Mustafa Kemal Üniversitesi Tarım Bilimleri Dergisi 24 (Özel Sayı):149-154, 2019
(*Mustafa Kemal University Journal of Agricultural Sciences* 24 (Special Issue):149-154, 2019)
e-ISSN: 2667-7733
http://dergipark.org.tr/mkutbd

1st International Congress on Biosystems Engineering 2019 ARAŞTIRMA MAKALESİ/RESEARCH ARTICLE

Estimation of monthly precipitation based on machine learning methods by using meteorological variables

Fatemeh Shaker SUREH[1], Mohammad Taghi SATTARI[1], Ahmet IRVEM[2]

[1]*University of Tabriz, Faculty of Agriculture, Department of Water Engineering, Tabriz, Iran.*
[2]*Hatay Mustafa Kemal University, Faculty of Agriculture, Department of Biosytems Engineering, Hatay, Turkey*

MAKALE BİLGİSİ / ARTICLE INFO

Makale tarihçesi / Article history:
Geliş tarihi /Received:09.10.2019
Kabul tarihi/Accepted:16.12.2019

Keywords:
Data mining, nearest neighbours, precipitation modelling, support vector regression, Chabahar.

Corresponding author: M. Taghi SATTARI
✉: mtsattar@gmail.com

ÖZET / ABSTRACT

Aims: The aim of this study is to estimate monthly precipitation by support vector regression and the nearest neighbourhood methods using meteorological variables data of Chabahar station.

Methods and Results: Monthly precipitation was modelled by using two support vector regression and the nearest neighbourhood methods based on the two proposed input combinations.

Conclusions: The results showed that the support vector regression method using normalized polynomial kernel function has higher accuracy and it has lower estimation error than the nearest neighbour method.

Significance and Impact of the Study: Precipitation is one of the most important parts of the water cycle and plays an important role in assessing the climatic characteristics of each region. Modelling of monthly precipitation values for a variety of purposes, such as flood and sediment control, runoff, sediment, irrigation planning, and river basin management, is very important. The modelling of precipitation in each region requires the existence of accurately measured historical data such as humidity, temperature, wind speed, etc. Limitations such as insufficient knowledge of precipitation on spatial and temporal scales as well as the complexity of the relationship between precipitation-related climatic parameters make it impossible to estimate precipitation using conventional inaccurate and unreliable methods.

Atıf / Citation: Sureh FM, Sattari MT, Irvem A (2019) Estimation of monthly precipitation based on machine learning methods by using meteorological variables. *MKU. Tar. Bil. Derg.* 24 (Özel Sayı) :149-154

INTRODUCTION

Precipitation plays an important role in the global water and energy cycle. More than 54% of the world's population lives in areas where the water crisis is serious (Boston et al., 2012). Iran is located in a dry belt, and its average rainfall is only equivalent to one-fifth of the global average (Anonymous, 2016). Naturally, it is quite critical to know the factors affecting rainfall patterns in such areas.

Meteorological parameters are one of the most important factors affecting precipitation. They have a significant impact on agriculture and society. Therefore, precipitation prediction is very important, especially in developing countries that have large populations and economies that rely on agriculture. During the process of plant growth, drought or heavy precipitation can cause a significant reduction in the yield of the product. Accurate prediction of precipitation helps inform effective decisions for proper planning and sustainable water resources management, flood control, environmental protection, tourism development, and urban development planning.

However, precipitation prediction is difficult as it exhibits large spatial and temporal variability. The intrinsic nonlinear and intricate patterns of precipitation have made it an appealing application for simulation (Asghari and Nasseri, 2015).

Ingsrisawang et al. (2008), used Support Vector Machine, Decision Trees, and Artificial Neural Networks for predicting short-term precipitation and developed classification and prediction models in northern Thailand. Data were collected during the years 2004 to 2016.

