Water Resour Manage (2015) 29:1575–1590
DOI 10.1007/s11269-014-0895-5

Estimation of Failure Rate in Water Distribution Network Using Fuzzy Clustering and LS-SVM Methods

Mahmut Aydogdu · Mahmut Firat

Received: 11 November 2014 / Accepted: 30 November 2014 /
Published online: 6 December 2014

Abstract In this study, a novel approach combining fuzzy clustering and Least Squares Support Vector machine (LS-SVM) methods is developed for estimation of failure rate in water distribution networks and for determination of the relationship between failure rate-effective factors. For this aim, failure data observed Malatya water distribution network during 2006–2012 was selected as study area. In first phase, estimation model was developed and tested for the complete data set in estimating the failure rate by LS-SVM method. Then, in order to develop a more sensitive estimation model and to improve the performance of LS-SVM, 9 sub-regions were defined with similar characteristics by using fuzzy clustering method. Then failure rate estimation was carried out for each of the sub-regions using by LS-SVM method. Feed Forward Neural Network (FFNN) and Generalized Regression Neural Network (GRNN) methods were also used for estimation of failure rate and the results were compared with those of LS-SVM. The criteria such as Correlation Coefficient (R), Efficieny (E) and Root Mean Square Error (RMSE) were used to evaluate the performance of models. The results showed that LS-SVM model gives better results in comparison with the FFNN and GRNN models. It was also determined that LSSVM model results for the sub-regions defined by clustering analysis are better and that the clustering analysis increases the estimation model performance in addition to the fact that the estimation results have become better. In conclusion, it can be possible to develop a more sensitive estimation models using fuzzy clustering and LSSVM methods.

Keywords Water distribution network · Failure rate · LS-SVM · Fuzzy clustering

1 Introduction

Urban infrastructural systems can be pointed out as the most important structural elements of a city. That is why it is important to put forth their current status and to continue monitoring them. Transferring the infrastructural systems to the digital medium is not sufficient by itself

M. Aydogdu
Darende MYO, Inonu University, Malatya, Turkey

M. Firat (✉)
Civil Engineering Department, Inonu University, Malatya, Turkey
e-mail: mahmut.firat@inonu.edu.tr

for the sustainable management of infrastructural systems, to put forth the risk factors in the network and to manage the network. Knowing the frequency of failures in the network, determining the factors that cause these failures, determining the locational intensity of these failures and the determination of the failure rate according to material properties are important for a good infrastructural management as well as water loss management. A good network management is required to ensure a better management of the water provided to the network in regions where there is a fresh water problem and the failures should be minimized. Cooper et al. (2000) generated a failure estimation model in the Geographic Information System (GIS) environment using statistical based methods and have suggested a regression equation. Pelletier et al. (2003) examined the data from three different cities in order to estimate the pipe failures in water networks and to put forth the structural status after which a model has been developed using the different distribution functions. Sinske and Zietsman (2004) developed decision-support systems for the determination of pipe cracks in the water distribution networks as well as for pipe maintenance and enhancement. Yamijala (2007) carried out a study to estimate pipe fractures and to determine the parameters that are most effective in pipe fractures. Park et al. (2008) applied the log-linear ROCOF and the power law process to model the failure rate and estimate the economically optimal replacement time of the individual pipes. Wang et al. (2009) carried out a study using multi-regression analysis to estimate the annual failure rates based on various input variables such as the pipe diameter, age, and length. Carrión et al. (2010) analyzed failure data registered in normal operation conditions in a water supply network in order to evaluate the pipes failure probability. Tsitsifli et al. (2011) applied the Discriminant Analysis and Classification (DAC) method to achieve the above monitoring, repairing and replacing components of network and predict the future behavior of network pipes. Christodoulou and Deligianni (2010) carried out pipe failure analysis using fuzzy logic method. Oliveira et al. (2011) used density-based fuzzy clustering method to define the pipe failures that cause leakage in drinking water networks positionally. Christodoulou (2011) investigated the parameters that might cause failures and risks and has applied a multi-decision support system. Christodoulou et al. (2012) examined the factors that pose risks in the network for leakage analysis and management in the network and have carried out a risk evaluation. Rogers (2011) has suggested a failure estimation model using different parameters such as pipe material properties and surface properties. Fragiadakis et al. (2013) presented a methodology for the seismic risk assessment of water distribution networks based on general seismic assessment standards. Shi et al. (2013) have carried out a study in which they used the global clustering method for the positional change of pipe failures in drinking water networks. Aydın et al. (2014) presented a study for determining sustainability indices based upon performance criteria including reliability, resiliency, and vulnerability for pressure and water age in water distribution systems.

Clustering methods such as hierarchical and non-hierarchical clustering algorithms (K-means, Ward method etc.) have been widely used to identify the homogenous regions. Some specific applications of cluster methods include the identification of homogeneous regions for regional flood frequency analysis (Burn, 1989; Burn and Goel 2000; Lecce, 2000; Thandaveswara and Sajikumar, 2000), defining of the homogeneous precipitation regions (Smithers and Schulze, 2001; Nasseri and Zahraie, 2011). Data sets are separated into sub-regions with similar characteristics by clustering analysis and more sensitive estimation models are set up for these sub-regions. Fuzzy clustering method has been developed in recent years different from the hard clustering methods (K-means, Ward's method etc.) which have been used in modeling of water resources and hydrological processes (Shu and Burn 2004; Basu and Srinivas 2014). Fuzzy Clustering based on the fuzzy logic method was suggested by Dunn (1974) and developed by Bezdek (1981). The most important features that set fuzzy clustering method from classical methods is that a feature vector has different membership levels and the fact that it is assigned to the set with the highest membership level. That is, an element is not evaluated as "either belongs to a cluster or not" as is the case in classical clustering method. With this property of fuzzy clustering, it can be stated that its results include more information on explaining water resources and hydrological processes better than the conventional methods (Rao and Srivinas 2006; Dikbas et al. 2012). Kulkarni and Kripalani (1998) used Indian precipitation data to define homogeneous sub-regions via fuzzy clustering method. Rao and Srivinas (2006) carried out a study in which fuzzy clustering method has been used to determine the homogeneous hydrologic basins in the analysis of regional flood frequency. Dikbas et al. (2012) applied the fuzzy clustering method for defining of homogeneous precipitation regions.

