Data mining process for modeling hydrological time series

M. Erol Keskin, Dilek Taylan and Ecir Ugur Kucuksille

ABSTRACT

The main purpose of this study was to develop an optimum flow prediction model, based on data mining process. The data mining process was applied to predict river flow of Seyhan Stream in the southern part of Turkey. Hydrological time series modeling was applied using monthly historical flow records to predict Seyhan Stream flows. Seyhan Stream flows were modeled by Markov models and it was seen that it adapted AR(2). Hence, F_{t-2} and F_{t-1} flows in (t–2) and (t–1) months were the taken inputs. For monthly streamflow predictions, data were taken from the General Directorate of Electrical Power Resources Survey and Development Administration. Used data covered 35 years between 1969 and 2003 for monthly streamflows. Furthermore, for the effect of monthly periodicity in hydrological time series cos ($2\pi_i/12$), sin ($2\pi_i/12$) ($l = 1, 2,\ldots, 12$) were included as inputs. Then, F_t flows in (t) months were modeled by data mining process. It was concluded that with using data mining process for streamflow prediction, it was possible to estimate missing or unmeasured data.

Key words | AR models, data mining process, flow prediction

M. Erol Keskin
Ecir Ugur Kucuksille
Faculty of Engineering-Architecture,
Suleyman Demirel University,
Isparta 32260,
Turkey

Dilek Taylan (corresponding author)
Faculty of Engineering,
Suleyman Demirel University,
Isparta 32260,
Turkey
E-mail: *dilektaylan@sdu.edu.tr*

INTRODUCTION

In the planning of water structures, future predictions based on past records are necessary for the assessment of design criteria. The identification of suitable generation models for future streamflows is an important precondition for successful planning and management of water resources.

Recently, the dominance of deterministic models in hydrology has gradually weakened as a number of factors have affected the constitution of hydrological events; therefore, the random nature of hydrological variables needs to be studied. Hipel (1985) showed that a simple stochastic approach gave better results than a more complex deterministic model. When available observation records are insufficient, the generating synthetic flow series can help a designer to carry out the analysis. Stochastic models, which reproduce the essential properties of the real process, are generally used for the generation of synthetic series and the prediction of future flows. Evaluating a large number of alternatives is necessary to reduce risk. For instance, Sert (1991) generated synthetic streamflows in order to obtain input for a simulation model aimed at operating the Keban–Karakaya–Atatürk Reservoir system and investigated risks resulting from the stochastic character of hydrological events. Generated series should maintain the same statistical characteristic of the historical series, such as the mean, standard deviation, skewness and autocorrelation coefficient. In this study, an autoregressive (AR) model is used for stochastic modeling of streamflow prediction.

Data mining (DM) is a hybrid technique that integrates technologies of databases, statistics, machine learning, signal processing, and high-performance computing. This 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 the commercial and research communities include summarization, association, classification, prediction and clustering (Zhou 2003).

A good relational database management system will form the core of the data repository, and adequately reflect both the data structure and the process flow; therefore, the database design would anticipate the kind of analysis and

doi: 10.2166/nh.2012.003

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 & Wang 2004).

DM is a powerful technique for extracting predictive information from large databases. The automated analysis offered by DM goes beyond the retrospective analysis of data. DM tools can answer questions that are too time-consuming to resolve with methods based on the 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 & Apostolakis 2003). DM is applied in a wide variety of fields for prediction, e.g. stock-prices, customer behavior, production control. In addition, DM has been applied to other types of scientific data such as bio-informatical, astronomical, and medical data (Li & Shue 2004).

This study has two steps: the first step was to develop a time series model using autoregressive model for Seyhan Stream and to find model degree; then, the second step was to form a model of monthly flow prediction of the stream using the DM process.

METHODS

Autoregressive modeling

The time series analysis and autoregressive models (AR) are used in many disciplines. In time series analysis, the relationship between the observed sample values (realization) and the underlying stochastic process is analogous to the relationship between the sample and the population in statistical hypothesis testing. Then, the time series are samples from the underlying stochastic process that have been generated from the series. The purpose of the time series analysis is to identify a model of an AR process by realization of a process.

The time series used in hydrological studies are annual, seasonal, monthly or weekly. For annual time series, the pattern is as follows:

$$Z_t = \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p} + \varepsilon_t \quad (1)$$

where t is the time (years), Z is a variable of interest, ϕ_p is autoregression coefficient; and ε_t is a residual part (Salas *et al.* 1980). The pattern can be random, seasonal, trend-type, periodic, etc., or a combination of these components (DeLurgio 1998).