Zaw (2008), developed a prediction model for the precipitation of Myanmar using Multiple Linear Regressions. There, 15 predictors were used; the results of the study indicated that the predicted precipitation values closely matched the actual values.

Wang and Sheng (2010), used a Generalized Regression Neural Network model for predicting annual precipitation in ZhengZhou. They showed that the Generalized Regression Neural Network has advantages for predicting precipitation compared to Back Propagation Neural Networks and regression analysis methods. The results of the Generalized Regression Neural Network simulation for annual precipitation are better than the Back-Propagation Neural Network.

Sivaramakrishnan and Meganathan (2012), tried to predict the precipitation at the Trichirappalli station in southern India using association rule mining. Data were first filtered using a discretization approach based on the best fit range and then predicted using the Apriori algorithm. Finally, the data were verified using the Kstar classifier approach. The results showed that the overall classification accuracy for the occurrence of rainfall on dry and wet days was satisfactorily using the data mining technique.

Olaiya and Adeyemo (2012), predicted maximum temperature, precipitation, evaporation, and wind speed using data mining techniques. This study was conducted using an Artificial Neural Network, Decision Tree algorithms, and meteorological data between 2000 and 2009 from the city of Dibadan, Nigeria. There, a data model was developed to train the classifier algorithms. The performances of these algorithms were compared using standard performance metrics and algorithms that obtained the best results were used to generate classification rules for the mean weather variables. The results showed that, with enough data, the data mining technique could be used to predict the weather.

Dutta and Tahbilder (2014), predicted precipitation using Multiple Linear Regressions with six-years of meteorological data for Guwahati, India. Maximum and minimum temperatures, humidity, pressure and sea level were used as predictive parameters. The results showed that the prediction model based on Multiple Linear Regression achieved 63% accuracy in prediction of precipitation.

Sethi and Garg (2014), used Multiple Linear Regression to predict precipitation using a 30-year dataset from 1973-2002. The climate data included precipitation, vapor pressure, mean temperature and cloud cover of Udaipur city in India. Experimental results showed a close agreement between the predicted and actual values.

Ji et al. (2012), predicted precipitation using CART (Classification and Regression Tree) and C4.5 algorithms. The study consisted of two steps: (1) determining the chance of precipitation; (2) predicting hourly precipitation when there is a chance of precipitation. Input parameters included wind gusts, wind speed, outdoor humidity, outdoor temperature, evaporation, solar radiation, wind chill, dew point, pressure, cloud base, air density, and vapor pressure. The results showed both methods have high accuracy and are efficient algorithms.

Bushara (2016), predicted precipitation in Sudan using monthly meteorological data and Computational Intelligence (Gaussian Processes, Linear Regression, Multilayer Perceptron, IBK, KStar, Decision Table, M5 Rules, M5P, REPTree, Additive Regression, Bagging, MultiScheme, Ensemble Methodology, and ANFIS). The proposed models were evaluated and compared using correlation coefficients and mean absolute errors. The results indicate that the ANFIS approach gave satisfactory performance.

In the present study, a site with low water resources was selected (the city of Chabahar). The precipitation in the city of Chabahar is low on average compared with the other cities in Iran. Annual precipitation is primarily a result of several sudden events. It also has a high temperature with low volatility, large solar insolation, and poor vegetation.

Significant year-to-year variation in precipitation is the main characteristics of precipitation. An alternative approach is to divide the response space into sub-sections and smaller units as used in regression models based on data mining. Predictions are made faster by using data mining regression and allow a user to identify important parameters.

Consequently, this study aims to determine the accuracy of different models (Support Vector Regression, Nearest Neighborhood Model) for modeling the monthly precipitation of the Chabahar Synoptic Station in the southeastern region of Iran by using meteorological parameters such as maximum temperature, mean temperature, and wind speed.

