 July 2012

Academic Editor: Tzung P. Hong

It is important to accurately estimate rainfall for effective use of water resources and optimal planning of water structures. For this purpose, the models were developed to estimate rainfall in Isparta using the data-mining process. The different input combinations having 1-, 2-, 3- and 4-input parameters were tried using the rainfall values of Senirkent, Uluborlu, Eğirdir, and Yalvaç stations in Isparta. The most appropriate algorithm was determined as multilinear regression among the models developed with various data-mining algorithms. The input parameters of Multilinear Regression model were the monthly rainfall values of Senirkent, Uluborlu and Eğirdir stations. The relative error of this model was calculated as 0.7%. It was shown that the data mining process can be used in estimation of missing rainfall values.

1. Introduction

The meteorological events affect permanently human life. Considering the meteorological phenomena, which have no possibility of intervention, they cause the important results in human life, accurate estimation and analysis of these variables are also very important. Precipitation, which is generating flow, is an important parameter. The occurrence of extreme rainfall in a short time causes significant events that affect human life such as flood. However, in the event of insufficient rainfall in long period occurs drought. Thus, rainfall estimation is very important in terms of effects on human life, water resources, and water usage areas. However, rainfall affected by the geographical and regional variations and features is very difficult to estimate. Nowadays, there are many researches about artificial intelligence methods used in the estimation of rainfall [1–7]. Partal et al. [8] developed rainfall estimation models using artificial neural networks and wavelet transform methods. Bodri and Čermák [9] evaluated the applicability of neural networks for precipitation prediction. Chang et al. [10] applied a modified method, combining the inverse distance method and fuzzy theory, to precipitation interpolation. They used genetic algorithm to determine the parameters of fuzzy membership functions, which represent the relationship between the location without rainfall records and its surrounding rainfall gauges. They worked to minimize the estimated error of precipitation with the optimization process.

One of the aims of storing this data in databases and receiving data from many sources is to convert raw data into information at present. This process is called as data-mining (DM) process of converting data into information. In recent years, the use of data-mining process in the field of hydrology is increasing. The studies have been performed using DM process in many areas [11–13]. Keskin et al. [14] developed integrated evaporation model using DM process for three lakes in Turkey. Terzi [15] developed the models to forecast flow of Kızılırmak River using rainfall and flow parameters with DM process. Terzi et al. [16] proposed various solar radiation models with DM process using air temperature, relative humidity, wind speed, and air pressure parameters and evaluated performance of the models. Teegavarapu [17] evaluated the use of association rule mining (ARM) in conjunction with a spatial interpolation technique to estimate of missing precipitation data and to overcome one of the major limitations of spatial interpolation techniques. Solomatine and Dulal [18] investigated the comparative performance of artificial neural networks (ANNs) and model trees (MTs) in rainfall—runoff transformation. They determined that both ANNs and MTs produce excellent results for 1-h ahead prediction, acceptable results for 3-h ahead prediction and conditionally acceptable result for 6-h

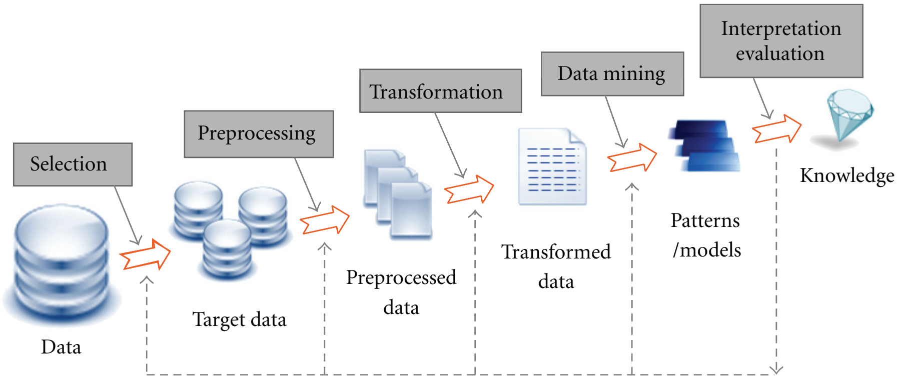


Figure 1: Knowledge discovery process.

ahead prediction. They obtained almost similar performance for 1-h ahead prediction of runoff, but the result of the ANN is slightly better than the MT for higher lead times from these techniques. Keskin et al. [19] applied data-mining process to river flow prediction. They determined that it was possible using data-mining process for river flow prediction. Teegavarapu and Chandramouli [6] developed a model that uses artificial neural network concepts and a stochastic interpolation technique. They tested the model for estimation of missing precipitation data.