Artificial intelligence techniques such as such as Artificial Neural Networks (ANN) and Fuzzy Logic (FL) have been recently accepted and used as an efficient alternative tool for modeling of complex water resources systems (Chen and Chau, 2006; Muttil and Chau, 2006; Chau 2007). Therefore, in this study, two different ANN techniques such as Generalized Regression Neural Networks (GRNN) and Feed Forward Neural Networks (FFNN) have been used to evaluate the performance of LS-SVM models. Some specific water resources system applications of ANN include modeling prediction of ground water level (Taormina et al. 2012), stream flow and discharge prediction (Cheng et al. 2005; Wu et al. 2009) and water consumption modeling (Firat et al. 2009). LSSVM method has been developed based on the SVM method and has first been suggested by Suykens and Vandewalle (1999). Some specific applications of LS-SVM method to hydrology include rainfall-runoff modeling, river flow estimation (Samsudin et al. 2011; Shabri and Suhartono 2012), sediment estimation (Kisi 2012), evapotranpration estimation (Samui 2011; Goyal et al. 2014).

In this study, a novel approach combining fuzzy clustering and LS-SVM methods is developed for estimation of failure rate in water distribution networks and for the determination of the relationship between failure rate-effective factors. In first phase, estimation model was developed and tested for the complete data set in estimating the failure rate by LS-SVM method. Then, in order to develop a more sensitive failure rate estimation model and to improve the performance of LS-SVM model, 9 sub-regions were defined with similar characteristics by using fuzzy clustering method. Then failure rate estimation was carried out for each of the sub-regions using by LS-SVM method. In addition, FFNN and GRNN methods were also used in the estimation of the failure rate and the results were compared with those of the LS-SVM models.

2 Material and Method

2.1 Fuzzy Clustering Method

In this study, the Fuzzy clustering method proposed by Dunn (1974) based on Fuzzy logic method and developed by Bezdek (1981), is used for identification of sub-failure rate regions. The main advantage and feature of FCM from hard clustering methods is that a feature vector in FCM can belong to several groups with the degree of belongingness specified by membership degree between 0 and 1. The X data set consisting of N feature vectors can be given as $X=\{x_k|k=1,2,.....N\}$ and the matrix can be written as in Equ.(1). Moreover, the objective function to be minimized in FCM and the membership matrix, U, showing the membership degrees of the feature vectors is given as in Equ. (2) and (3), respectively (Burn 1989; Rao and Srivinas, 2006; Dikbas et al. 2012; Aydogdu 2014).

$$X = \begin{bmatrix} x_{11}\ldots\ldots\ldots\ldots x_{1n} \\ \cdot \quad \cdot \quad \cdot\cdot \ \cdot \\ \cdot \quad\quad \cdot\cdot \ \cdot\cdot \\ x_{N1}\ldots\ldots\ldots\ldots x_{Nn} \end{bmatrix} \tag{1}$$

$$J(U, V : X) = \sum_{i=1}^{c} \sum_{k=}^{N} (u_{ik})^m d_{ik}^2(x_k, v_i) \tag{2}$$

$$U = \begin{bmatrix} u_{11}\ldots\ldots u_{12}\ldots\ldots u_{1c} \\ \cdot \\ u_{21} \quad \cdot \quad \cdot\cdot \ \cdot \\ \cdot \quad\quad \cdot\cdot \ \cdot\cdot \\ u_{N1}\ldots\ldots u_{N2}\ldots\ldots u_{Nc} \end{bmatrix}_{Nxc} \tag{3}$$

Where, N is the number of feature vectors, n is the number of variables, c is the number of clusters, V is the matrix containing the cluster centers, $u_{ik} \in [0,1]$ is the kth feature vector of ith cluster; x_k is the membership degree; $d_{ik}^2(x_k, v_i)$ is the distance between the kth feature vector and the ith cluster center; m $\in[1,\infty]$ is the fuzziness weight term. In FCM, the optimal number of groups was identified by using various validity indices such as; Partition Coefficient (PC) (Bezdek, 1981), Classification Entropy (CE) (Bezdek, 1981), Partition Index (SC), Separation Index (S), Xie and Beni Index (XB), Dunn Index (DI) and Alternative Dunn Index (ADI) measures (Valente de Oliveira and Pedrycz, 2007). The optimal number of clusters is determined by the lower value of CE, SC, S, and XB indices (Rao and Srivinas, 2006; Valente de Oliveira and Pedrycz, 2007; Dikbas et al. 2012).

2.2 LSSVM Method

The least squares support vector machine (LSSVM) method is firstly proposed by Suykens and Vandewalle (1999) based on support vector machine. The LSSVM is an effective method especially for the prediction and classification of nonlinear problems (Kumar and Kar 2002; Kisi 2012). The LSSVM model is trained and tested by using observed data set consisting of input (X) and output (Y) variables. In this model, the nonlinear prediction function for regression can be given as (Shabri and Suhartono 2012; Kisi 2012);

$$y(X) = w^T \varphi(x) + b \tag{4}$$

where, the output variable of LSSVM model is Failure Rate (y=FR), while the input vector used for prediction of output variable includes the input variables such as; Pipe Diameter (PD), Pipe Length (PL) and Pipe Age (PA), namely x=[PD,PL,PA]. Thus, equation (4) expresses the relationship between input and output variables. In this equation, b is the bias term; w is the weight vector; φ is a function mapping the inputs in m dimensional feature vector. The optimization problem of LSSVM model for regression prediction and constraint can be defined as given in equations (5) and (6), respectively (Samsudin et al. 2011; Shabri and Suhartono 2012);

$$\min \mathrm{R}(w,e) = \frac{1}{2}w^T w + \frac{\gamma}{2}\sum\nolimits_{i=1}^{n} e_i^2 \tag{5}$$

$$y(X) = w^T \varphi(x_i) + b + e_i \tag{6}$$

The Lagrange function constituted to solve the optimization problem given in equation (5) is shown in equation (7) (Shabri and Suhartono 2012);

$$L(w,b,e,a) = \frac{1}{2}w^T w + \frac{\gamma}{2}\sum\nolimits_{i=1}^{n} e_i^2 - \sum\nolimits_{i=1}^{n} \alpha_i \left\{ W^T \varphi(x_i) + e_i - y_i \right\} \tag{7}$$

Where, γ is the Kernel parameter, e_i is the slack variables for inputs and α_i is the Lagrange multiplier. The final LSSVM model used for regression prediction is obtained by solving the partial differential of equation (7) and is given in equation (8)

$$FR = y(x) = \sum\nolimits_{i=1}^{n} \alpha_i K(x_i, x) + b \tag{8}$$

In literature, many Kernel functions such as linear, polynomial, radial basis, sigmoidal has been proposed for LSSVM method. In this study, Radial Basis Function Kernel function given in equation (9) was applied for prediction of failure rates in water distribution system. Previous works using LSSVM method for prediction and estimation of problems have shown that the performance of Radial Basis Kernel Function, which is the most popular Kernel Function, is better than other Kernels such as Linear and Polynomial (Kisi 2012; Shabri and Suhartono 2012).

$$K(x_i, x) = \exp\left(-\|x - x_i\|/2\sigma^2\right) \tag{9}$$

Where, γ is the Kernel parameter and σ^2 is the regularization constant. The Kernel parameter defines the structure of feature vector and should be carefully chosen. Grid search algorithm is used to tune the regularization constant and width of RBF kernel parameters.