MATERIAL and METHODS

Area of Study

The port of Chabahar is located on the Makran coast of Sistan and Baluchistan Province, next to the Gulf of Oman and at the mouth of Strait of Hormuz. It is the only Iranian port with direct access to the Indian Ocean. Being close to Afghanistan and the Central Asian countries of Turkmenistan, Uzbekistan, etc., it has been termed the "Golden Gate" to these land-locked countries. Chabahar Port is located at an area of 11 km^2 at an elevation of 7 meters above sea level.

Chabahar located in 25°17′ latitudes and 60°37′ longitudes in most years around 100 millimeters (3.9 in) will fall; however, a positive Indian Ocean Dipole in 1997/1998 led to a record total of 470 millimeters (18.5 in); in contrast between July 2000 and June 2002 only 57.5 millimeters (2.3 in) fell in two years. (Anonymous, 2012). Fig. 1 shows the studied station.

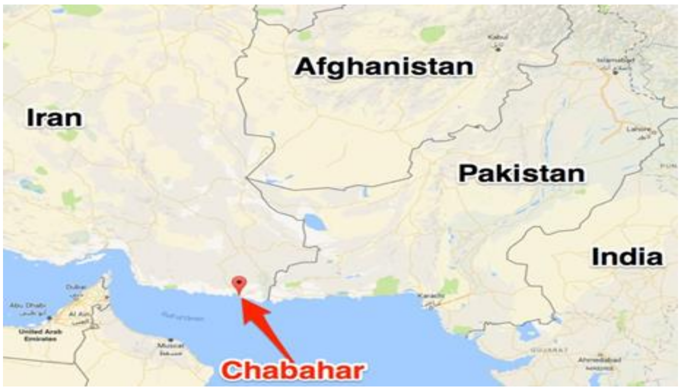

Fig. 1. Location of Chabahar station in Iran

Data Used

In this study, monthly data including precipitation, mean temperature and humidity and direction and wind speed of Chabahar station between 1986 and 2017 were used. Of the total data, 80% of the data were used for model training and the remaining 20% for testing. The statistical data of the data used in this study are presented in Table 1 and the precipitation chart of Chabahar Meteorological Station is shown in Fig. 2.

Table 1. Total heating value of agricultural residues of Hatay

Variable	Minimum	Maximum	Mean	Std. deviation
T_{max} (°C)	22.00	34.90	29.43	3.16
T_{min} (°C)	14.40	30.30	23.24	4.48
T_{mean} (°C)	18.70	32.20	26.32	3.76
RH (%)	38.00	93.00	74.37	9.21
Wind speed (m/s)	5.00	22.00	9.80	2.45
P (mm)	0.00	198.60	9.66	25.47