When the intervals are shorter than 1 year (such as a month, a week, or a day) the series are called periodic. In such series, the statistical characteristics vary within the same year. In such cases, the time interval must be considered in the presentation of time. While time interval is shown as τ, the year is shown as ν. Thus the variables in Equation (1) change to $Z_{\nu,\tau}$ and $\varepsilon_{\nu,\tau}$, and Equation (1) becomes as follows:

$$Z_{\nu,\tau} = \phi_{1,\tau} Z_{\nu,\tau-1} + \ldots + \phi_{p,\tau} Z_{\nu,\tau-p} + \sigma_{\varepsilon,\tau} \xi_{\nu,\tau} \quad (2)$$

where $\phi_{j,\tau}$'s are the periodic autoregression coefficients, $\sigma_{\varepsilon,\tau}$ is the variance of residuals, and $\xi_{\nu,\tau}$ are standard normal random numbers.

The following steps are common to all empirical modeling and are as follows:

- model identification;
- parameter estimation;
- model diagnostics;
- forecast verification and reasonableness.

Model identification

Periodic historical series are controlled whether they are normal by using either skewness coefficient or Chi square test. If series are not normal, they are transformed into the normal distribution with an appropriate transformation function. Computing μ_τ periodic mean values and s_τ periodic standard deviations of periodic series, series are transformed into the standard form, and periodic condition is removed. Autocorrelation function (ACF) and partial autocorrelation function (PACF) belonging to these series are obtained and prior evaluation is made for degree of model.

Parameter estimation

ϕ_j autoregressive parameters are computed for the selected model. The stationarity condition is controlled. Variance of residuals is determined.

Model diagnostics

ε_t residuals are computed by using historical records. It is controlled whether residuals are independent or not. The Portmanteau test is used for this purpose. It is then determined whether residuals fit in normal distribution or not, the decision is made in respect of skewness coefficient or normal distribution chart. However, it is possible to relax the criteria about normal distribution (Salas *et al.* 1980). The Akaike Information Criterion (AIC) is used to investigate appropriateness of degree of the selected model. The best fitting model should have min AIC, and the ACF of the selected model should be consistent with the ACF of historical series.

Forecast verification and reasonableness

Synthetic series are generated and statistical characteristics of these series are compared to those of historical series. These are mean, standard deviation and ACF. Standard normal random numbers are easily obtained from uniform random numbers generated by computer.

DATA MINING (DM) PROCESS

A systematic approach is essential to obtain satisfactory results for the DM analysis. Nowadays, a number of versions of DM tools exist. The most widespread application amongst the tools is CRISP-DM. CRISP-DM (**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining) and is a data mining process model that describes commonly used approaches that expert data miners use to tackle problems. CRISP-DM was developed by a consortium which consists of NCR System Engineering (USA–Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringen en Bank Groep B.V. (The Netherlands) (Chapman *et al.* 2000; Fernandez *et al.* 2002). CRISP-DM is a process which defines the basic stages of DM, as can be seen in Figure 1.

Business understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective,

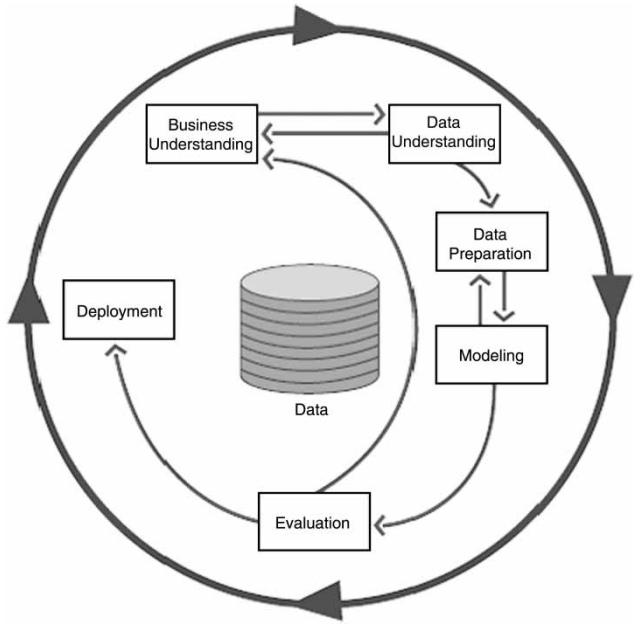

Figure 1 | CRISP-DM Data Mining Process (Chapman *et al.* 2000).

and then converting this knowledge into a data mining problem definition, and a preliminary project plan designed to achieve the objectives (Wirth & Hipp 2000; Fernandez *et al.* 2002).

Data understanding

The second phase of the DM is the understanding of the data. The process starts with collecting data. In applications, data are stored in different media. For example, Microsoft saves the data in hundreds of OLTP (Online Transaction Processing) databases and more than 70 data warehouses. The first step is to choose appropriate data from a database or data warehouse for the application (Tang & MacLennan 2005). Data quality is determined and first insights into the data are discovered in this step (Fernandez *et al.* 2002).

