Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 7 (4): 593-596, 2007
ISSN 1812-5654

Data Mining Approach for Estimation Evaporation from Free Water Surface

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

Abstract: Evaporation is a fundamental parameter in the cycle of hydrology. In the present study, data mining method is used to developed evaporation models. Before modeling, air temperature, water temperature, solar radiation and relative humidity parameters are selected as parameters affecting evaporation. Decision Table, KStar, M5P, Pace Regression, M5'Rules, Neural Network, Regression, Simple Linear Regression and SMO Regression algorithms are used for modeling. Finally, the developed models are compared with measured daily pan evaporation values and Penman method. The comparisons show that there is a good agreement between results of M5P model and measured daily pan evaporation values.

Key words: Pan evaporation, data mining, penman method

INTRODUCTION

Evaporation is one of the fundamental elements in the hydrological cycle, which affects the yield of river basins, the capacity of reservoirs, the consumptive use of water by crops and the yield of underground supplies. In general, there are two approaches in the evaporation estimation, namely, direct and indirect. Indirect methods of estimation based on meteorological data are in use by many researchers for free surface water bodies. Stewart and Rouse (1976) determined the summer-time evaporation from shallow lakes using the energy budget and equilibrium models. They showed that the actual evaporation could be determined within 10% over periods of two weeks using these models. de Bruin (1978) used the simplified model by combining Priestley-Taylor and Penman equations to estimate evaporation. He indicated that the model would produce good results for periods of 10 days or more. Morton (1979) modified a model to estimate annual evaporation from lake-based monthly observations of temperature, humidity and sunshine duration. The results of the model were compared with those of the water budget for lakes and showed that there is a good agreement between the suggested and water budget models. Singh and Xu (1997) evaluated and compared 13 evaporation equations that belonged to the category of the mass transfer method and a generalized model form for that category was also developed. On the other hand, the sole direct method is the U.S. Weather Bureau Class A pan measurement, which gives records of evaporation amount by time. The direct methods are used and compared in evaporation estimation studies by various researchers (Choudhury, 1999; McKenzie and Craig, 2001; Vallet-Coulomb *et al*., 2001; Abtew, 2001). Terzi and Keskin (2005) used Gene Expression Programming (GEP) to model Penman evaporation using air temperature, solar radiation and relative humidity parameters. They suggested that GEP approach is an alternative model of Penman method.

Knowledge discovery uses data mining and machine learning techniques that have evolved through a synergy in artificial intelligence, computer science, statistics and other related fields. Although there are technical differences, the terms machine learning, data mining and Knowledge Discovery and Data mining (KDD) are often used interchangeably (Goodwin *et al*., 2003).

Data Mining is often defined as the process of extracting valid, previously unknown, comprehensible information from large databases in order to improve and optimize business decisions (Braha and Shmilovici, 2002). In others definition, data mining is defined as the identification of interesting structure in data, where structure designates patterns, statistical or predictive models of the data and relationships among parts of the data (Fayyad and Uthurusamy, 2002). Data mining is applied in a wide variety of fields for prediction, e.g., stock-prices, customer behavior, production control. In addition, data mining has also been applied to other types of scientific data such as bioinformatical, astronomical and medical data (Li and Li Shue, 2004).

Terzi *et al*. (2005) used data mining methods to model pan evaporation. They applied genetic algorithm to select dominant parameter affecting evaporation and determined air temperature, water temperature and relative humidity parameters. They developed models using M5 rules, Kstar and decision table methods and showed that Kstar model gives better results comparing with the other models.

The present study has three objectives (1) to develop suitable DM evaporation models using Decision Table, KStar, M5P, Pace Regression, M5'Rules, Neural Network, Linear Regression, Simple Linear Regression and SMO Regression algorithms to estimate daily pan evaporation

from air temperature, water temperature, solar radiation and relative humidity parameters selected using bestfirst method, (2) to compare the DM evaporation models to Penman model and (3) to evaluate the potential of DM evaporation models.

MATERIALS AND METHODS

Study region and data: Lake Egirdir (lat. 37.80° and 38.43°N, lon. 30.30° and 31.37°E) is a freshwater lake located in Lakes District of Turkey which is the second largest freshwater lake in the country with a surface area and volume as 470 km^2 and 4360 hm^3, respectively. It is being used as water supply and irrigation purposes. This Lake is of tectonic origin in the northern part of the Egirdir County. The altitude of the lake is about 916 m above mean sea level. Geographically, the lake lies on a 50 km stretch on the north-south direction. The distance between east and west shores is 3 km, at which the depth is around 1.8 m. The mean depth of the lake is 8 to 9 m and the deepest point is 15 m. In the southern part, the width of the lake reaches a maximum of 16 km.