The aim of the study is to evaluate the use of data-mining process to estimate rainfall of Isparta in Turkey. This study is performed using rainfall data of Senirkent, Uluborlu, Eğirdir, and Yalvaç stations in Isparta city.

2. Data-Mining Process

Knowledge discovery is a process that extracts implicit, potentially useful or previously unknown information from the data. The knowledge discovery process is described in Figure 1.

Let us examine the knowledge discovery process in the diagram in Figure 1 in details.

(i) Data coming from variety of sources is integrated into a single data store called target data.

(ii) Data then is preprocessed and transformed into standard format.

(iii) The data-mining algorithms process the data to the output in form of patterns or rules.

(iv) Then those patterns and rules are interpreted to new or useful knowledge or information.

The ultimate goal of knowledge discovery and data-mining process is to find the patterns that are hidden among the huge sets of data and interpret them to useful knowledge and information. As described in process diagram above, data-mining is a central part of knowledge discovery process.

The data-mining definition is defined as "the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions" [20]. This data-mining definition has business flavor and for business environments. However, data-mining is a process that can be applied to any type of data ranging from weather forecasting, electric load prediction, product design, among others.

Data-mining also can be defined as the computer-aid process that digs and analyzes enormous sets of data and then extracting the knowledge or information out of it. By its simplest definition, data-mining automates the detections of relevant patterns in database [21].

The emergence of knowledge discovery in databases (KDD) as a new technology has been brought about with the fast development and broad application of information and database technologies. The process of KDD is defined as an iterative sequence of four steps: defining the problem, data preprocessing (data preparation), data-mining, and postdata-mining.

2.1. Defining the Problem. The goals of a knowledge discovery project must be identified. The goals must be verified as actionable. For example, if the goals are met, a business organization can then put the newly discovered knowledge to use. The data to be used must also be identified clearly.

2.2. Data Preprocessing. Data preparation comprises those techniques concerned with analyzing raw data so as to yield quality data, mainly including data collecting, data integration, data transformation, data cleaning, data reduction, and data discretization.

2.3. Data-Mining. Given the cleaned data, intelligent methods are applied in order to extract data patterns. Patterns of interest are searched for, including classification rules or trees, regression, clustering, sequence modeling, dependency, and so forth.

2.4. Postdata-Mining. Post data-mining consists of pattern evaluation, deploying the model, maintenance, and knowledge presentation.

The KDD process is iterative. For example, while cleaning and preparing data, it might be discovered that data from a certain source is unusable, or that data from a previously unidentified source is required to be merged with the other

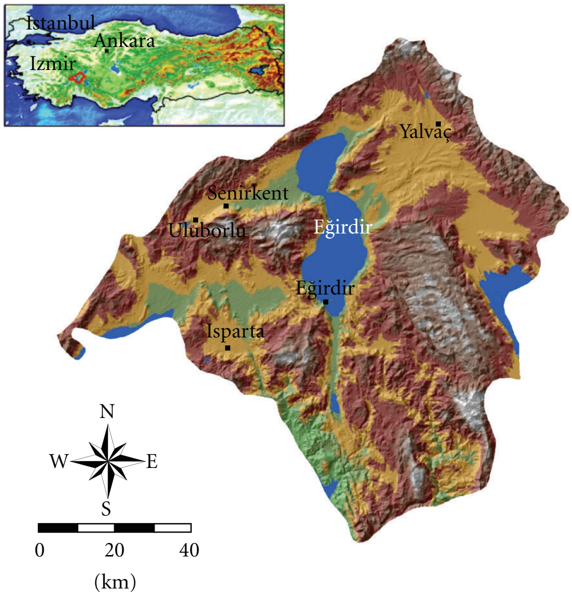


Figure 2: Locations of rain gauges in Isparta.

data under consideration. Often, the first time through, the data-mining step will reveal that additional data cleaning is required [22].