3 Study Area and Data

3.1 Study Area

Malatya water distribution network was selected as the application region for this study. The spring that feeds the Malatya water distribution network is a carstic spring with a maximum flow rate of 4000 *l*/s and an average flow rate of 2600 *l*/s. The Malatya drinking water spring supplies the demand of 19 districts and the municipality in addition to the center with a population of 550.000. The length of the network of the application region is about 440 km and the network includes type of pipes such as Polyvinyl Chloride Pipe (PVC), Asbestos Cement Pipe (ACP), Cast-iron Pipe, Steel Pipe and Polyethylene pipe (PE). The general appearance of the network is shown in Fig. 1.

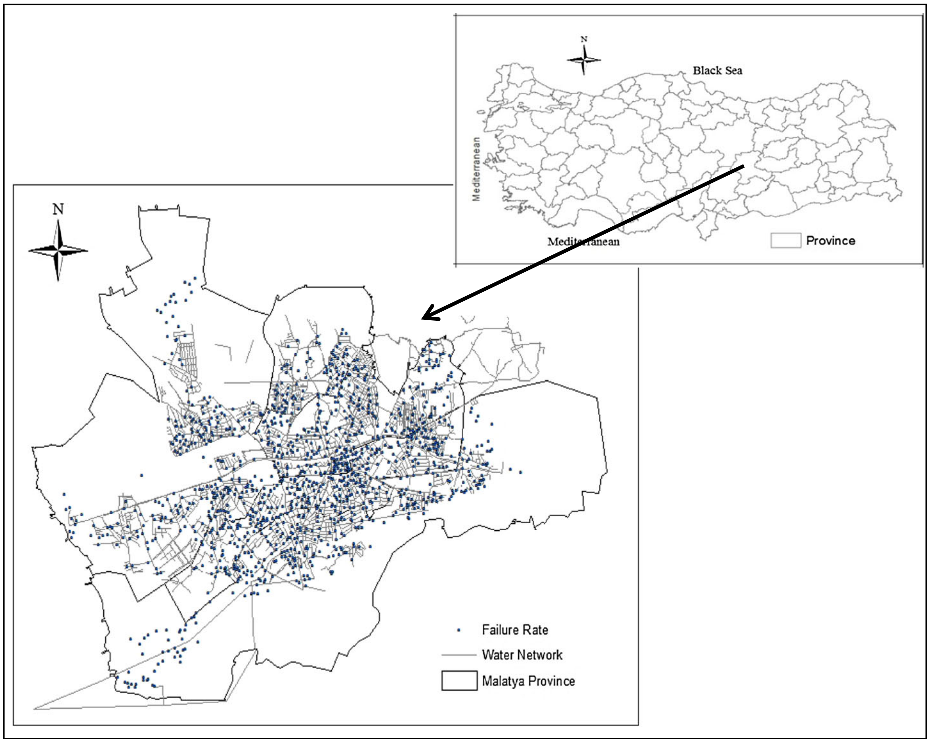

Fig. 1 Study Area (MASKI)

3.2 Data

Failure data records observed in water distribution network of Malatya during 2006–2012 were taken into account in order to estimate the failure rates in water distribution networks and to define the relationship between the effective variables. Pelletier et al. (2003) stated that the main difficulty in developing mathematical models for this type of problem is the lack of data on both the water pipe breaks. It is seen in literature that some studies related to modeling of pipe failures have been used a limited data as very little information is available for all pipe segments. Rogers (2011) utilized GIS to predict the failure risk based on pipe diameter, pipe material and pipe age. Shi et al. (2013) carried out the failure factor analysis between failure rate and four factors: pipe diameter, pipe age, material and temperature. Wang et al. (2009) considered the pipe characteristics such as pipe material, diameter, age, and length to develop the deterioration models that predict the failure rates by multiple regression analysis. In this study, the pipe characteristics such as pipe diameter, pipe length and pipe age were considered as effective factors for estimation of failure rate. Here, the ground level and traffic density variables have not been taken into account since healthy and reliable data for these parameters could not be acquired (Aydogdu 2014).

In the application region, it is difficult to achieve the accurate and reliable data recorded in water distribution system. The quality of data affects the performance of estimation models. Therefore, in this study, only failures in normal operation conditions have been considered, excluding those caused by abnormal events such as failures caused by other institutions (Telekom, Sewage unit, Electricity Company etc.), installation of new lines, cancelled service connections, and the replacement work carried out at the lines. The almost 21,000 failure data have been examined within the scope of this study in accordance with the criteria given above and the about 5111 failure data have been used. In this study, failure rate variable has been defined as the ratio of the number of failures to pipe length instead of the number of failure data. For the application area, 1231 failure rates have been calculated for clustering and estimation models by taking into account this failure rate (Aydogdu 2014).

4 Analysis and Discussion

4.1 Statistical Evaluation of Failure Records

In this section, the relationship between failure data and pipe characteristics such as pipe age, length and diameter has been evaluated statistically and graphically. When 5111 failures data are examined, it was observed that different rates of failures have occurred in different pipe types and diameters. It was been observed that the highest rate was for PVC pipes with 59.53 %, followed by ACP with 21.19 %, Cast-iron pipes with 17.68 % and PE pipes with 1.60 %. It is seen that the failure rates are high for PVC pipes. However, the total length of the PVC pipes in the network is 304.88 km and they constitute about % 69.10 of the total pipeline. That is why the *Failure Rate* defined as the ratio of the number of failures to the length of the pipe has been used in order to carry out a more accurate evaluation and comparison. Failure rate has been calculated as 0.00964 for PVC pipes, 0.0085 for ACP and 0.0846 for Cast-iron pipes. Here, the change of failure rate with respect to pipe characteristics such as pipe length, pipe diameter and pipe age in the network was demonstrated in Fig. 2.