Data preparation

The data preparation phase covers all activities to construct the final dataset or the data that will be fed into the modeling tool(s) from the initial raw data. Tasks include table, record, and attribute selection, as well as transformation

and cleansing of data for modeling tools. The five steps in data preparation are the selection of data, the cleansing of data, the construction of data, the integration of data, and the formatting of data (Shearar 2000). The purpose of the cleansing of data is to pick inappropriate data within the database (Fernandez *et al*. 2002). The purpose of the integration of data is to transform the present data into different formats. For example, linguistic data can be converted to numerical data.

Modeling

In this phase, various modeling techniques are selected, applied and their parameters are calibrated to optimal values. Typically, several techniques exist for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase may be necessary. Modeling steps include the selection of the modeling technique, the generation of test design, the creation of models, and the assessment of models (Shearar 2000). If the purpose was exactly comprehended, then a true algorithm can be easily chosen. Each point consists of different algorithms and one cannot obtain the model which gives the best result without modeling (Tang & MacLennan 2005).

Evaluation

Before proceeding to final deployment of the model built by the data analyst, it is important to more thoroughly evaluate the model and review the model's construction to be certain it properly achieves the business objectives. It is critical to determine if some important business issue has not been sufficiently considered. At the end of this phase, the project leader should then decide exactly how to use the data mining results. The key steps here are the evaluation of results, the process review, and the determination of next steps (Shearar 2000). Different tools can be utilized for this purpose. For example, if numerical data exist and if one requires checking the accuracy of the model, MAPE (Mean Absolute Percentage Error) or R^2 (Coefficient of Determination) can be used (Fernandez *et al*. 2002).

Deployment

Model creation is generally not the end of the project. The knowledge gained must be organized and presented in a way that the customer can use it, which often involves applying 'live' models within an organization's decision-making processes, such as the real-time personalization of Web pages or repeated scoring of marketing databases (Shearar 2000). Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. Even though it is often the customer, not the data analyst, who carries out the deployment steps, it is important for the customer to understand up front what actions must be taken in order to actually make use of the created models. The key steps here are plan deployment, plan monitoring and maintenance, the production of the final report, and review of the project (Shearar 2000).

APPLICATION

AR modeling

Model identification

Transformation function (log $(Z_{v,\tau} - 7)$) was applied to monthly flows which did not fit normal distribution in respect of skewness coefficient. The historical $Y_{v,\tau}$ obtained is given in Figure 2.

Afterwards, μ_τ periodic means and s_τ standard deviations of periodic series were determined. The standard series were then obtained by removal of these characteristics. ACF and PACF for $Z_{v,\tau}$, which were non-periodic series, are given in Figures 3 and 4, respectively, within a 95% confidence interval. It appeared that Z_t was a dependent series in respect of ACF.

Parameter estimation

ACFs of AR(1), AR(2) and AR(3) were constituted for model degree and compared with ACF of Z_t.

For AR(1); $\phi_1 = r_1 = 0{,}740$

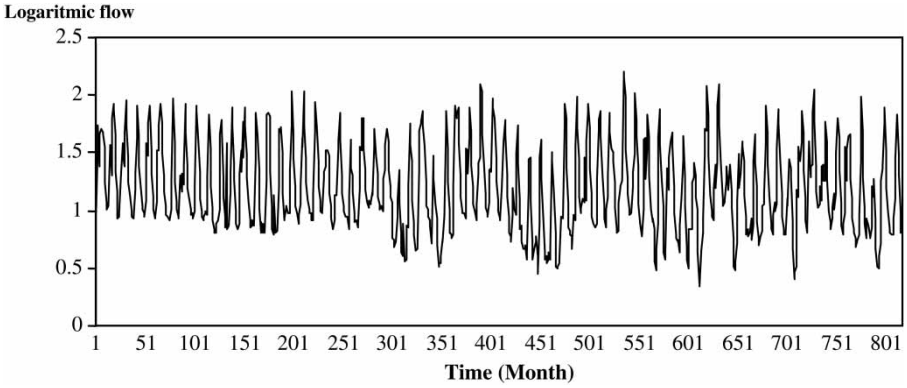

Figure 2 | Monthly transformed flow values.

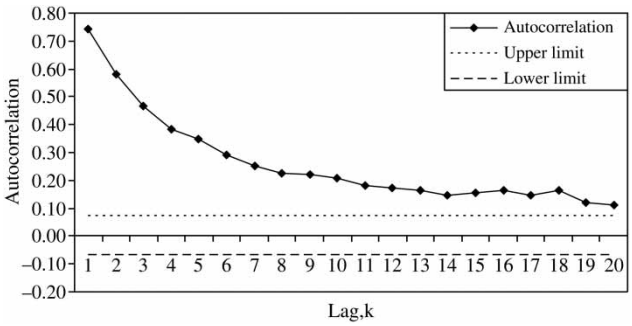

Figure 3 | ACF of Z_t and 95% confidence intervals.

