Chapter 20
Electrical Energy Consumption Forecasting to Improve Energy Efficiency of Water Distribution Systems

Gheorghe Grigoras

Abstract The energy management problem is essential in optimal planning and operation of water distribution systems. This problem involves establishing the operation schedule for all water hydrophore stations from the system. One of the available policies for improving energy efficiency is related to the decrease of electricity consumption. This supposes two important aspects: the upgrading of the water distribution system and use of software packages order to ease the decision making process. Two approaches based on clustering and decision trees are proposed for electrical energy consumption forecasting in the water distribution systems. The comparative studies were realized using a database of 85 urban water hydrophore stations belonging to a Romanian water distribution company. Based on these results, it can be considered that the new proposed approaches have the ability to constitute an IT infrastructure which can be actively used for improving energy efficiency in water distribution systems.

Keywords Electrical energy consumption · Clustering · Decision trees · Hydrophore stations · Water distribution system

Abbreviation and Acronyms

AI	Artificial Intelligence
ANN	Artificial Neural Networks
CART	Classification and Regression Trees
EP	Evolutionary Programming
FT	Fuzzy Techniques
RLP	Representative Loading Profiles
SGI	Silhouette Global Index
MAPE	Mean Absolute Percentage Error
WF	Daily Water Flow

G. Grigoras (✉)
Department of Power System, Faculty of Electrical Engineering,
“Gheorghe Asachi” Technical University of Iasi,
Bd. Dimitrie Mangeron, No. 21-23, 700050 Iasi, Romania
e-mail: ggrigor@tuiasi.ro

N. Bizon et al. (eds.), *Energy Harvesting and Energy Efficiency*,
Lecture Notes in Energy 37, DOI 10.1007/978-3-319-49875-1_20

PG Pressure Growth
W Daily Electrical Energy Consumption

20.1 Energy Efficiency of Water Distribution Systems

The linkage between two important components of our life, namely water and energy, sometimes called the water-energy nexus, is increasingly tighter due to massive demand for both water and energy. Water is used in electrical energy production and electrical energy is necessary to extract, treat and distribute water and to clean the waste water.

These aspects explain the water-energy nexus, such that both components must be addressed together in an efficiency energy policy [1]. Even if there is a relationship between the two components, these have been considered and managed independently until recently.

In planning or operating water supply systems, traditionally the electrical energy consumptions were mostly neglected or given little importance [2, 3]. The interdependency between these two important components is shown in Fig. 20.1 [4].

These interdependencies must be considered, because the constraints of one resource introduce constraints in the other. For example, the droughts and heat waves can lead to water constraints that, subsequently, can become constraints in the hydropower sector. In the same way, the outages from power systems can represent constraints in the water and wastewater sectors [3, 5].

High energy amounts are needed for pumping water, treating the water to drinking quality, distribution of water, pumping and treating wastewater. Thus, obtaining potable water and cleaning wastewater using less electrical energy amount should represent an important target for water companies. For example, there are countries where energy used for water treatment is of about 3% of the nation's electricity use (in USA and UK). In some European countries, Sweden for example, the energy amount is approximately 1%. However, there are countries

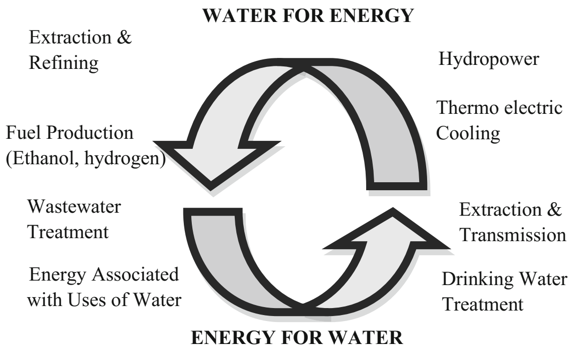


Fig. 20.1 The water-energy nexus [4]

with a big percent, around 10% (for example Israel) [1, 6]. From the entire chain, water pumping requires the highest energy amount.

In this moment, some countries have already launched nation-wide energy efficiency programs in water supply systems. For example, in the USA, in California region, energy efficiency must increase by 20% according to the California Water Plan. A similar percentage was approved in China and Sweden, where the wastewater treatment energy saving target is set for at least 20–30% from the total electrical energy requirement [1].