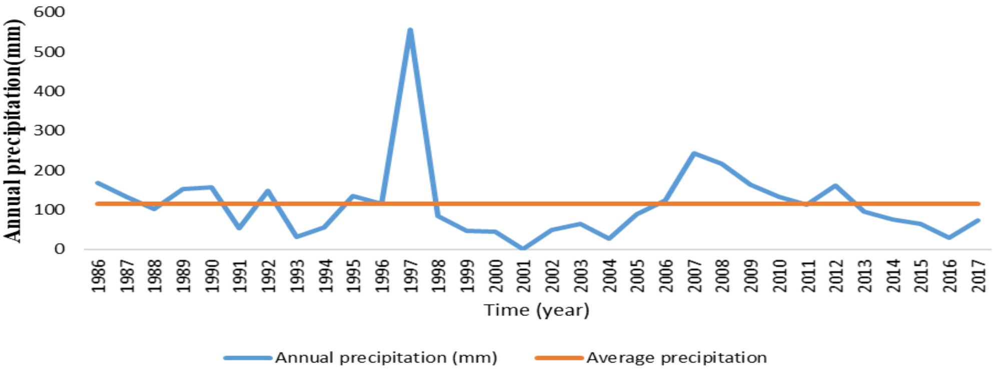

Fig. 2. Chabahar meteorological station precipitation chart

Support Vector Regression (SVR) Model

Support Vector Machine is one of the supervised learning methods introduced in 1992 by Vapnik and Chervonenkis based on the theory of statistical learning. Regression means to obtain a superscript that fits the desired data. The distance of each point from this super file indicates the error of that particular point (Shahrabi and Shojaee, 2011). The best method suggested for linear regression is the least-squares method. However, for regression cases, it is possible that the use of the least-squares estimator in the presence of outliers may not be complete. As a result, the regressor performs poorly. Therefore, a robust estimator should be developed that is not sensitive to small changes in the model. The kernel functions are used to solve the problem of very high dimensions and convert the nonlinear form to linear. In this way, different kernel functions such as polynomial kernels, normalized polynomials, and Pearson can be used. Therefore, it is sufficient to use the kernel of input values instead of the function itself in nonlinear problems.

Nearest Neighborhood Model

Nearest neighborhood model technique is based on the concept of similarity. The results of memory-based reasoning are based on similar situations that have happened in the past. This means that the algorithm stores all the available items and makes numerical predictions based on the similarity measurements. In general, steps can be taken to improve the performance of the nearest neighbor models, including selecting a method for estimating the best neighbors, developing information transfer functions, and developing distance functions. To estimate the best neighborhoods in the nearest neighbor method, different methods have been proposed which can be used depending on the accuracy and the complexity and volume of the problem. One of these methods is the use of empirical relation $K = \sqrt{n}$ where n is the time series and k is the best number of neighbors used in this method. The efficiency of this relationship increases with the increasing length of time series. Trial-Error is another method that can be used to find the best neighbors by selecting different neighborhoods in the best neighborhood and extracting prediction errors. Another function that plays a key role in the performance of the nearest-neighbor method is the distance function, including Euclidean, Minkowski, Manhattan, and Chebyshev (Sharif and Burn, 2006).

In this study, we used WEKA software developed by the Waikato University of New Zealand to model the precipitation of the Chabahar station using support vector regression and nearest neighbor regression methods.

Feature Selection by Relief Algorithm

Relief method uses a statistical solution to select a feature, as well as a weight-based method inspired by sample-based algorithms. One of the most important features and prominences of the Relief algorithm is its suitability for use in a dataset with a small number of educational samples. Therefore, in this research, due to the length of the statistical data, this method was also used to determine the effective parameters for drought modeling. It should be noted that this test was carried out in Weka software.

Feature Selection Based on Correlation (CFS)

Correlation feature selection is a commonly used method for selecting input variables and reducing problem dimensions and is introduced by Hall (1999). The correlation method gives the subsets that have the characteristics with the highest correlation coefficient with the sample class, and the variables with the highest score are considered as the main variable. This algorithm has a high ability to quickly detect unrelated, unnecessary and error-free data, which generally results in the removal of half the data. This feature increases the productivity of the models by reducing the dimensions of the problem.

Model Evaluation Criteria

The purpose of the evaluation criteria is to estimate the error rate and determine the patterns and the structures that have the least amount of rainfall forecasting error. In this study, three evaluation criteria were used; correlation coefficient (R) (Eq. 1), root mean square error (RMSE) (Eq. 2) and mean absolute error (MAE) (Eq. 3).

$$R=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad \text{Eq. (1)}$$

$$RMSE=\sqrt{\frac{\sum_{i=0}^{n}(y_i-x_i)^2}{N}} \qquad \text{Eq. (2)}$$

$$MAE=\frac{1}{n}\sum_{i=0}^{n}\left|x_i-y_i\right| \qquad \text{Eq. (3)}$$

In these relations, xi (yi) are observational (computational) values. The higher (lower) the correlation coefficient values are, (RMSE and MAE) the more accurate the model is.

RESULTS and DISCUSSION

Combinations of input parameters and their combined ability to modeling precipitation are identified. Table 2 shows the input parameters selected for modeling precipitation based on the RELIEF and CFS algorithms.