3. Study Region and Data

In this study, the data used to developed rainfall estimation models are the monthly rainfall data of Isparta, Senirkent, Uluborlu, Eğirdir, and Yalvaç stations. The Isparta city is located in the Lakes Region located in the north of the Mediterranean Region, and between 30°20′ and 31°33′ east longitudes and 37°18′ and 38°30′ north latitudes. The altitude of Isparta having a surface area of 8933 km^2 is the average of 1050 m. The average annual total rainfall of Isparta is 440.3 kg/m^2. The most of rainfall (72.69%) has occurred in winter and spring months. The summer and autumn months are quite dry (29.31% of total rainfall). While it is observed usually rain, occasional snow in winter in the region, it is observed in the form of rainstorm the in spring and summer months. The study region and the locations of rain gauges are shown in Figure 2.

The monthly rainfall data for 1964–2005 years used in this study were obtained from Turkish State Meteorological Service. The various rainfall estimation models were developed for Isparta using the rainfall values of Senirkent, Uluborlu, Eğirdir, and Yalvaç stations as input parameters. It was investigated whether or not there are any missing data. Then, the mean values were used for substitution of missing values. The training dataset consisted of the 1964–1996 years was used to develop the models. The trained models were used to run the testing dataset for 1997–2005 years.

4. Model Performance Criteria

In the model assessment stage, after it has built a set of models using different algorithms, these models were evaluated in terms of accuracy. There are a few popular criteria to evaluate the quality of a model. It was chosen coefficient of determination (R^2) and root mean-squared error (RMSE) which are the most well known and the commonly used performance criteria [23–25]. The R^2 is the proportion of variability in a dataset that is accounted for by the statistical model. The RMSE is valuable and because it indicates error in the units (or squared units) of the constituent of interest, which aids in analysis of the results. The coefficient of determination based on the rainfall estimation errors is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(P_{i(\text{rainfall})} - P_{i(\text{model})}\right)^2}{\sum_{i=1}^{n}\left(P_{i(\text{rainfall})} - P_{\text{mean}}\right)^2}, \qquad (1)$$

where $n$is the number of observed data, $P_{i(\text{rainfall})}$ and $P_{i(\text{model})}$ are monthly rainfall measurement and the results of the developed model, respectively, and P_{mean} is mean rainfall measurements.

The root mean square error represents the error of model and defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_{i(\text{rainfall})} - P_{i(\text{model})}\right)^2}, \qquad (2)$$

where parameters have been defined above.

5. Rainfall Estimation Models

For rainfall estimation, Decision Table, KStar, Multilinear Regression, M5'Rules, Multilayer Perceptron, RBF Network, Random Subspace, and Simple Linear Regression algorithms were used in this study. The fifteen models were developed using different input combinations with the rainfall values of Senirkent, Uluborlu, Eğirdir and Yalvaç stations to estimate rainfall of Isparta station. These models including 1-input, 2-input, 3-input and 4-input parameters were given in Tables 1, 2, 3, and 4, respectively.

Firstly, the relationships between rainfall data of Isparta station and them of other stations (Senirkent, Uluborlu, Eğirdir, and Yalvaç) were investigated using statistical analyses. The effective variables on Isparta station were ranked in the order of Senirkent, Uluborlu, Eğirdir, and Yalvaç stations. The performance criteria of the models developed with 1-input parameters were given in Table 1 for testing set.

Examining the models given in Table 1, it was determined as the highest R^2 value was 0.745 and lowest RMSE value was 48.44 mm for models developed using Multilinear Regression (MLR), M5'Rules, and Simple Linear Regression algorithms with rainfall data of Senirkent station. These models have the same R^2 and RMSE values. The worst model with the highest RMSE (141.50) was developed with decision table. When the developed models by using MLR, M5'Rules, and Simple Linear Regression algorithms were analyzed, the input parameter of the best performing model was rainfall of Senirkent station. Later, the best models were generally ranked in Uluborlu, Eğirdir, and Yalvaç stations. In Table 2, it was given the R^2 and RMSE values of developed models with 2-input parameters.