In the context of energy efficiency, the Watergy concept appeared. The term "Watergy" was coined by the Alliance to Save Energy in order to describe the strong link between water and energy in urban water distribution systems [7, 8].

In a statement given in the Alliance to Save Energy report [8] it is stated that between 2 and 3% of the total world electricity consumption is used for the pumping and treatment of water used by residential and industrial consumers.

For the water supply companies, these costs with consumed energy are the largest expenditure. These costs can reach up to 65% of the annual operating budget of a water supply company [9]. Also, water losses can augment the total cost. These losses may reach high levels in cities from countries where investments in water supply systems are very small. The examples can include countries from Latin America (Mexico, Brazil) or India where the losses are in the range of 40–50% [8].

The major steps involved in cities with clean water supply are presented in Fig. 20.2 [10]. It can be observed that in each step the electrical energy consumption is significant.

The electrical energy consumption from water supply systems is influenced by factors such as: topographical factors, geographical factors (water sources and consumers location), technical factors (pipe dimensions and configurations, rated powers of the pumps), treatment standards, and consumption factors (consumers' number, consumers' types, and consumption category). This varies from city to city, depending on the factors described above [7–9, 11–15].

Other factors that can also significantly affect electrical energy consumption depend on operational strategies and technology choices at national level [14].

An analysis of Fig. 20.2 can lead to the discovery of the technical factors that lead to a high electrical energy consumption: pump stations poor design, installation or maintenance, old pipes with high head loss, bottlenecks in the supply networks, excessive supply pressure, or inefficient operation strategies of various supply facilities [9, 11, 16–23].

There are many solutions for energy-saving in water supply systems, which do not require huge investment: from decreasing the volume of water pumps (e.g., adjusting pressure zone boundaries) to reducing the price of energy (e.g., avoiding peak hour pumping and making effective use of storage tanks) or increasing pump efficiency (e.g., ensuring that pumps are operating near their best efficiency point). These solutions can recover their costs in a time period ranging from a few months to 3 years [19–24].

In addition to these technical solutions, the companies can implement modern solutions based on the supervisory, control, and acquisition data systems. These

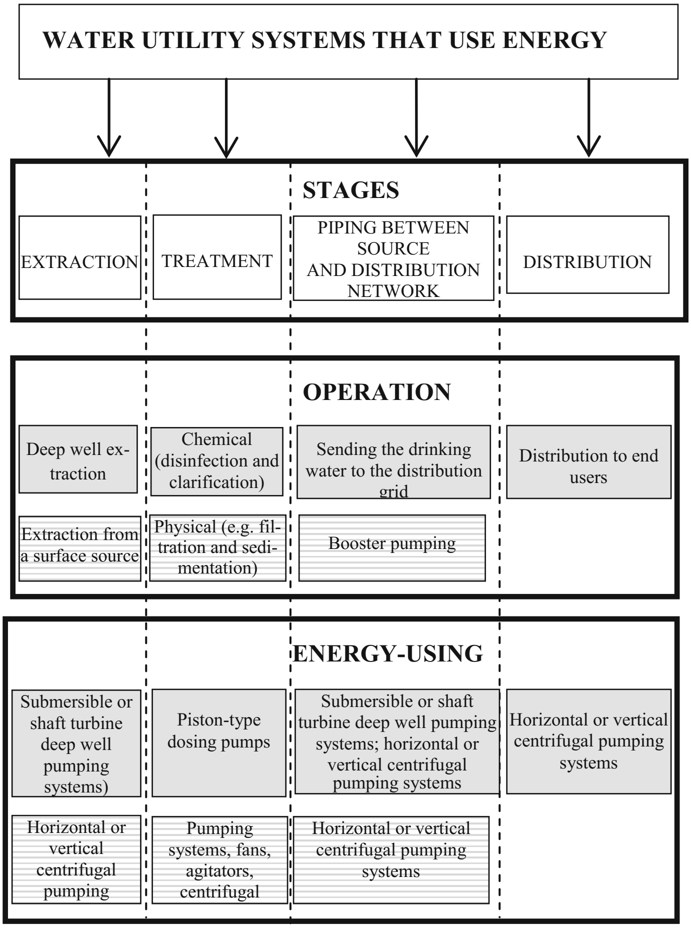


Fig. 20.2 Water utility system that use energy [10]

solutions use integrated information systems, able to realize complex monitoring and automation functions, implemented through customized software packages. Also, the SCADA systems can provide an efficient energy consumption management and can improve all activities from water supply systems. The motivation for introducing such systems is due the following factors [9, 17–23].