Table 2. Input parameters determined by feature selection methods

Algorithm	**Variables**
Cfs	T_{max},T_{mean}, Wind speed
Relief	T_{max}, RH, Wind direction

Investigation of Precipitation Modeling Results

Due to the importance of selecting such cases as kernel function in support vector regression and distance function in nearest neighbor method, the performance of these models was first evaluated for different functions using the default values of the parameters of each function to find the optimal function for each. The method of selection is then modeled for precipitation with each binary combination based on the optimal functions. For this purpose, 80% of the data was used for model training and 20% was used for model testing. The results of the support vector regression and nearest neighbor regression methods using different functions in default mode for the structure of each function are presented in Table 3.

Based on the results presented in Table 3, it can be seen that the best results are obtained in the vector support regression method by using normalized polynomial function and in the nearest neighbor method the best results are obtained using Manhattan function. The nearest-neighbor method was used to determine the optimal neighborhood number to increase the accuracy of the model by trial and error method. However, the combination of the input parameters determined by the correlation method (CFS) is relatively more accurate and less error-free than the combination of the parameters specified by the Relief algorithm. To better understand the results obtained in this study and to visualize the accuracy of the models used in rainfall modeling, the distribution graphs of computational values of data mining methods compared to the observed values of precipitation for the best input parameter combination are shown in Fig. 3.

Table 3. Input parameters determined by feature selection methods

Method	Function used	Combination (Cfs)			Combination (Relief)		
		MAE	RMSE	R	MAE	RMSE	R
Support Vector Regression	Polynomials	7.02	15.08	0.75	7.03	15.32	0.74
	Normalized polynomials	6.73	14.58	0.78	7.51	15.45	0.71
	Pearson	7.38	14.98	0.76	9.04	18.10	0.56
Nearest Neighborhood	Euclidean	16.13	36.83	0.24	9.93	24.33	0.27
	Chebyshev	13.85	28.61	0.31	10.47	23.18	0.24
	Manhattan	14.76	34.66	0.32	10.20	24.27	0.28

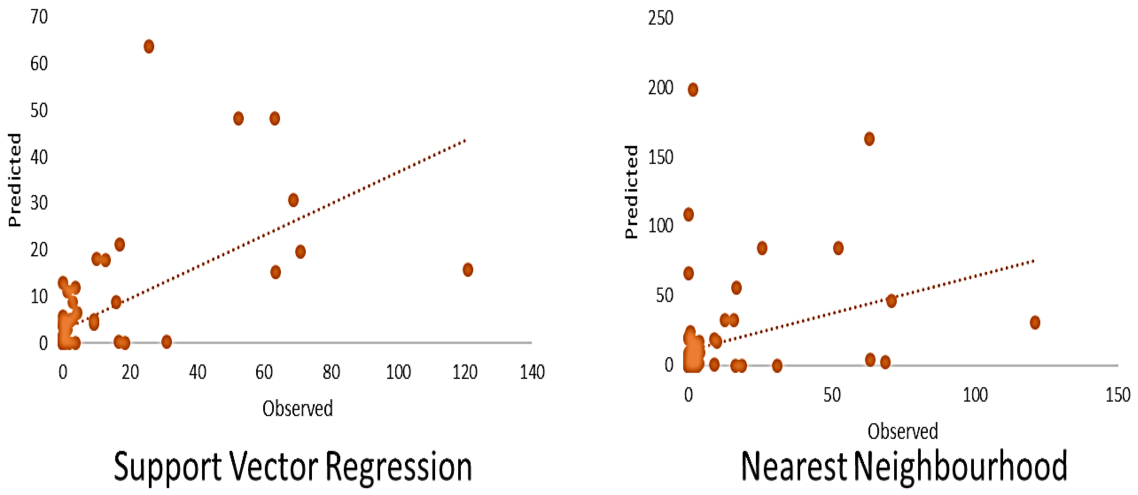

Fig. 3. Predicted and observed precipitation distribution graph.