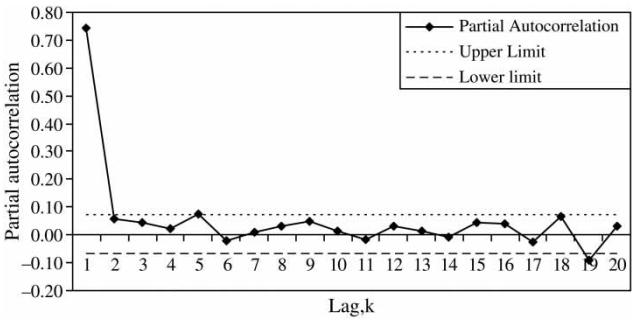

Figure 4 | PACF of Z_t and 95% confidence intervals.

For AR(2); $\phi_1 = 0{,}687$ $\phi_2 = 0{,}072$

For AR(3); $\phi_1 = 0{,}684$ $\phi_2 = 0{,}043$ $\phi_3 = 0{,}041$ and ACFs are shown in Figure 5.

It could be concluded that AR(2) process was the most appropriate. It was shown that AR(2) provided stationarity condition as follows:

$$\phi_1 + \phi_2 = 0,687 + 0,072 = 0,759 < 1$$

$$\phi_2 - \phi_1 = 0,072 - 0,687 = -0,615 < 1$$

$$-1 < \phi_1 < 1 \text{ and } -1 < \phi_2 < 1$$

Residual variance $\sigma_\varepsilon{}^2 = 0{,}431$ was obtained.

Model diagnostics

ε_t residual series of AR(2) were tested with the Portmanteau in order to investigate whether it was dependent or not. According to Portmanteau, it was independent and fits normal distribution. Then, AIC test was applied to AR(1), AR(2) and AR(3); and AR(2) was selected.

Forecast verification and reasonableness

Fifty synthetic series generated for AR(2), firstly, were changed from standard condition into normal condition and then inverted transformation function was applied to them. Thus original series were obtained. After calculations of ACF, mean and standard deviation, the same characteristics were computed for synthetic series within a 95% confidence interval. The conclusions are shown in Figures 6–8.

It was recorded that historical ACF, mean and standard deviations were within the confidence interval. In this case, the model preserved the characteristics of the historical series.

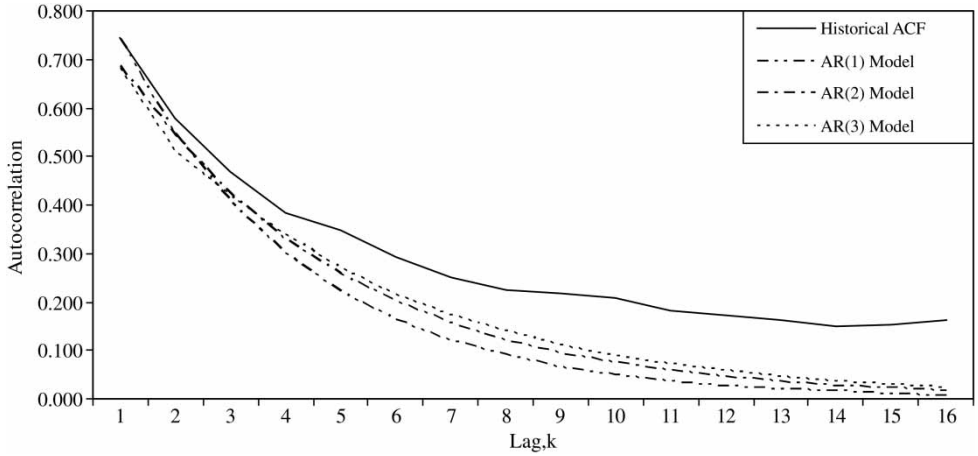

Figure 5 | Harmony between historical ACF and ACFs of AR(1) and AR(2).

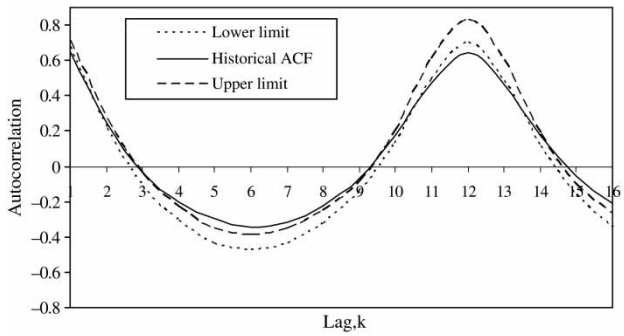

Figure 6 | Historical ACF and 95% confidence interval.

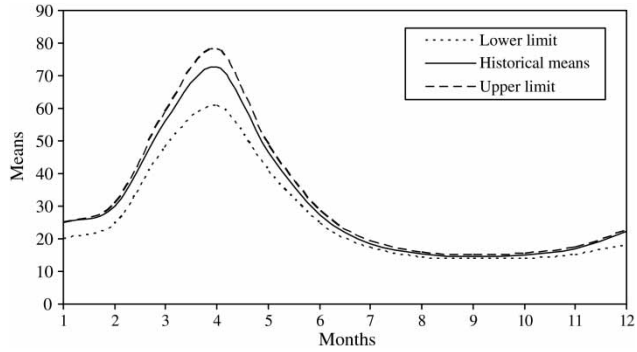

Figure 7 | Historical means and 95% confidence limit.