- the structure of the water supply systems became increasingly complex due to rising demands and the diversity of water sources;
- an aging infrastructure that leads to high operating costs;
- the availability and reliability of capable hardware components and software packages.

The energy efficiency policies in water supply systems must take account both of technical aspects which lead to infrastructure upgrades and information aspects that enable the supervisory and control of all activities through SCADA systems.

Energy efficiency policies correlated with of electricity prices growth have stimulated water supply companies, to find solutions to minimize their electrical energy consumption. An electrical energy consumption forecast with small errors can guarantee an efficient management plan leading to water supply of the customers at all times with minimum energy cost, considering the various technical constraints (reservoir levels, water pressure in the distribution system, water quality, and supply restrictions). Because in electricity markets, the price of electricity can change hourly, the water supply companies could move a part from the daily electrical energy consumption from a certain hour to other hours when the price is lower. But to achieve this target the water companies must have accurate electricity consumption forecasts. If the errors are high, they can lead to purchasing a supplementary energy amount with a high price from the day-ahead market or the intra-day market.

For electrical energy consumption forecasting, many methods were proposed in the literature. These methods are based mainly on classical approaches: heuristic, linear regression, or similar day. From these, the heuristic approach leads to the smallest errors (between 2.5 and 3%). In order to apply this type of method, a significant input database is required, with measurements: weather forecast, day type (working or weekend), knowledge of future events, and other historical operational data.

Out of these methods, few are taking into account the technical characteristics of the water hydrophore stations and the uncertainties that may exist in the system due to lack of measurements. The higher uncertainty degree leads to lower quality decisions made by the human operator. Therefore, the attempts to reduce uncertainty must be well substantiated. However, even with the reduction of the information uncertainty, considerable efforts and expenses are necessary. Thus, the targets of the forecasting process are to determine those parameters and characteristics which influence most the results and to improve the quality of the corresponding models and the relevant information.

In order to consider all these factors, the Artificial Intelligence (AI) techniques based on decision trees and clustering techniques are proposed for the electrical energy consumption forecasting in water distribution systems. These approaches are original because, to the author's best knowledge, no other methods proposed previously in the literature use decision trees or clustering for electricity consumption forecasting in water distribution systems. A comparative study using a database that contains records from 85 urban hydrophore stations belonging to a water distribution company from Romania will be presented. Using these approaches, the forecasting errors of the electrical energy consumption are better (approximately 1%) than the ones currently expected by water companies (approximately 3%). Thus, the results obtained demonstrate the ability of the proposed approaches to become an important step in ensuring an efficient management in the water distribution systems.

The remainder of this chapter is organized as follows. Section 20.2 presents a short review of AI techniques. Section 20.3 presents the implementation details of the proposed methods and the results of their testing on a real water distribution system from Romania. Section 20.4 contains the concluding remarks.

20.2 AI Techniques in Electrical Energy Consumption Forecasting

20.2.1 General Aspects

Today, information technology has drastically changed education, the nature of work, society, and decision-making processes all over, among other things. More information is becoming available, faster than ever before. The changing face of information technology and communication are proof that the economy, and in particular the utilities systems sector (power, water, gas, etc.), have entered a new era of fundamental change. The impact is being felt in the market, in the control and operation of the power, water or gas networks. Under these circumstances of momentous change, AI has the potential to play a more important role [25–27].

The main AI applications developed worldwide in the last years are:

- Load/energy forecasting,
- Optimal flow,
- Operating and expansion planning,
- Tariff selection,
- Alarm processing and fault diagnosis,
- Consumption control (power, water, gas, etc.).

The most used AI techniques are: Neural Networks, Fuzzy Logic, or Evolutionary programming. Other techniques based on machine learning are also used for classification and prediction. From these, the Decision Trees and Clustering algorithms are considered the most popular approaches. In the following, a summary of some theoretical issues relates to the AI techniques are presented.