Meteorological data for models were obtained from an Automated GroWeather Meteorological Station setup near Lake Egirdir on August, 2000. Meteorological parameters included air and water temperature, relative humidity, solar radiation, wind speed and air pressure were logged. Class A pan evaporation values used as output in the models are measured daily by 18th District Directorate of State Hydraulic Works. The data used to develop data mining models included 874 daily observations for 2000-03 years. Training dataset is consisted of years 2000-01-02. The trained models are used to run a set of test data for year 2003.

Data mining: Data mining process generally involves phases of data understanding, data preparation, modeling and evaluation (Li and Li Shue, 2004). It is a hybrid disciplinary that integrates technologies of databases, statistics, machine learning, signal processing and high performance computing. This rapidly emerging technology is motivated by the need for new techniques to help analyze, understand or even visualize the huge amounts of stored data gathered from business and scientific applications. The major data mining functions that are developed in commercial and research communities include summarization, association, classification, prediction and clustering (Zhou, 2003).

Data understanding starts with an initial data collection and proceeds with activities to get familiar with the data, to identify data quality problems and to discover first insights into the data. Data preparation covers all activities that construct the final dataset to be modeled from the initial raw data. The tasks of this phase may include data cleaning for removing noise and inconsistent data and data transformation for extracting the embedded features (Li and Li Shue, 2004). Successful mining of data relies on refining tools and techniques capable of rendering large quantities of data understandable and meaningful (Mattison). The modeling phase applies various modeling techniques, determines the optimal values for parameters in models and finds the one most suitable to meet the objectives. The evaluation phase evaluates the model found in the last stage to confirm its validity to fit the problem requirements. No matter which areas data mining is applied to, most of the efforts are directed toward the data preparation phase (Li and Li Shue, 2004).

A good relational database management system will form the core of the data repository and adequately reflect both the data structure and the process fow and the database design will anticipate the kind of analysis and data mining to be performed. The data repository should also support access to existing databases allowing retrieval of supporting information that can be used at various levels in the decision making process (Rupp and Wang, 2004).

Data mining is a powerful technique for extracting predictive information from large databases. The automated analysis offered by data mining goes beyond the retrospective analysis of data. Data mining tools can answer questions that are too time-consuming to resolve with methods based on first principles. In data mining, databases are searched for hidden patterns to reveal predictive information in patterns that are too complicated for human experts to identify (Hoffmann and Apostolakis, 2003).

EVAPORATION MODELS

Bestfirst method is applied to select dominant parameters affecting evaporation among parameters from the meteorological station. The dominating parameters affecting evaporation are selected air temperature, water temperature, solar radiation and relative humidity. The air pressure and wind speed parameters with the least effects are neglected in evaporation models. In order to estimate daily pan evaporation from Lake Egirdir, Decision Table, KStar, M5P, Pace Regression, Additive, M5'Rules, Neural Network, Linear Regression, Simple Linear Regression and SMO Regression algorithms are considered. The Penman method, which is presented by Penman (1948) as a theory and formulae for the estimation of evaporation from weather data, is considered for comparison in this study.

Various statistical parameters and a comparison of the results for training and testing sets are presented in Table 1. It is obvious that the Penman method and the developed models in this study indicate after their comparisons with the pan evaporation that the M5P

algorithm in particular gives better results however the neural network model does not give suitable results among the developed models according to the mean square error (MSE) and the coefficient of determination (R^2) parameters. The results of M5P model is plotted against measured daily pan evaporation for training and testing sets in Fig. 1.