Table 1: The performance criteria of the models having 1-input parameter.

Input parameters	Eğirdir		Senirkent		Uluborlu		Yalvaç	
Models	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Decision Table	0.254	141.5	0.695	57.90	0.638	68.62	0.531	89.10
KStar	0.686	59.60	0.641	68.14	0.648	66.82	0.543	86.70
Multilinear Regression	0.671	62.49	0.745	48.44	0.717	53.63	0.616	72.84
M5'Rules	0.671	62.49	0.745	48.44	0.717	53.63	0.616	72.84
Multilayer Perceptron	0.711	54.89	0.649	66.58	0.653	65.81	0.578	80.06
RBF Network	0.533	88.67	0.641	68.13	0.672	62.28	0.495	95.81
Random Subspace	0.617	72.71	0.634	69.56	0.590	77.77	0.492	96.43
Simple Linear Regression	0.671	62.49	0.745	48.44	0.717	53.63	0.616	72.84

Table 2: The performance criteria of the models having 2-input parameters.

Input parameters	Eğirdir-Uluborlu		Eğirdir-Yalvaç		Eğirdir-Senirkent		Senirkent-Uluborlu		Senirkent-Yalvaç		Uluborlu-Yalvaç	
Models	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Decision Table	0.638	68.62	0.254	141.52	0.695	57.90	0.695	57.90	0.695	57.90	0.638	68.62
KStar	0.765	44.52	0.732	50.83	0.751	47.21	0.727	51.81	0.668	62.93	0.684	60.07
Multilinear Regression	0.807	36.65	0.743	48.80	0.792	39.40	0.765	44.60	0.745	48.44	0.717	53.63
M5'Rules	0.807	36.65	0.743	48.80	0.792	39.40	0.765	44.60	0.745	48.44	0.717	53.63
Multilayer Perceptron	0.796	38.64	0.743	48.69	0.746	48.29	0.678	61.08	0.662	64.12	0.670	62.64
RBF Network	0.663	63.87	0.550	85.41	0.568	81.96	0.647	67.02	0.556	84.19	0.567	82.21
Random Subspace	0.782	41.45	0.620	72.05	0.725	52.27	0.695	57.93	0.610	74.12	0.638	68.65
Simple Linear Regression	0.717	53.63	0.671	62.49	0.745	48.44	0.745	48.44	0.745	48.44	0.717	53.63

Table 3: The performance criteria of the models having 3-input parameters.

Input parameters	Senirkent-Uluborlu-Eğirdir		Senirkent Uluborlu-Yalvaç		Senirkent-Yalvaç-Eğirdir		Uluborlu-Yalvaç-Eğirdir	
Models	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Decision Table	0.695	57.90	0.695	57.90	0.695	57.90	0.638	68.62
KStar	0.771	43.54	0.693	58.20	0.745	48.33	0.771	43.43
Multilinear Regression	0.813	35.43	0.765	44.60	0.792	39.40	0.798	38.38
M5'Rules	0.808	36.43	0.765	44.60	0.792	39.40	0.711	54.89
Multilayer Perceptron	0.774	42.83	0.726	51.98	0.772	43.33	0.797	38.55
RBF Network	0.622	71.67	0.560	83.48	0.583	79.23	0.574	80.90
Random Subspace	0.760	45.62	0.680	60.83	0.714	54.31	0.757	46.12
Simple Linear Regression	0.745	48.44	0.745	48.44	0.745	48.44	0.717	53.63

Table 4: The performance criteria of the models having 4-input parameters.