20.2.2 Artificial Neural Networks

Artificial Neural Networks (ANN) are used increasingly in modelling and controlling industrial processes, gaining momentum in recent years in front of the conventional methods. Also called connectionist networks, ANNs reproduce certain functions of the human brain by simulating a model of the brain which is as yet extremely simplified. Because they are able to generalize linear statistic methods, ANNs are extremely well adapted for solving shape recognition problems, for which they give excellent results. Much of this interest has come about with the successful of real-word applications of ANNs and their ability to learn [26].

ANNs, as universal approximators, can give a systematic approach to the modelling of industrial processes, especially in those complex processes where solving the operation equations is very difficult using conventional mathematical methods. The most widely used ANN is the Multilayer Perceptron with the associated back propagation error learning, for minimizing the observed sum of squared errors over a given set of data.

In the literature excellent results were obtained using RNAs in some technical areas, but there have been also some problems that could not be successfully resolved using these techniques. From a recent analysis of applications in electric systems the main conclusions are:

- ANNs are extremely well developed to forecasting short-term consumption, especially if their architecture is optimized (for example, by using genetic algorithms);
- encouraging results have been obtained in modelling and controlling production units, alarm processing, and distribution networks control;
- ANNs are not the best choice for problems concerning network safety, production plan research and electrical networks planning.

The success in industrial process modelling can be firstly represented by the ability to extract information about the model structure and the relationships between its input and output data from trained network. Thus, these can become very important in validation of the models and the optimization and control of process. Furthermore, when the model of the process is not known, the objective is to learn as much about the system, and the obtained information to be used in the analysis of the process and in determining the disturbing factors that affect it.

20.2.3 Fuzzy Techniques

The fuzzy techniques (FT) were presented for the first time in the mid of the 60s by L. A. Zadeh. They had a huge success and have been used immediately in various fields. The broad development of mathematical theory, especially in areas of Possibility Theory, Fuzzy Control, ANN, and Pattern Recognition provided the basis for resolving different problems. Today, these techniques are implemented in many software packages and hardware components [25–28].

The advantage of these techniques is represented by the ability to work and perform calculations with uncertainties. These uncertainties can result from human observations based on descriptions and abstract information obtained by modelling the process. A classification of the uncertainty sources and the mathematical models corresponding to these is shown in Fig. 20.3.

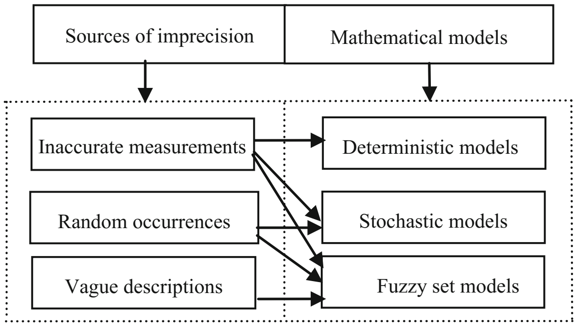


Fig. 20.3 Mathematical models for imprecision [28]

If for the first two sources of imprecision (inaccurate measurements, random events) appropriate mathematical models have been determined, it's not the same for the last source of imprecision (description language), although we use everyday expressions such as "the energy consumption is small/big" etc. Most linguistic descriptions such as Small, Medium, and High are fuzzy in nature. These vague descriptions are part of modelling process and the algorithm.

The decision maker must make the difference between the various linguistic classes, e.g., when these classify the operation regimes of the supervised system according to certain operational characteristics [25–28]. The linguistic description is a part of the modelling process and the operators/experts' language. The experts' knowledge is usually translated in natural language. Therefore, it requires a way of encoding and manipulating knowledge expressed in natural language. This approach could be based on the use of linguistic variables and nuanced sets. Moreover, the experts' knowledge is affected by the ambiguity and imprecision of the natural language. The fuzzy techniques try to resolve these problems through a coherent modelling of the uncertainty, while providing the possibility to operate with such models.

The fuzzy sets are conceptually linked to the uncertain intrinsic properties of the model. A fuzzy set presents the following:

- an area that represents the fuzzy population;
- the semantic representation associated fuzzy population;
- a function that defines the degree of membership of an element from the fuzzy population to the set.

Unlike bivalent logic, in which membership degree of an element to the set can have two values, in fuzzy logic it can have any real value in the [0, 1] range.

The major advantage of this logic is that it operates with sets which by definition are characterized by continue and normal membership functions.