DATA MINING MODELS

Data understanding

A stochastic model was set up by an AR model for data from the 18-01 station on the Seyhan Stream in the east Mediterranean part of Turkey. Monthly flow data were used, covering the time span of 68 years (816 months), i.e. the observation period between 1936 and 2003. It was observed that the AR(2) process of Markov models provided stationarity conditions. Hence, it was shown that autocorrelation coefficients of AR(2) model were independent of time. The residual variance was obtained as $\sigma_\varepsilon{}^2 = 0.431$. The residual series of AR(2) ε_t were tested using the Portmanteau test in order to investigate whether or not it was dependent. According to this test, it was seen that residual series were

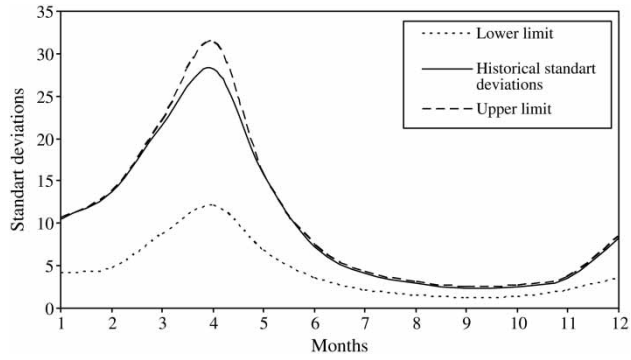

Figure 8 | Historical standard deviations and 95% confidence limit.

independent and fit in to normal distribution. Then, the AIC test was applied to AR(1), AR(2) and AR(3), which resulted in –666.07, –682.78 and –682.67, respectively; thus the AR(2) model was selected. Since AR(2) process

was appropriate, F_{t-2} and F_{t-1} flows in (t–2) and (t–1) months and cos ($2\pi i/12$), sin ($2\pi i/12$) ($i = 1, 2,\ldots, 12$) for the effect of monthly periodicity in hydrological time series were selected as input variables and F_t was output variable.

Data preparation

This study investigated whether or not there were any missing data. For substitution of missing data, the mean values were used. Hence, it was investigated if there were inaccurate data assuming normal distribution and outliers were deleted.

Modeling

In order to estimate monthly flow for Seyhan Stream Linear Regression, MultiLayerPerceptron, Pace Regression, KStar, SMOReg, M5P (M5 Model Tree Algorithm), REPTree and decision table algorithms were used in data mining process in Weka. Detailed explanations of these algorithms are given below.

Linear regression

Linear regression is a well known method of mathematical modeling of the relationship between a dependent variable and one or more independent variables. Regression uses existing (or known) values to forecast the required parameters. In the simplest case, regression employs standard statistical techniques such as linear regression. Unfortunately, many real world problems are not simply linear projections of previous values. For instance, sales volumes, stock prices, product failure rates and models of engineering systems are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g. logistic regression, decision trees or neural networks) may be necessary to forecast future values (Read 1999).

Multilayer perceptron

The most popular and powerful learning algorithms in neural networks are the back-propagation and its variants. This algorithm is based on the error-correction learning rule. Basically, the error back-propagation process consists of two passes through the different layers of the network: a forward pass and a backward pass. In the forward pass, an activity pattern (input vector) is applied to the sensory nodes of the network, and its effect propagates through the network, layer by layer. Finally, a set of outputs is produced as the actual response of network. During the forward pass, the synaptic weights of the network are all fixed. During the backward pass, on the other hand, the synaptic weights of the network are all adjusted in accordance with the error-correction rule. Specifically, the actual response of network is subtracted from a desired (target) response to produce an error signal. This error signal is then propagated backward through the network (Alam *et al.* 2009).

Pace regression

The basic idea of regression analysis is to fit a linear model to a set of data. The classical ordinary least squares estimator is simple, computationally cheap and has well established theoretical justification. Nevertheless, the models produced are often not satisfactory. Pace regression improves the classical ordinary least squares regression by evaluating the effect of each variable and using a clustering analysis to improve the statistical basis for estimating their contribution to the overall regressions. Under regularity conditions, pace regression is provably optimal when the number of coefficients tends to infinity. It consists of a group of estimators that are either overall optimal or optimal under certain conditions (Witten & Frank 2000; Panda *et al.* 2006).

KStar

K^* is an instance-based classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. The underlying assumption of instance-based classifiers is such as K^* (Wu *et al.* 2004).