Modeller	R^2	RMSE
Decision Table	0.695	57.90
KStar	0.761	45.33
Multilinear Regression	0.806	36.89
M5'Rules	0.766	44.35
Multilayer Perceptron	0.774	42.91
RBF Network	0.573	80.95
Random Subspace	0.757	46.17
Simple Linear Regression	0.745	48.44

As seen from Table 2, the highest R^2 (0.807) and lowest RMSE (36.65) values were obtained for MLR and M5'Rules models developed using rainfall values of Eğirdir and Uluborlu stations. Table 2 shows that increasing of number of the model input parameter improved the performance of the models. While R^2 value of the best model with one input parameter was 0.745, performance of the model with two input parameters is 0.807. The models having 3-input parameters are shown in Table 3.

It was shown that the R^2 values of the models having 3-input parameters (Senirkent—Uluborlu—Eğirdir) were 0.813 and 0.808 for MLR and M5'Rules models in Table 3, respectively. The MLR (Senirkent—Uluborlu—Eğirdir)

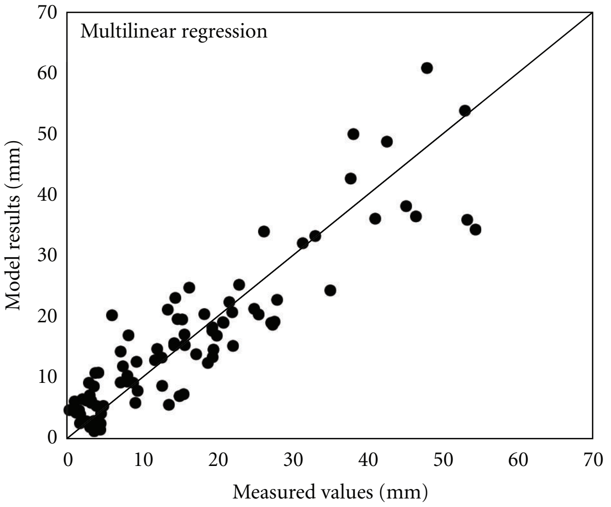


Figure 3: Comparison plot for MLR model.

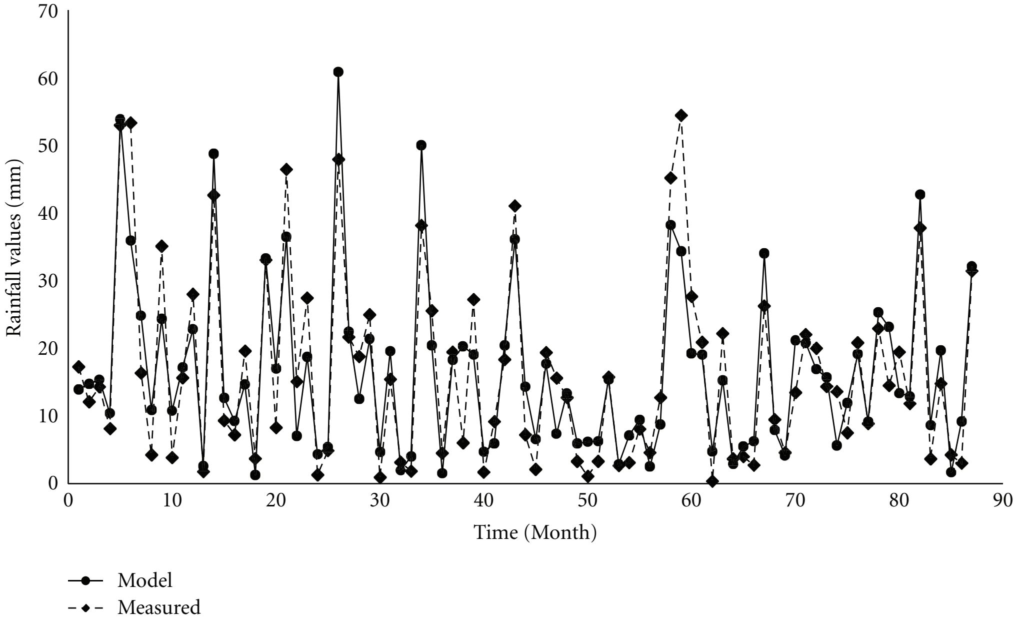


Figure 4: Time series for MLR model.

model showed the best performance. The models developed with Senirkent, Uluborlu, and Eğirdir stations ranked according to statistical analysis showed generally the better performance. The model with the worst performance was Radial Basis Function (RBF) network model. The models having 4-input parameters are shown in Table 4.