According to Fig. 2, it is observed that the highest failure rate occurs on the pipeline with a length of 0–200 m. It has also been determined that the failure rate is high for pipes in the 200–400 m interval. When the results for different pipe materials are examined, it has been observed that the highest failure rate occurs among all pipes of ACP, Cast-iron pipe and PVC with a length of 0–200 m. This can be evaluated to be related with the fact that the number of pipes with lengths of 0–200 m is high. It is observed in Fig. 2 that the failure rates in pipes with a diameter of 110 mm are greater in comparison with others. When the failure rate with respect to diameter in PVC pipes with the highest pipeline length is examined, it is observed again that the highest value is for the pipes with diameter of 110 mm. The fact that the failure rate of PVC pipes is close to the results obtained for the whole line can be explained by the fact that the length of PVC pipes in the pipeline is high and that the 110 mm diameter pipe ratio among PVC pipes is also high. When we consider the other pipe types, the highest ratio is observed in 90 mm diameter pipes for Cast-iron pipes whereas the highest failure rate was observed for 150 mm diameter pipes among ACP. It is seen in Fig. 2 that the highest failure rate is for the pipes in the age interval of 15–20, whereas the lowest failure rate was observed for the pipes in the age interval of 25–30. As can be seen from the graph, the failure rate for pipes in the age interval of 10–15 is quite high. On the other hand, when we make an evaluation for other pipe types it has been observed that the highest failure rate was for the pipes in the 15–20 age intervals for PVC and ACP. According to the results obtained, the highest failure rate was not observed in older pipes contrary to what was expected. This can be explained as follows; since the old pipes in the network are replaced with newer ones, the ratio of pipes over the age of 25 decreases in the network thereby decreasing the failure rate.

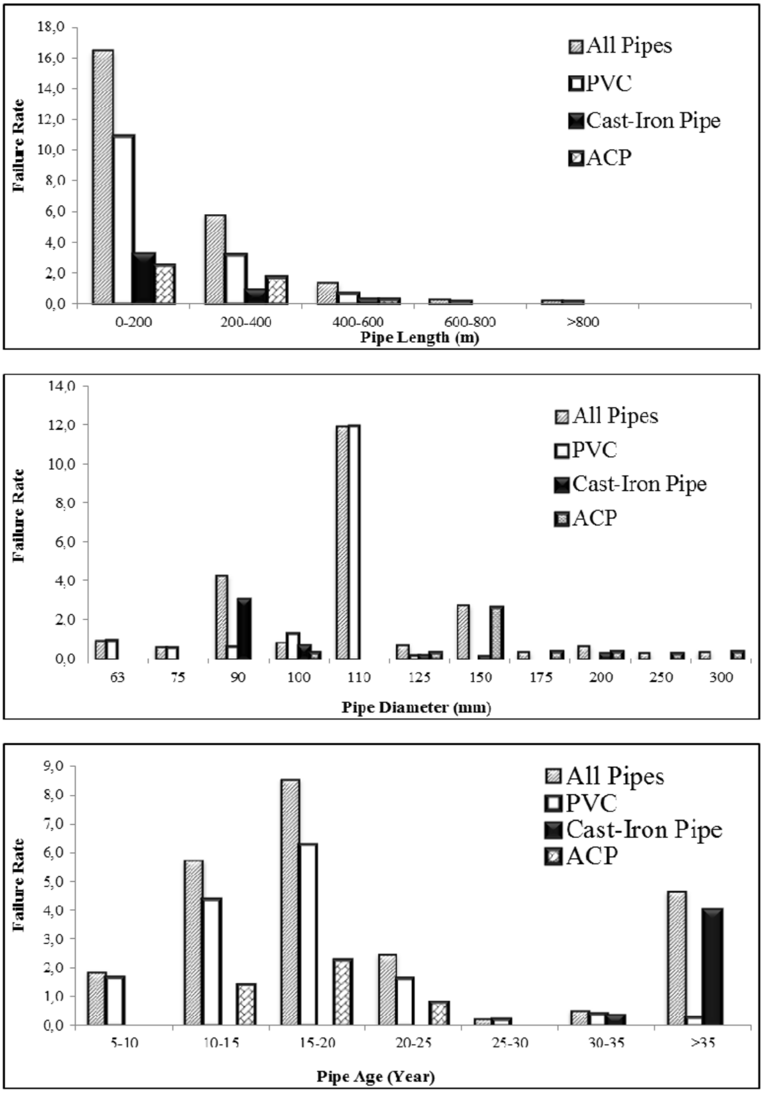

Fig. 2 The changes of Failure Rates with pipe characteristics

4.2 Clustering Analysis Results

Clustering was carried out for the failure rate in water distribution network by using the total of 1231 data. The analysis was started by choosing the number of clusters to be at least 2 and the optimum number of clusters were tried to be found by increasing the number of clusters. Various factors were calculated to decide on the most suitable number of clusters and the change of these calculated criteria with respect to the number of cluster has been shown in Fig. 3.

According to Fig. 3, it is seen th. small changes in the SC indice after cluster 9. It is observed that the S indice decreases until cluster 9 and that the lowest value was obtained for