SMOReg

The self-organizing map (SMO) algorithm for regression implements a non-linear method for sequential minimal optimization to train a support vector regression using

polynomial or radial basis function (RBF) kernels. Multi-class problems are solved using pairwise classification. To obtain the proper probability estimates, we use the option that fits logistic regression models to the outputs of the support vector machine (Chiu *et al.* 2008).

M5P (M5 Model Tree Algorithm)

This is an algorithm for generating M5 model trees (Gelly *et al.* 2006). M5 builds a tree to predict numeric values for a given instance. The algorithm requires the output attribute to be numeric while the input attributes can be either discrete or continuous. For a given instance, the tree is traversed from top to bottom until a leaf node is reached. At each node in the tree, a decision is made to follow a particular branch based on a test condition on the attribute associated with that node. Each leaf has a linear regression model associated with it of the form

$$w_0 + w_1 a_1 + \ldots\ldots\ldots + w_k a_k \tag{3}$$

based on some of the input attributes a_1, a_2, .., a_k in the instance respective weights w_0, w_1,….., w_k are calculated using standard regression. As the leaf nodes contain a linear regression model to obtain the predicted output, the tree is called a model tree (Qinlan 1992).

The M5P Model Tree algorithm in Weka is available in the Java class 'weka.classifiers.trees.M5P'. The two main parameters are described below:

1. **buildRegressionTree**: If True then the algorithm builds a regression tree rather than a model tree (Yuan *et al.* 2000).
2. **minNumInstances:** The minimum number of instances to allow at a leaf node (Yuan *et al.* 2000).

REPTree

Quinlan (1987) first introduced reduced error pruning (REP) as a method to prune decision trees. REP is a simple pruning method though it is sometimes considered to overprune the tree. A separate pruning dataset is required, which is considered a downfall of this method because data are normally scarce. However, REP can be extremely powerful when it is used with either a large number of examples or in combination with boosting. The used pruning method is the replacement of a subtree by a leaf representing the majority of all examples reaching it in the pruning set. This replacement is done if this modification reduces the error, i.e. if the new tree would give an equal or fewer number of misclassifications (Licamele & Getoor 2006).

Decision table

The decision table summarizes the dataset with a 'decision table', a decision table contains the same number of attributes as the original dataset, and a new data item is assigned a category by finding the line in the decision table that matches the non-class values of the data item. This implementation employs the wrapper method (Kohavi & John 1997) to find a good subset of attributes for inclusion in the table. By eliminating attributes that contribute little or nothing to a model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller, more condensed decision table (Cunnigham & Holmes 1999).

Evaluation

R^2 determination coefficients were taken into account to define the success rate of the formed models at previous stage based on the flow forecasting errors, as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(F_{i(\text{measured})} - F_{i(\text{model})}\right)^2}{\sum_{i=1}^{n} \left(F_{i(\text{measured})} - F_{\text{mean}}\right)^2} \tag{4}$$

where n = number of measured flow data; $F_{i(\text{measured})}$ and $F_{i(\text{model})}$ = measured flow value and the developed model flow estimations, respectively; and F_{mean} = monthly mean flow. Furthermore, the root mean square error (RMSE) is used to decide the best model. A strong relationship (with a 1-to-1) between R^2 and RMSE (both are based on the sum of square residuals) was observed. Thus, only R^2 was used as performance criteria.

R^2 coefficients for each model are given in Table 1. To examine the performance of each model, the results of the