It was shown that the R^2 value of the model having three 4-input parameters were 0.806 for MLR model in Table 4. When Yalvaç station was added to the best 3-input model, the 4-input model performance had decreased slightly. The MLR and M5'Rules algorithms in all the DM algorithms gave generally the best results and had the almost same performance except the 4-input model. While the RBF network from artificial neural network algorithms showed the worst performance in all DM models, MLR had relatively good results. Considering all the DM models, MLR model with 3-input parameters ($R^2 = 0.813$) showed the best performance. Examining RMSE values of the model, the model (Senirkent—Uluborlu—Eğirdir) had the lowest error. Thus, the monthly rainfall results of MLR model (Senirkent—Uluborlu—Eğirdir) are shown in Figures 3 and 4 as comparison plot and time series for testing data set. Figure 3 shows that the MLR model comparison plot was uniformly distributed around the 45° straight line implying that there were no bias effects. It was apparent a good relationship between estimated and measured rainfall values. The relative error between the measured values and the value of the developed MLR model was calculated as 0.7%.

It was shown that, for Isparta region, the developed MLR model gave the best results to estimate rainfall. They cannot be used to estimate rainfall of another region, because the MLR models were developed for Isparta region. For a different region, the models need to be reestablished or need to be calibrated according to data of a new region. In the future, when more data are obtained, the developed models need to be revised. The other methods can give better results than MLR when adding more data or developing model for different region.

6. Conclusions

The rainfall which is an important factor for the use of water resources is a difficult variable to estimate. In this study, data-mining process was used to estimate monthly rainfall values of Isparta. The monthly rainfall data of Senirkent, Uluborlu, Eğirdir, and Yalvaç stations were used to develop rainfall estimation models. When comparing the developed models to measured values, multilinear regression model from data-mining process gave more appropriate results than the developed models in this study. The input parameters of the best model were the rainfall values of Senirkent, Uluborlu, and Eğirdir stations. Consequently, it was shown that the data-mining process, producing a solution more quickly than traditional methods, can be used to complete the missing data in estimating rainfall.

References

[1] T. B. Trafalis, M. B. Richman, A. White, and B. Santosa, "Data mining techniques for improved WSR-88D rainfall estimation," *Computers and Industrial Engineering*, vol. 43, no. 4, pp. 775–786, 2002.

[2] K. C. Luk, J. E. Ball, and A. Sharma, "An application of artificial neural networks for rainfall forecasting," *Mathematical and Computer Modelling*, vol. 33, no. 6-7, pp. 683–693, 2001.

[3] M. Zhang, J. Fulcher, and R. A. Scofield, "Rainfall estimation using artificial neural network group," *Neurocomputing*, vol. 16, no. 2, pp. 97–115, 1997.

[4] T. Shoji and H. Kitaura, "Statistical and geostatistical analysis of rainfall in central Japan," *Computers and Geosciences*, vol. 32, no. 8, pp. 1007–1024, 2006.

[5] M. C. V. Ramírez, H. F. C. Velho, and N. J. Ferreira, "Artificial neural network technique for rainfall forecasting applied to the São Paulo region," *Journal of Hydrology*, vol. 301, no. 1–4, pp. 146–162, 2005.

[6] R. S. V. Teegavarapu and V. Chandramouli, "Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records," *Journal of Hydrology*, vol. 312, no. 1–4, pp. 191–206, 2005.

[7] Y.-M. Chiang, F. J. Chang, B. J. D. Jou, and P. F. Lin, "Dynamic ANN for precipitation estimation and forecasting from radar observations," *Journal of Hydrology*, vol. 334, no. 1-2, pp. 250–261, 2007.

[8] T. Partal, E. Kahya, and K. Cığızoğlu, "Estimation of precipitation data using artificial neural networks and wavelet transform," *ITU Journal*, vol. 7, no. 3, pp. 73–85, 2008 (Turkish).