CONCLUSIONS

Iran suffers from a shortage of precipitation and consequently a shortage of surface and groundwater resources. Important parts of Iran, especially the city of Chabahar, are facing a water crisis. In such circumstances, forecasting and modeling of precipitation values are necessary to provide proper planning in water resources management. In this study, the efficiency of the Support Vector Regression (SVR) and Nearest Neighbor models in the monthly precipitation modeling of the Chabahar Station was evaluated. In this study, monthly precipitation of Chabahar was modeled using different combinations as input variables. In general, the best performance of the model occurred when meteorological parameters such as maximum temperature, mean temperature, and wind speed was considered as inputs of the models. The results showed that the performance of the Support Vector Regression (SVR) model was better than that of the nearest neighbor model. The results also showed that the combination of parameters based on Relief algorithm in the modeling process with the negligible difference is very close competition with the selection of parameters based on correlation method and can be an efficient method for determining effective parameters in hydrological processes especially modeling.

Precipitation is generally an easily measured parameter. estimation models used are not satisfactory. This study should be concluded with emphasis on direct measurement of precipitation.

DECLARATION OF CONFLICTING INTERESTS

The authors declared no potential conflicts of interest for this study.

REFERENCES

Anonymous, (2012) Chaharmahal and Bakhtiari Meteorological Administration. Retrieved June 26, 2019, from https://www.chaharmahalmet.ir/ .

Anonymous, (2016) Water challenges of Iran. Retrieved June 16, 2019, from https://water.fanack.com/iran/water-challenges-of-iran/ .

Asghari K, Nasseri M (2014) Spatial rainfall prediction using optimal features selection approaches. Hydrol. Res. 46 (3): 343-355.

Bostan PA, Heuvelink GB, Akyurek SZ (2012) Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. Int. J. Appl. Earth. Obs. 19: 115-126.

Breiman L (2001) Random Forests. Mach. Learn. 45(1): 5-32.

Bushara NO (2016) Rainfall forecasting in Sudan using computational inteligence. PhD Thesis, Sudan University of Science and Technology, 177 p.

Eyvazi M, Mosaedi A (2012) An investigation on spatial pattern of annual precipitation in Golestan province by using deterministic and geostatisticcs model. J. Water Soil, 26 (1): 53-64.

Hall MA (1999) Correlation-based feature selection for machine learning. PhD Thesis, University of Waikato, Department of Computer Science, 178 p.

Ji SY, Sharma S, Yu B, Jeong DH (2012) Designing a rule-based hourly rainfall prediction model. IEEE 13th International Conference on Information Reuse & Integration (IRI). pp. 303-308.

Khandelwal N, Davey R (2012) Climatic assessment of Rajasthan's region for drought with concern of data mining techniques. Int. J. Eng. Res. App. 2 (5): 1695-1697.

Olaiya F, Adeyemo AB (2012) Application of data mining techniques in weather prediction and climate change studies. Int. J. Inf. Eng. Elec. Bus. 4 (1): 51.

Sethi N, Garg K (2014) Exploiting data mining technique for rainfall prediction. Int. J. of Comp. Sci. Information Tech. 5 (3): 3982-3984.

Sharif MH, Burn D (2006) Simulating climate change scenarios using an improved K-nearest neighbor model. J. Hydrol. 325: 179-196.

Sivaramakrishnan TR, Meganathan S (2012) Point rainfall prediction using data mining technique. Res. J. Ap. Sci. Eng. Tech. 4 (13): 1899-1902.

Wang ZL, Sheng HH (2010) Rainfall prediction using generalized regression neural network: case study Zhengzhou. International conference on computational and information sciences. pp. 1265-1268.

Zaw WT, Naing TT (2008) Empirical statistical modeling of rainfall prediction over Myanmar. W. Acad. Sci., Eng. Techn. 2 (10): 500-504.