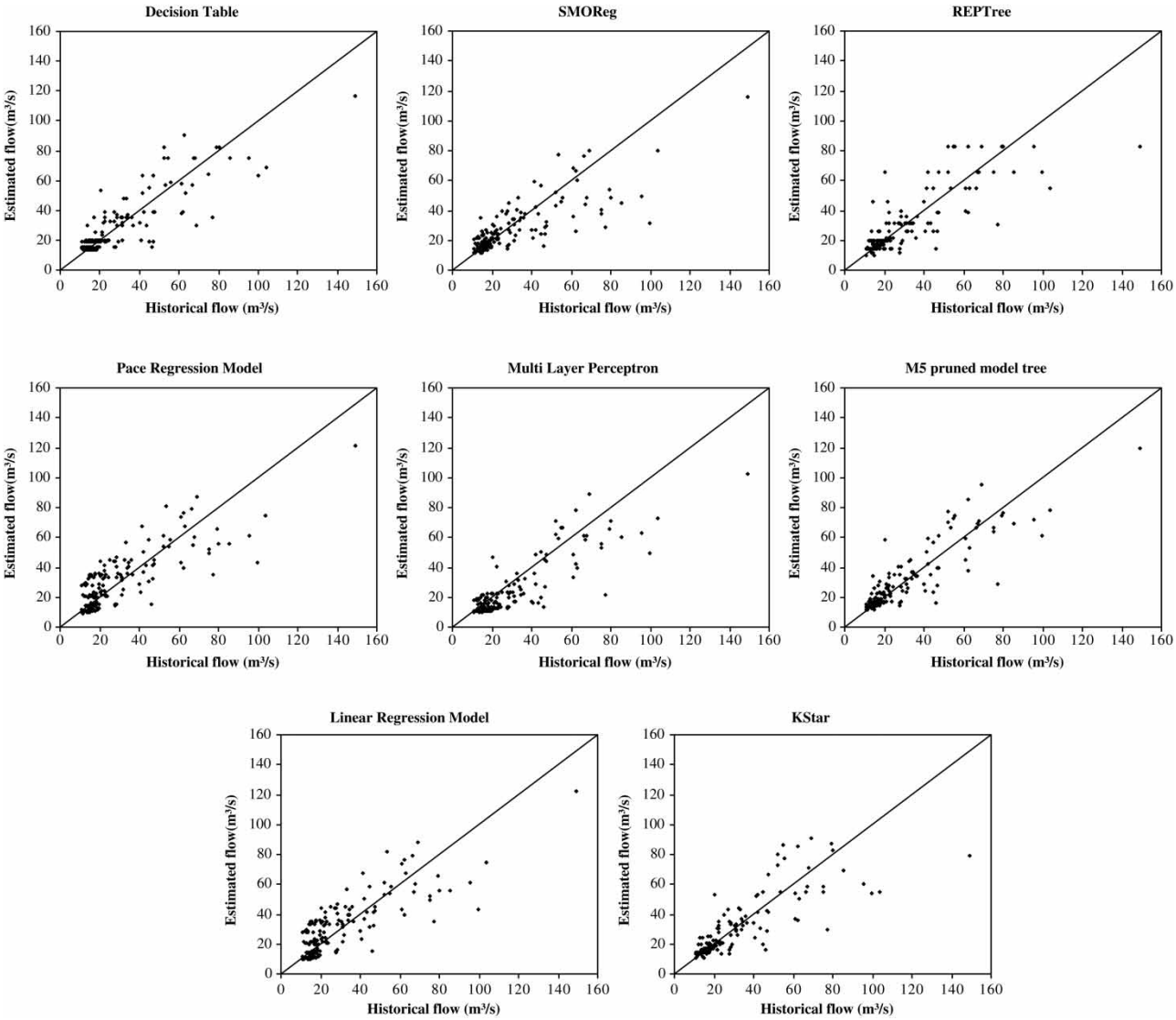

Figure 9 | Comparison of monthly flow prediction models for Seyhan Stream for testing data.

models were plotted against the historical monthly flow data for Seyhan Stream. Figure 9 shows that the M5 pruned model tree comparison plot is uniformly distributed around the 45° straight line, implying that there are no bias effects. As seen in Table 1, the M5 pruned model tree has maximum R^2 values, it is then concluded that this model is optimum for flow prediction for Seyhan Stream.

The results of statistical model for AR(2) model are given in Figure 10, which was drawn between the historical flow data and model. For this model R^2 value was

Table 1 | R^2 values between each model and monthly flows of Seyhan Stream for testing set

Algorithms	R^2
Linear regression model	0.65
Multilayer perceptron	0.68
Pace regression model	0.65
SMOReg	0.63
KStar	0.67
M5 pruned model tree	0.77
REPTree	0.68
Decision table	0.74

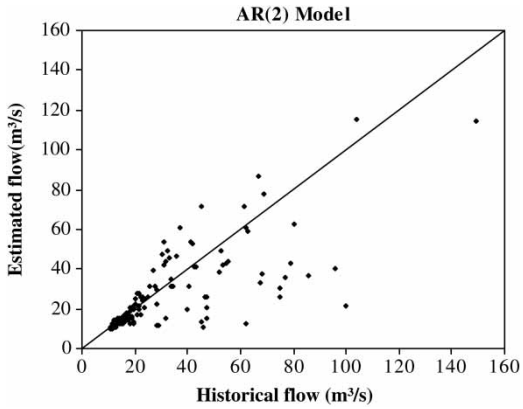

Figure 10 | The scatter diagram of AR(2) model for testing data.

calculated as 0.54. It was seen that M5 pruned model tree was the best model for flow prediction for Seyhan Stream.

CONCLUSIONS

Prediction of flow records is important for water resources planning and management for purposes such as flood control, use and operation of water, determination of settlement and energy production. Here, an alternative model is proposed for prediction of monthly flow records using data mining process for Seyhan Stream in Mediterranean Region, Turkey. The data mining process has been applied to Seyhan Stream which meets vital components such as irrigation, drinking water and power generation. The various models based on flow data are developed and compared to measured flow data. The most appropriate algorithm is determined according to the model performance criteria for testing dataset. It was also included that the AR(2) model was formed as hydrological time series modeling and results of the model were presented. The comparisons show that there is better agreement between monthly flow data and the results of the M5 pruned model algorithm in data mining process than others and AR(2). The performance of the developed models suggests that the flow could be successfully forecasted from available historical flow data using a data mining process, and the model is used in water resources planning and management. Also, this model is useful as only historical flow values are used. Therefore, the different hydrological variables are not necessary for the model. Hence, the developed model can be used to estimate monthly flow of Seyhan Stream in which the flow measurement system has failed or to estimate the missing monthly flow records.