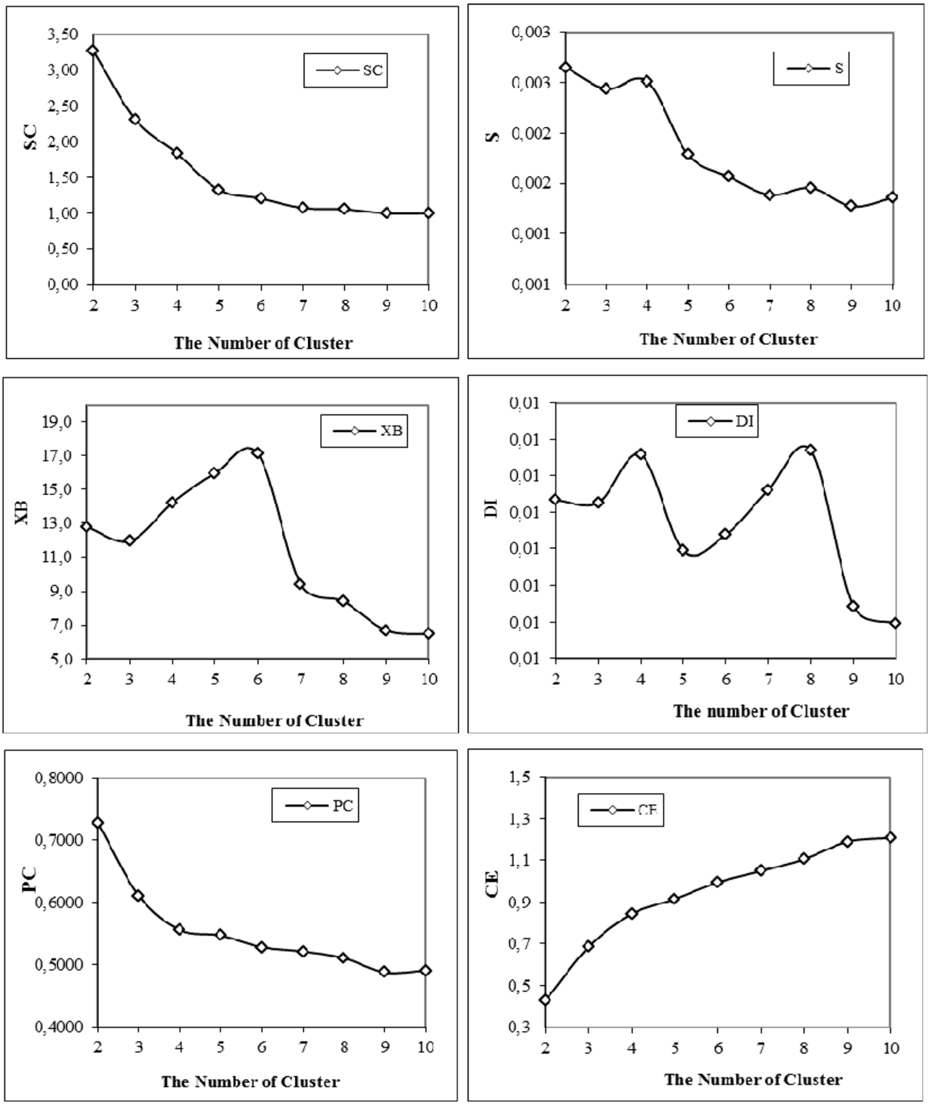

Fig. 3 Variation of the index values according to the number of clusters

the cluster 9. On the other hand, the lowest value of the XB indice was obtained for the cluster number 9. In addition, when we examined the DI criteria, it is observed that the lowest value is obtained for the cluster number 10. According to the evaluations, the optimum number of clusters was determined to be 9 for failure rate with fuzzy clustering method. The distribution of the failure rates of the 9 sub-regions defined by fuzzy clustering method is shown in Fig. 4.

4.3 Estimation of Failure Rate

The estimation model was set up initially for the Data Set I with 1231 data for which clustering has not been made. 1009 data were used for this data set to train the LSSVM model and the

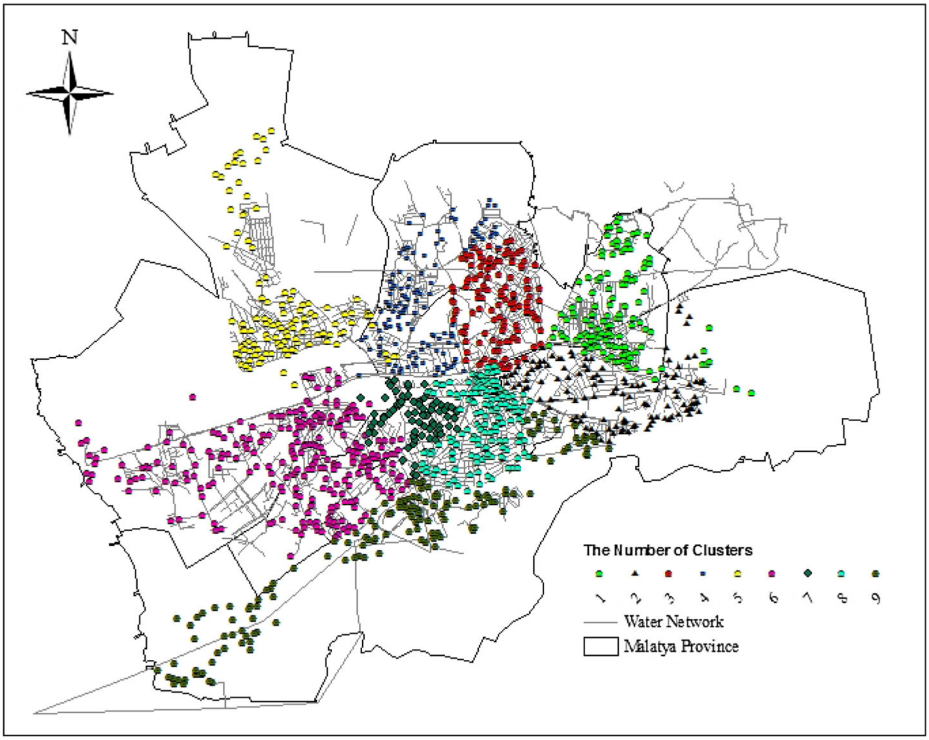

Fig. 4 The location of failure rates in clusters identified by Fuzzy Clustering

remaining 222 data were used to test the model. Secondly, the data set was divided into groups using fuzzy clustering method in order to improve the performance of the LS-SVM model and to obtain more sensitive and reliable results. Estimation models for each of the 9 sub-regions (Set II C1-C9) defined by fuzzy clustering method were developed and tested using LSSVM method. 80 % of the data were used for the training of estimation models in each data set and the remaining 20 % was used for the testing of the models. The data used in the testing of each model was selected randomly from among the total data set and these data were not used in the training of the models. The optimal values of model parameters of LSSVM method, (γ, σ2) were determined by grid search algorithm. In this study, for LS-VM estimation models, grid search of model parameters with σ^2 in the range 1 to 32 and γ in the range 0.01 to 14 was determined. In the search space, a 3-fold cross validation method on the training set was applied to each sub-group data set for computing the validation of LS-SVM model. The software package LS-SVMlab 1.8 developed by Pelckmans et al. (2011) in Matlab was used in the training and testing of the LSSVM models.

In literature, various performance evaluation criteria are used to compare the results of estimation models. The performance of models developed for modeling of water resources and hydrological processes are commonly evaluated based on statistical criteria such as, Correlation Coefficient (R), Efficiency (E), and Root Mean Square Error (RMSE) (Samsudin et al. 2011; Goyal et al. 2014). Therefore, in this study these statistical criteria were considered for evaluating of performance of LS-SVM estimation models. The comparison of estimation models is given in Table 1.

Table 1 Comparison of performances of the models

Data Set		LS-SVM			FFNN			GRNN		
		R	E	RMSE	R	E	RMSE	R	E	RMSE
Set I		0.436	0.401	0.0138	0.331	0.307	0.0224	0.291	0.315	0.0185
Set II	C1	0.858	0.864	0.0086	0.721	0.658	0.0115	0.412	0.366	0.0137
	C2	0.672	0.629	0.0062	0.538	0.505	0.0081	0.598	0.502	0.0069
	C3	0.657	0.618	0.0040	0.608	0.536	0.0053	0.656	0.571	0.0042
	C4	0.723	0.890	0.0061	0.635	0.509	0.0069	0.692	0.755	0.0064
	C5	0.703	0.635	0.0038	0.547	0.519	0.0065	0.665	0.538	0.0060
	C6	0.656	0.607	0.0114	0.562	0.555	0.0149	0.503	0.552	0.0143
	C7	0.615	0.596	0.0057	0.580	0.541	0.0064	0.589	0.522	0.0098
	C8	0.732	0.698	0.0147	0.631	0.566	0.0165	0.669	0.658	0.0154
	C9	0.705	0.659	0.0057	0.681	0.625	0.0071	0.685	0.632	0.0067

$$\mathrm{E} = \frac{\mathrm{E}_1 - \mathrm{E}_2}{\mathrm{E}_1} \quad \mathrm{E}_1 = \sum_{i=1}^{N} \left(FR_{ocserved} - \overline{FR_{obsreved}} \right)^2, \ \mathrm{E}_2 = \sum_{i=1}^{N} \left(FR_{observed} - FR_{estimate} \right)^2 \tag{10}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(FRi_{observed} - FRi_{estimate} \right)^2} \tag{11}$$

where, $FR_{estimate}$ is the estimated failure rate, $FR_{observed}$ is the observed failure rate, $\overline{FR_{observed}}$ is the mean of the observed failure rate data. The correlation coefficient is a commonly used statistic and provides information on the strength of linear relationship between the observed and the estimated values. The Efficiency E is a statistic employed to evaluate model performance. Values of R and E close to 1.0 indicate good model performance.