[9] L. Bodri and V. Čermák, "Prediction of extreme precipitation using a neural network: application to summer flood occurrence in Moravia," *Advances in Engineering Software*, vol. 31, no. 5, pp. 311–321, 2000.

[10] C. L. Chang, S. L. Lo, and S. L. Yu, "Applying fuzzy theory and genetic algorithm to interpolate precipitation," *Journal of Hydrology*, vol. 314, no. 1–4, pp. 92–104, 2005.

[11] C. Damle and A. Yalcin, "Flood prediction using time series data mining," *Journal of Hydrology*, vol. 333, no. 2–4, pp. 305–316, 2007.

[12] K.-W. Chau and N. Muttil, "Data mining and multivariate statistical analysis for ecological system in coastal waters," *Journal of Hydroinformatics*, vol. 9, no. 4, pp. 305–317, 2007.

[13] E. P. Roz, *Water quality modeling and rainfall estimation: a data driven approach [M.S. thesis]*, University of Iowa, Iowa City, Iowa, USA, 2011.

[14] M. E. Keskin, Ö. Terzi, and E. U. Küçüksille, "Data mining process for integrated evaporation model," *Journal of Irrigation and Drainage Engineering*, vol. 135, no. 1, pp. 39–43, 2009.

[15] Ö. Terzi, "Monthly river flow forecasting by data mining process," in *Knowledge-Oriented Applications in Data Mining*, K. Funatsu, Ed., InTech, Rijeka, Croatia, 2011.

[16] Ö. Terzi, E. U. Küçüksille, G. Ergin, and A. İlker, "Estimation of solar radiation using data mining process," *SDU International Technologic Science*, vol. 3, no. 2, pp. 29–37, 2011 (Turkish).

[17] R. S. V. Teegavarapu, "Estimation of missing precipitation records integrating surface interpolation techniques and spatio-temporal association rules," *Journal of Hydroinformatics*, vol. 11, no. 2, pp. 133–146, 2009.

[18] D. P. Solomatine and K. N. Dulal, "Model trees as an alternative to neural networks in rainfall-runoff modelling," *Hydrological Sciences Journal*, vol. 48, no. 3, pp. 399–412, 2003.

[19] M. E. Keskin, D. Taylan, and E. U. Kucuksille, "Data mining process for modeling hydrological time series," *Hydrology Research*. In press.

[20] E. Simoudis, "Reality cheek for data mining," *IEEE Expert-Intelligent Systems and their Applications*, vol. 11, no. 5, pp. 26–33, 1996.

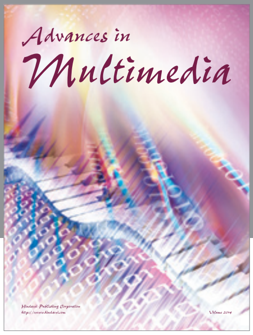
[21] http://www.dataminingtechniques.net/data-mining-tutorial/what-is-data-mining/.

[22] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 375–381, 2003.

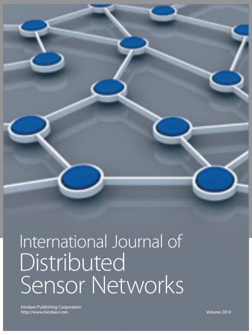
[23] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, "Model evaluation guidelines for systematic quantification of accuracy in watershed simulations," *Transactions of the ASABE*, vol. 50, no. 3, pp. 885–900, 2007.

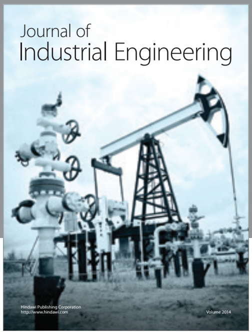
[24] J. Piri, S. Amin, A. Moghaddamnia, A. Keshavarz, D. Han, and R. Remesan, "Daily pan evaporation modeling in a hot and dry climate," *Journal of Hydrologic Engineering*, vol. 14, no. 8, pp. 803–811, 2009.