REFERENCES

Alam, S., Kaushik, S. C. & Garg, S. N. 2009 Assessment of diffuse solar energy under general sky condition using artificial neural network. *Appl. Energy* **86**, 554–564.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearar, C. & Wirth, R. 2000 *CRISP-DM 1.0 Step-by-Step Data Mining Guide*. The CRISPDM Consortium/SPSS Inc. © Copyright 2000 SPSS Inc. CRISPWP-0800.

Chiu, S.-H., Chen, C.-C. & Lin, T.-H. 2008 Using support vector regression to model the correlation between the clinical metastases time and gene expression profile for breast cancer. *Artif. Intell. Med.* **44** (3), 221–231.

Cunnigham, S. J. & Holmes, G. 1999 Devoloping innovative application in agriculture using data mining. In *Proceedings of the Southeast Asia Regional Computer Confederation Conference*, Singapore (available on CD-ROM).

DeLurgio, S. A. 1998 *Forecasting Principles and Applications*. McGraw-Hill, New York, USA.

Fernandez, I. B., Zanakis, S. H. & Walczak, S. 2002 Knowledge discovery techniques for predicting country investment risk. *Comput. Indust. Eng.* **43**, 787–800.

Gelly, S., Mary, J. & Teytaud, O. 2006 *Learning for Stochastic Dynamic Programming*. ESANN 2006 European Symposium on Artificial Neural Network, 26–28 April, 2006. D-Side Publishers, Bruges, Belgium, pp. 191–196.

Hipel, K. W. 1985 Time series analysis in perspective. *Water Resour. Bull.* **21**, 609–623.

Hoffmann, D. & Apostolakis, J. 2003 Crystal structure prediction by data mining. *J. Molec. Struct.* **647**, 17–39.

Kohavi, R. & John, G. H. 1997 Wrappers for feature subset selection. *Artif. Intell.* **97** (1–2), 273–324.

Li, S. T. & Shue, L. Y. 2004 Data mining to aid policy making in air pollution management. *Expert Syst. Appl.* **27** (3), 331–340.

Licamele, K. & Getoor, L. 2006 Predicting Protein-Protein Interactions Using Relational Features. In *Proceedings of the ICML Workshop on Statistical Network Analysis*. ICML-SNA.

Panda, S. S., Singh, A. K., Chackraborty, D. & Pal, S. K. 2006 Drill wear monitoring using back propagation neural network. *J. Mater. Process. Technol.* **172**, 283–290.

Qinlan, J. R. 1992 Learning with continuous classes. In *Proceedings of the AI'92, 5th Australian Joint Conference on Artificial Intelligence*. World Scientific, Singapore, 343–348.

Quinlan, J. R. 1987 Simplifying decision trees. *Int. J. Man-Machine Studies* **27**, 221–234.

Read, B. J. 1999 Data mining and science? Knowledge discovery in science as opposed to business. 12th ERCIM Workshop on Database Research, Amsterdam.

Rupp, B. & Wang, J. 2004 Predictive models for protein crystallization. *Methods* **34**, 390–407.

Salas, J. D., Delleur, J. W., Yevjevich, V. & Lane, W. L. 1980 *Applied Modeling of Hydrologic Time Series*. Water Resources Publications, Colorado.

Sert, M. 1991 Simulation of the multi-reservoir systems operation in water resources planning. *Doğa J.* **15**, 145–158 (in Turkish).

Shearar, C. 2000 The CRISP-DM Model: The New BluePrint for Data Mining. *J. Data WareHous.* **5** (4), 13–22.

Tang, Z. & MacLennan, J. 2005 *Data Mining with Sql Server 2005*. John Wiley & Sons, New York.

Wirth, R. & Hipp, J. 2000 CRIPS-DM: towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, Manchester, UK, pp. 29–39.

Witten, I. H. & Frank, E. 2000 *Data Mining: Practical Machine Learning Tools with Java Implementations*. Morgan Kaufmann, San Francisco.

Wu, S., Zhao, X., Shao, H. & Ren, D. 2004 Cold rolling process data analysis based on Svm. In *Proceedings of the Third International Conference on Machine Learning and Cybernetics*. Institute of Electrical and Electronics Engineering, Shanghai.

Yuan, B., Wang, X. Z. & Morris, T. 2000 Software analyser design using data mining technology for toxicity prediction of aqueous effluents. *Waste Manage.* **20**, 677–686.

Zhou, Z.-H. 2003 Three perspectives of data mining. *Artif. Intell.* **143** (1), 139–146.

First received 3 January 2011; accepted in revised form 27 September 2011. Available online 11 June 2012