It is observed in Table 1 that the correlation between the results of the LS-SVM model and observation data for Set I has been calculated as 0.436 and that this value is quite small. When the performance of LS-SVM models for the 9 sub-regions defined by fuzzy clustering method is evaluated, it is seen that the results for Set II including the 9 sub-regions are better in general in comparison with those of Set I. It can be said that the correlation coefficients calculated for each sub-region are at acceptable levels. When the results of these sets are evaluated separately, the highest correlation (0.858) and E (0.864) values were obatined for Set II C1 LS-SVM model, whereas the lowest correlation (0.615) and E (0.596) values were calculated for Set II C7 LS-SVM model. On the other hand, it is observed that the correlation and E values of Set II LS-SVM models are greater than those calculated for Set I LS-SVM model. In addition, it has been determined that the RMSE values calculated for Set II LS-SVM models are in general lower than those calculated for Set I. The table also shows the FFNN and GRNN model results for both Set I and Set II. When the correlation coefficient and E values of FFNN model are examined, it is observed that the highest correlation coefficient (0.721) and E (0.658) was calculated for Set II C1 FFNN model, whereas the lowest correlation coefficient (0.538) and E (0.505) values were calculated for Set II C2 FFNN model. Moreover, when the correlation coefficient and E values of GRNN model are evaluated,it is seen in table that the highest correlation coefficient (0.692) and E (0.755) was calculated for Set II C4 GRNN model, whereas the lowest correlation coefficient (0.412) and E (0.366) values were calculated for Set II C1 GRNN model. When the results for the FFNN, GRNN and LS-SVM models are compared, it is observed in general that the correlation coefficient and E values calculated for the LS-SVM models are greater than those calculated for the FFNN and GRNN models. On the other hand, when we compare the RMSE values of these two methods, it has been determined that in general the values obtained for LS-SVM models are smaller. Accordingly, it can be stated that the results of the LS-SVM models developed for the sub-regions defined by clustering analysis are better and that clustering analysis increases the performance of the estimation models. Figure 5 shows the comparison of results of LSSVM models and observed data.

Figure 5 demonstrates that the performance of LS-SVM models for each sub-groups defined by clustering analysis is better than LS-SVM model developed for Set I and in general satisfactory. A good way of managing drinking water distribution systems, identifying the factors that cause failure and estimate the failure rate is very important for decision-makers and administrators. Defining spatial change of failures and assessing of current status of water distribution systems can