In order to examine the performance of the M5P model, its results were plotted against the Penman method. Figure 2 shows that the Penman approach gives

Table 1: Descriptive statistics of the evaporation models and Penman method

	Training set (2000-2001-2002)						Testing set (2003)					
					Pan evaporation comparison						Pan evaporation comparison	
Models	Mean (mm day^{-1})	Std. Dev.	Skewness	Kurtosis	MSE	R^2	Mean (mm day^{-1})	Std. Dev.	Skewness	Kurtosis	MSE	R^2
Pan Evap.	5.344	2.826	0.615	0.084	-	-	5.344	2.842	0.104	-1.309	-	-
Penman	4.539	2.118	0.011	-1.077	3.096	0.610	4.318	1.940	-0.058	-1.003	3.30	0.590
Linear Re.	5.344	2.345	-0.248	-1.044	2.482	0.688	5.344	2.786	-0.413	-1.074	1.695	0.789
KStar	5.324	2.637	0.460	-0.596	0.233	0.971	5.324	2.421	0.236	-1.355	1.805	0.776
N.Network	6.332	2.177	0.418	-1.111	3.156	0.604	6.332	2.215	0.305	-1.275	3.330	0.586
Pace Reg.	5.344	2.345	-0.248	-1.044	2.482	0.688	5.344	2.786	-0.413	-1.074	1.695	0.789
S. L. Reg.	5.344	2.231	-0.210	-1.043	3.003	0.623	5.344	2.557	-0.512	-.820	2.085	0.741
SMO_Reg	5.199	2.248	-0.246	-1.058	2.523	0.683	5.199	2.692	-0.418	-1.071	1.679	0.794
Additive	5.344	2.421	0.495	-0.578	2.090	0.738	5.344	2.572	0.353	-1.081	2.298	0.714
M5P	5.345	2.465	0.462	-0.693	1.685	0.789	5.345	2.453	0.212	-1.456	1.540	0.809
M5'Rules	5.362	2.509	0.624	-0.031	1.618	0.797	5.362	2.421	0.129	-1.419	1.608	0.800
D. Table	5.344	2.438	0.376	-0.939	2.034	0.745	5.344	2.117	0.246	-1.322	3.073	0.618

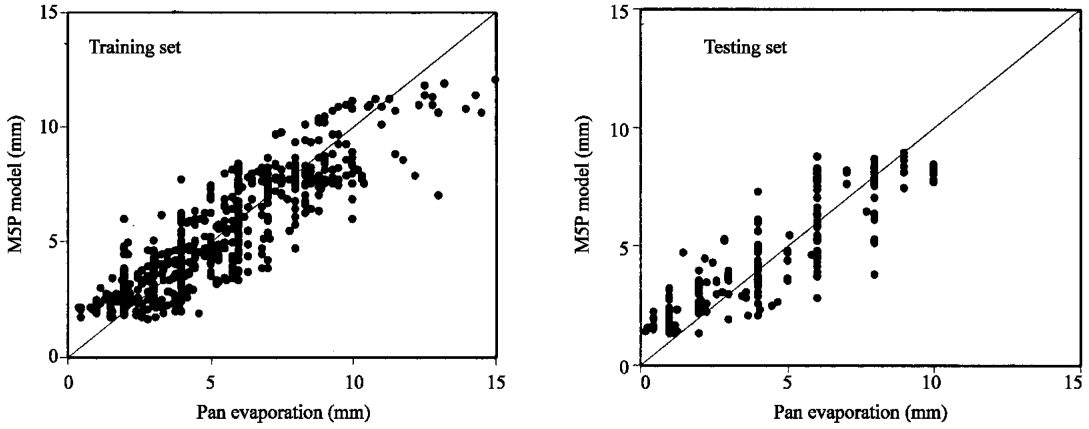

Fig. 1: Comparison daily pan evaporation with M5P model for training and testing sets