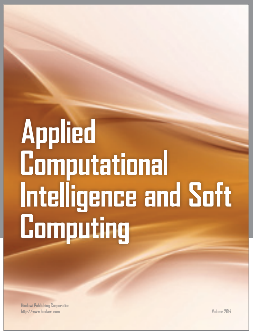
[25] S. Lallahem and J. Mania, "A nonlinear rainfall-runoff model using neural network technique: example in fractured porous media," *Mathematical and Computer Modelling*, vol. 37, no. 9-10, pp. 1047–1061, 2003.

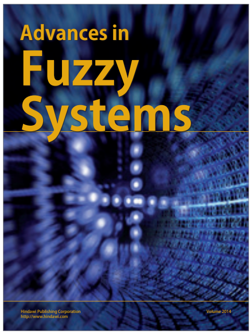

Advances in
Multimedia

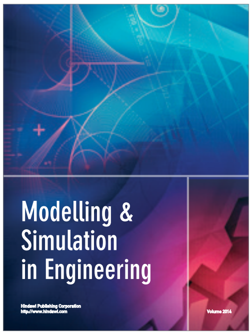
The Scientific
World Journal

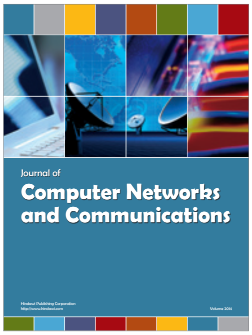

International Journal of
Distributed
Sensor Networks

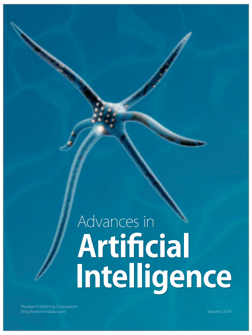

Journal of
Industrial Engineering

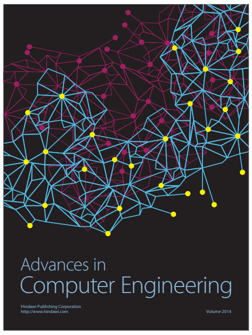

Applied
Computational
Intelligence and Soft
Computing


Advances in
Fuzzy
Systems

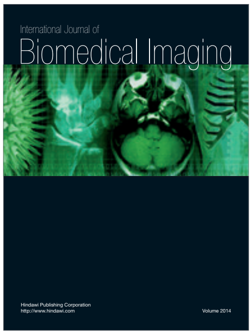

Modelling &
Simulation
in Engineering

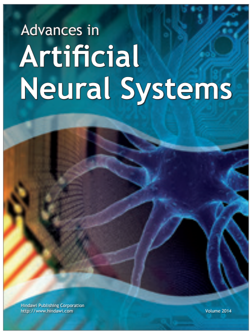

Journal of
Computer Networks
and Communications

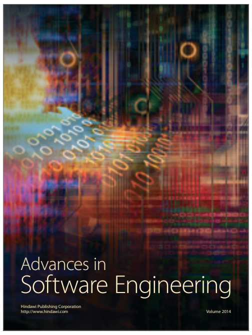

Advances in
Artificial
Intelligence


Advances in
Computer Engineering

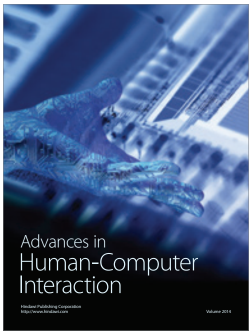
International Journal of
Computer Games
Technology

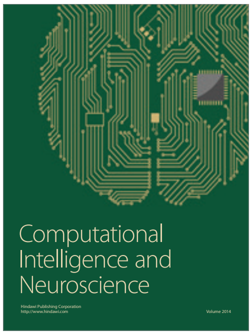

International Journal of
Biomedical Imaging


Advances in
Artificial
Neural Systems


Advances in
Software Engineering

Journal of
Robotics


Advances in
Human-Computer
Interaction


Computational
Intelligence and
Neuroscience

International Journal of
Reconfigurable
Computing

Journal of
Electrical and Computer
Engineering