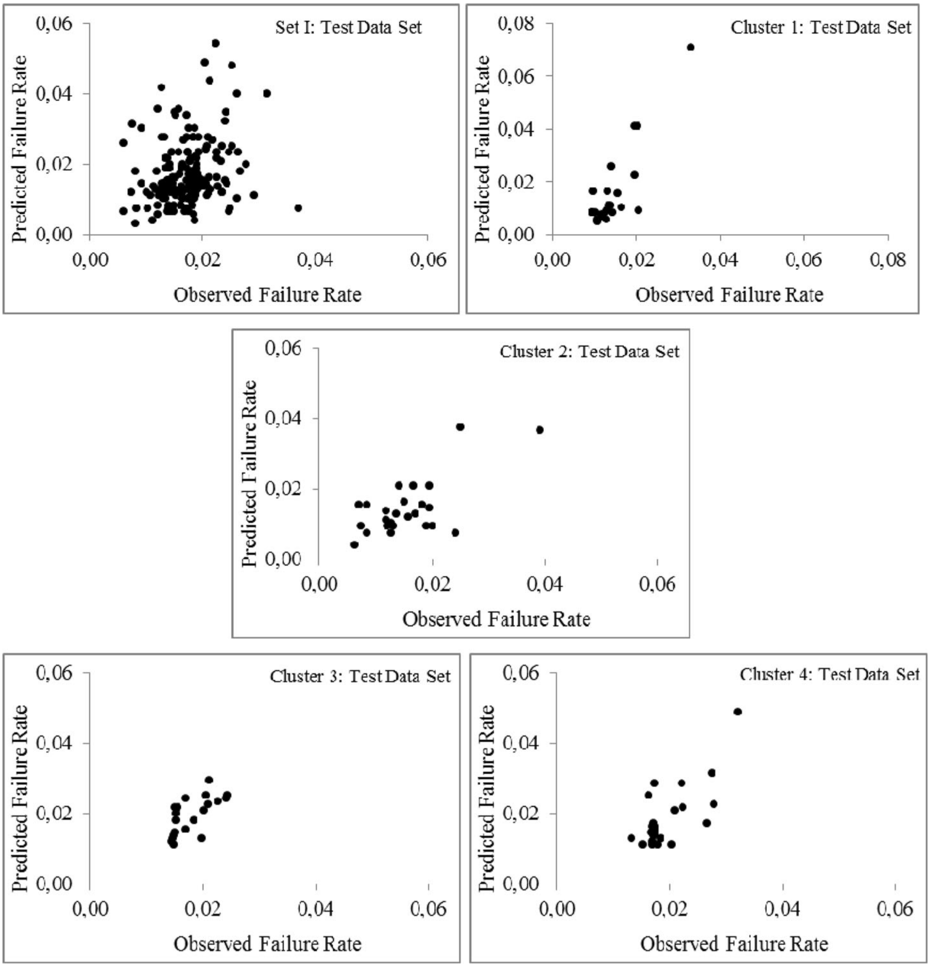

Fig. 5 Comparison of LSSVM Model Results and Observed Data

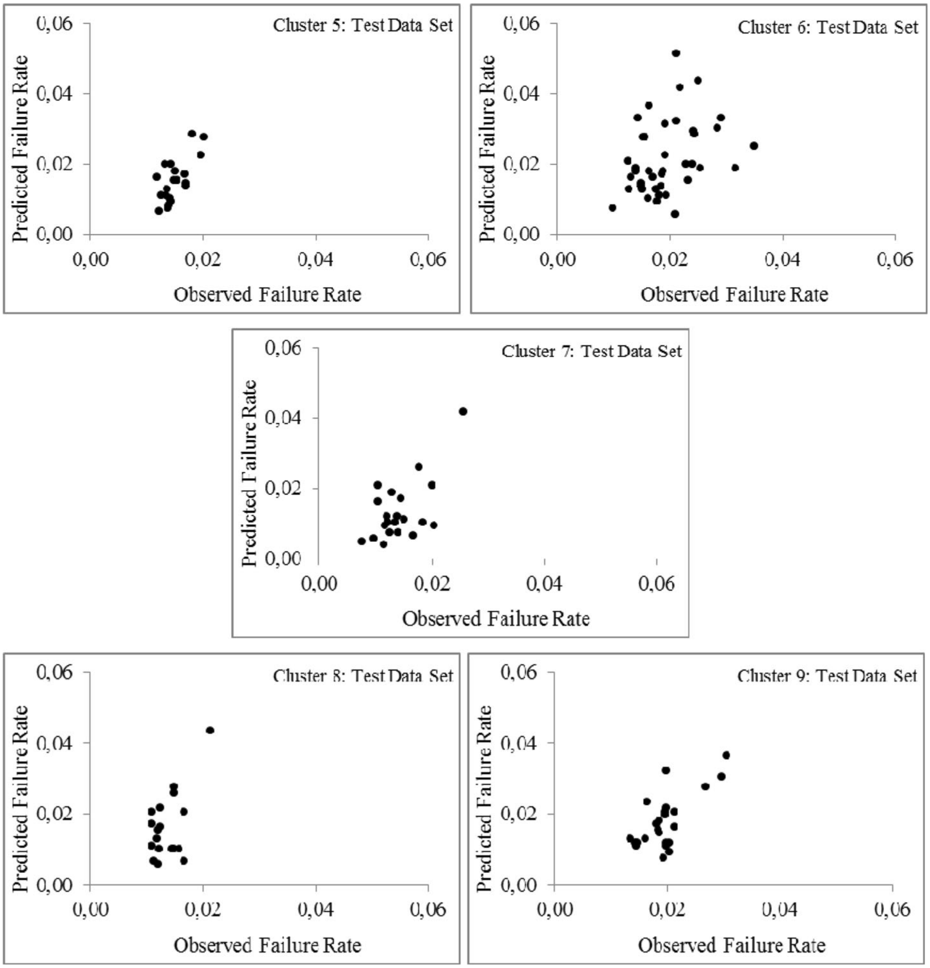

Fig. 5 (continued)

be affected for estimation of future status. According to results above, it is considered that Fuzzy clustering method used in this study can be used as an effective tool for defining of sub-groups with similar characteristics. Moreover, the use of clustering analysis and LS-SVM, which showed a good performance in the estimation of failure rate for each sub-groups defined by fuzzy clustering method, together is very important to be used as alternative methods for managing of water distribution system. It is important to develop the failure rate estimation models by using the variables such as traffic intensity, ground water level, rainfall and temperature and pressure in water distribution for good management of water distribution system in future works.

5 Results

In this study, a novel approach combining fuzzy clustering and LS-SVM methods applied for estimation of failure rate in water distribution networks and for the determination of the relationship between failure rate-effective factors. For this aim, the failure data records observed in the Malatya water distribution network during 2006–2012 were considered and the relationship between failure rate and effective factors has been evaluated for the application region prior to setting up the estimation models. It was observed that the highest failure rate was for the pipes with lengths of 0–200 m. Here, it has also been determined when we evaluated the failure rate-pipe diameter that the pipes with diameters of 110 m have the highest failure rate. Finally, when the relationship between failure rate-pipe ages was examined, it was observed that the highest failure rate was for the pipes with ages in the interval of 15–20. Estimation model was firstly developed and tested for the complete data set in estimating the failure rate by LS-SVM method. Then, in order to develop a more sensitive failure rate estimation model and to improve the performance of LS-SVM model, 9 sub-regions were defined with similar characteristics by using fuzzy clustering method. Failure rate estimation was carried out for each of the 9 sub-regions using the LS-SVM method. It is observed that the correlation coefficient between the results of LSSVM estimation model and observed data for Set I is quite low. On the other hand, when the results of LSSVM models for the 9 sub-regions defined by clustering method are examined, it is observed that the results for Set II are better in general in comparison with those of Set I. When the results of these Set II are evaluated separately, the highest correlation coefficient (0.858) and E (0.864) values are obtained for the Set II C1 LS-SVM model, while the lowest correlation coefficient (0.615) and E (0.596) values have been calculated for the Set II C7 LS-SVM model. In addition, it can be stated that the RMSE values calculated for Set II LSSVM models are in general at better levels in comparison with the values calculated for Set I. It can be said that the correlation coefficients calculated for estimation models of each subset are at acceptable levels. It is observed that the correlation and E values for Set II LS-SVM models are greater than the correlation and E values for Set I LS-SVM model. In addition, it was determined that the RMSE values for the Set II LSSVM models are in general lower in comparison with the values calculated for Set I. According to results, the LS-SVM estimation model performance increases with the application of clustering analysis, that the correlation and E values are greater for the data set values not subject to clustering and that the RMSE values are lower. On the other hand, FFNN and GRNN models have been trained and tested for both Set I and Set II. When the FFNN and GRNN and LSSVM model results are compared, it can be said that in general the LSSVM models have a better performance. When the correlation and E values for all methods are evaluated, the values for LSSVM model for each data set are at greater levels in comparison with the FFNN and GRNN models. Similarly, when the RMSE values are compared, it is observed that the values calculated for the LSSVM model are smaller than those calculated for the FFNN and GRNN models. The results showed that the LSSVM model results for the sub-regions defined by clustering analysis are better and that the clustering analysis increases the estimation model performance in addition to the fact that the estimation results have become better. In conclusion, it can be said that the use of clustering analysis and LS-SVM methods together brings about a better performance in the estimation of the failure rate of the developed models thus yielding successful results. In this study, the pipe characteristics such as pipe diameter, pipe length and pipe age were considered as effective factors for estimation of failure rate. In future works, in addition to the pipe characteristics, the variables such as traffic intensity, ground water level, rainfall and temperature and pressure in water distribution can be taken into account to develop the failure rate estimation models. Moreover, the estimation models can be developed by using different clustering and regression methods.

Acknowledgments Authors greatly appreciate the cooperation of MASKI (General Manager of water and sewerage works of Malatya) authorities for providing data and information. The authors are also grateful for Editor and Anonymous Reviewers for their helpful and constructive comments on an earlier draft of this paper.