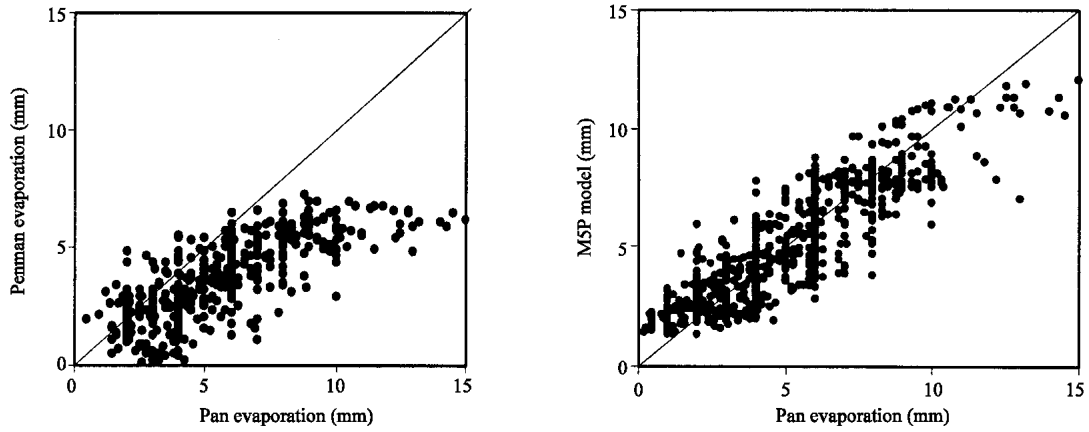

Fig. 2: Scatter diagrams between Penman method and M5P model *vs* daily pa

underestimated evaporation values, while the M5P model results lie around a 45° straight line implying, that there are no bias effects.

RESULTS

The aim of this research is to apply data mining (DM) method to estimate daily pan evaporation depending on meteorological data for Lake Eğirdir. Initially, the dominant factors affecting evaporation were determined as air temperature, water temperature, solar radiation and relative humidity. Then, alternative models were proposed to estimate evaporation using DM method (Decision Table, KStar, M5P, Pace Regression, M5 Rules, Neural Network, Linear Regression, Simple Linear Regression ve SMO Regression algorithms). The developed models are compared with the measured daily pan evaporation values and the Penman method. The comparisons show that there is a better agreement between the results of M5P model and pan evaporation values than others model. The evaporation could be estimated easily from available data using M5P algorithm. The model is also suitable for estimating missing daily pan evaporation values.

REFERENCES

Abtew, W., 2001. Evaporation estimation for Lake Okeechobee in South Florida. J. Irrig. Drainag. Eng., 127: 140-147.

Braha, D. and A. Shmilovici, 2002. Data mining for improving a cleaning process in the semiconductor industry. IEEE Transactions on Semiconductor Manufacturing, 15: 91-101.

Choudhury, B.J., 1999. Evaluation of an empirical equation for annual evaporation using field observations and results from a biophysical model. J. Hydrol., 216: 99-110.

de Bruin, H.A.R., 1978. A simple model for shallow lake evaporation. J. Applied Meteorol., 17: 1132-1134.

Fayyad, U.M. and R. Uthurusamy, 2002. Evolving data mining into solutions for insights. Commun. ACM, 45: 28-31.

Goodwin, L., M. VanDyne, S. Lin and S. Talbert, 2003. Data mining issues and opportunities for building nursing knowledge. J. Biomed. Inform., 36: 379-388.

Hoffmann, D. and J. Apostolakis, 2003. Crystal structure prediction by data mining. J. Mol. Struct., 647: 17-39.

Li, S.T. and L.Y. Li Shue, 2004. Data mining to aid policy making in air pollution management. Expert Syst. Applic., 27: 331-340.

Mattison, R. Data Warehousing: Strategies, Technologies and Techniques Statistical Analysis. SPSS Inc. White Papers.

McKenzie, R.S. and A.R. Craig, 2001. Evaluation of river losses from the Orange River using hydraulic modeling. J. Hydrol., 241: 62-69.

Morton, F.I., 1979. Climatological estimates of lake evaporation. Water Resour. Res., 15: 64-76.

Penman, H.L., 1948. Natural evaporation from open water, bare soil and grass. Proc. Roy. Soc. London, 193: 120-145.

Rupp, B. and J. Wang, 2004. Predictive models for protein crystallization. Methods, 34: 390-407.

Singh, V.P. and C-Y. Xu, 1997. Evaluation and generalization of 13 mass-transfer equations for determining free water evaporation. Hydrolog. Processes, 11: 311-323.

Stewart, R.B. and W.R. Rouse, 1976. A simple method for determining the evaporation from shallow lakes and ponds. Water Resour. Res., 12: 623-627.

Terzi, Ö. and M.E. Keskin, 2005. Evaporation estimation using gene expression programming. J. Applied Sci., 5: 508-512.

Terzi, Ö., E.U. Küçüksille and M.E. Keskin, 2005. Modeling of Daily Pan Evaporation Using Data Mining. Int. Symp. Innovations in Intelligent Syst. Applic., pp: 182-185.

Vallet-Coulomb, C., D. Legesse, F. Gasse, Y. Travi and T. Chernet, 2001. Lake evaporation estimates in tropical Africa (Lake Ziway, Ethiopia), J. Hydrol., 245: 1-18.

Zhou, Z.-H., 2003. Three perspectives of data mining. Artificial Intelligence, 143: 139-146.