References

Aydın NY, Mays L, Schmitt T (2014) Sustainability assessment of urban water distribution systems. Water Resour Manag 28:4373–4384

Aydogdu M (2014) Analysis of pipe failure occurred in water distribution system using cluster method. MSc Thesis. Inonu University, Turkey (in Turkish)

Basu B, Srinivas VV (2014) Regional flood frequency analysis using kernel-based fuzzy clustering approach. Water Resour Res 50(4):3295–3316

Burn DH (1989) Cluster analysis as applied to regional flood frequency. J Water Resour Plan Manag 115:567–582

Burn DH, Goel NK (2000) The formation of groups for regional flood frequency analysis. Hydrol Sci J 45(1): 97–112

Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press: New York

Carrión A, Solano H, Gamiz ML, Debon A (2010) Evaluation of the reliability of a water supply network from right-censored and left-truncated break data. Water Resour Manag 24(12):2917–2935

Chau KW (2007) An ontology-based knowledge management system for flow and water quality modeling. Adv Eng Softw 38(3):172–181

Chen W, Chau KW (2006) Intelligent manipulation and calibration of parameters for hydrological models. Int J Environ Pollut 28(3–4):432–447

Cheng CT, Chau KW, Sun Y, Lin J (2005) Long-term prediction of discharges in Manwan Reservoir using artificial neural network models. Lect Notes Comput Sci 3498:1040–1045

Christodoulou SE (2011) Water network assessment and reliability analysis by Use of survival analysis. Water Resour Manag 25:1229–1238

Christodoulou S, Deligianni A (2010) A Neurofuzzy Decision Framework for the Management of Water Distribution Networks. Water Resour Manag 24:139–156

Christodoulou S, Gagatsis A, Agathokleous A, Xanthos S, Kranioti S (2012) Urban water distribution network asset management using spatio-temporal analysis of pipe-failure data, 14th International Conference on Computing in Civil and Building Engineering Moscow, Russia.

Cooper NRG, Blakey C, Sherwin TJ, Whiter T, Woodward CA (2000) The use of GIS to develop probability-based trunk main burst risk model. Urban Water 2:97–103

de Oliveira D, Neill D, Jr Garrett J, Soibelman L (2011) Detection of Patterns in Water Distribution Pipe Breakage Using Spatial Scan Statistics for Point Events in a Physical Network. J Comput Civ Eng 25(1):21–30

Dikbas F, Firat M, Koc AC, Gungor M (2012) Classification of Precipitation Series using Fuzzy Cluster Method. Int J Climatol 32(10):1596–1603

Dunn JC (1974) A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters. J Cybern 3(3):32–57

Firat M, Yurdusev MA, Turan ME (2009) Evaluation of Artificial Neural Network Techniques for Municipal Water Consumption Modeling. Water Resour Manag 23(4):617–632

Fragiadakis M, Christodoulou SE, Vamvatsikos D (2013) Reliability assessment of urban water distribution networks under seismic loads. Water Resour Manag 7:3739–3764

Goyal MK, Bharti B, Quilty J, Adamowski J, Pandey A (2014) Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, fuzzy logic, and ANFIS. Expert Syst Appl 41:5267–5276

Kisi O (2012) Modeling discharge-suspended sediment relationship using least square support vector machine. J Hydrol 456–457:110–120

Kulkarni A, Kripalani RH (1998) Rainfall patterns over India: Classiffication with fuzzy C-means method. Theor Appl Climatol 59:137–146

Kumar M, Kar IN (2002) Non-linear HVAC computations using least square support vector machines. Energy Convers Manag 50:1411–1418

Lecce SA (2000) Spatial variations in the timing of annual floods in the southeastern United States. J Hydrol 235: (3-4) 151–169

Muttil N, Chau KW (2006) Neural network and genetic programming for modelling coastal algal blooms. Int J Environ Pollut 28(3–4):223–238

Nasseri M, Zahraie B (2011) Application of simple clustering on space-time mapping of mean monthly rainfall pattern. Inte Climatol 31(5):732–741

Pelckmans K, Suykens J, Van G, De Brabanter J, Lukas L, Hanmers B, De Moor B, Vandewalle J (2011). LS-SVMlab: a MATLAB/C toolbox for Least Square Support Vector Machines. http://www.esat.kuleuven.ac.be/sista/lssvmlab. Accessed 19 Sep 2013

Park S, Jun H, Kim BJ, Im GC (2008) Modeling of water main failure rates using the Log-linear ROCOF and the power Law process. Water Resour Manag 22(9):1311–1324

Pelletier G, Mailhot A, Villeneuve J-P (2003) Modeling water pipe breaks—three case studies. J Water Resour Plan Manag 129(2):115–123

Rao AR, Srivinas VV (2006) Regionalization of watersheds by fuzzy cluster analysis. J Hydrol 318:57–79

Rogers PD (2011) Prioritizing water main renewals: case study of the Denver water system. J Pipeline Systems Engineering and Practice 2(3):73–81

Samsudin R, Saad P, Shabri A (2011) River flow time series using least squares support vector machines. Hydrol Earth Syst Sci 15:1835–1852

Samui P (2011) Application of least square support vector machine (LSSVM) for determination of evaporation losses in reservoirs. Engineering 3:431–434

Shabri A, Suhartono (2012) Streamflow forecasting using least-squares support vector machines. Hydrol Sci J 57(7):1275–1293

Shi W-Z, Zhang A-S, Ho O-K (2013) Spatial analysis of water mains failure clusters and factors: a Hong Kong case study. Ann GIS 19(2):89–97

Shu C, Burn DH (2004) Homogeneous pooling group delineation for flood frequency analysis using a fuzzy expert system with genetic enhancement. J Hydrol 291:132–149

Sinske SA, Zietsman HL (2004) A spatial decision support system for pipe-break susceptibility analysis of municipal water distribution systems. Water SA 30(1):71–79

Smithers JC, Schulze RE (2001) A methodology for the estimation of short duration design storms in South Africa using a regional approach based on L-moments. J Hydrol 241:(1-2) 42–52

Suykens JAK, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9(3): 293-300

Thandaveswara BS, Sajikumar N (2000) Classification of river basins using Artificial Neural Network. J Hydrol Eng 5(3):290–298

Taormina R, Chau KW, Sethi R (2014) Artificial Neural Network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon. Eng Appl Artif Intell 25(8):1670–1676

Tsitsifli S, Kanakoudis V, Bakouros I (2011) Pipe networks risk assessment based on survival analysis. Water Resour Manag 25(14):3729–3746

Valente de Oliveira J, Pedrycz W (2007) Advances in fuzzy clustering and its applications. John Wiley & Sons Ltd. 457p

Wang Y, Zayed T, Moselhi O (2009) Prediction models for annual break rates of water mains. J Perform Constr Facil 23(1):47–57

Wu CL, Chau KW, Li YS (2009) Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. Water Resour Res 45(8), W08432

Yamijala S (2007) Statistical Estimation of Water Distribution System Pipe Break Risk, MSc Thesis, Texas A&M University USA