Fuzzy logic can be used with success in real applications, where there are many uncertainties in mathematical models (both in the objective function and constraints). But, these uncertainties can be modelled using the FTs. Designed to work with uncertain data, the fuzzy techniques can be applied in most industry sectors process control. Thus, the control and regulation of industrial processes are areas where fuzzy techniques found a breeding ground for their application.

20.2.4 Evolutionary Programming

Evolutionary programming (EP) searches an optimal solution by evaluating a set of candidate solutions, over a number of generations or iterations. In each iteration, another set is formed, starting from the existing set, using the mutation operator. Thus, a new solution is obtained, by disturbing each element of the existing solution by a random value. The optimality degree of o each candidate solution or individual is measured by its fitness, which can be defined based on the objective function of the problem. In this case, if a competition scheme is used, the individuals in each population compete with each other. The winning individuals will form a resultant set, which is regarded as the next generation [26].

The optimization process must be conducted so that the competitive scheme should favour the survival of the better solutions.

20.2.5 Decision Trees

Machine learning represents an important AI research domain. This is due to the following two aspects:

- the ability to learn represents a characteristic of an intelligent behaviour;
- the understanding of the intelligence should be done through the understanding of the learning model [29].

In the last years, the approach based on decision trees (DTs) was used increasingly in resolving classification and prediction problems. DTs use a flow chart-like tree structure. From the view point of data-mining tools, decision trees

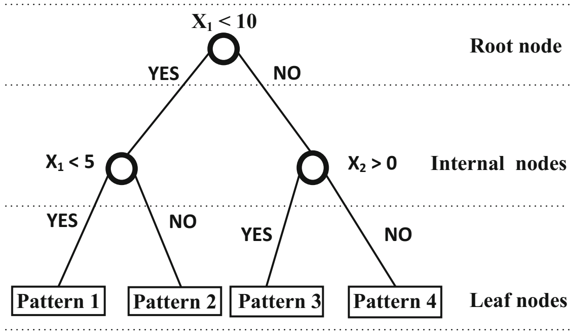


Fig. 20.4 The structure of a decision tree

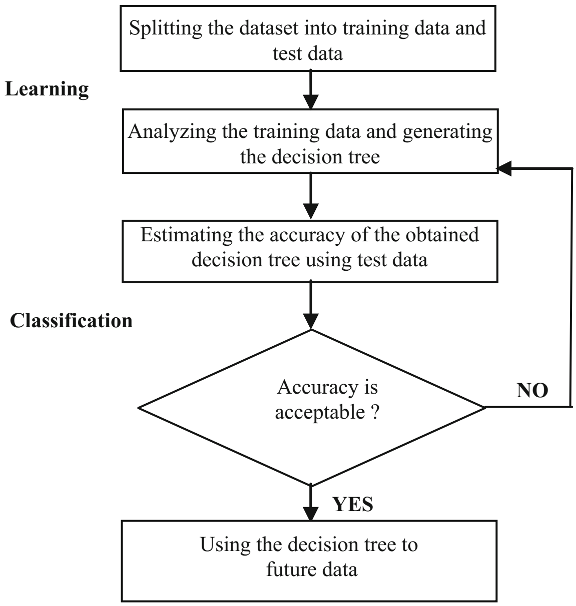
can extract useful information contained in large databases, which can be used in decision-making processes.

A decision tree can have three nodes categories: root, internal, and leaf. The structure of a decision tree can be seen in Fig. 20.4.

If a decision tree is viewed as a logical model, this can indicate a relationship between a target variable and a set of predictor variables [29, 30]. The classification process will take place in function by separation rules relating to the numerical or categorical values of the explanatory variables. Each node will have proper classification rules derived from a mathematical process that will minimize the impurity of the resulting nodes, using the available learning set. Using these classification rules, the leaf nodes will be finally reached [30, 31]. But, for the leaf nodes, the separation rules are not applied just like internal nodes. With the help of these nodes, the probability of each obtained pattern can be estimated. From the structure of a decision tree it can observed that the root and internal nodes indicate a binary split test on an attribute. The classification process depends mainly by first node (root) that contains all learning set.

Also, the leaf nodes represent the results of the classification process so that they are labelled as targets. From the point of view of the algorithm, in order to generate a decision tree, two steps should be covered, namely learning and classification [30]. These steps are shown in Fig. 20.5. The available data must be divided in two sets: the first set (the learning set) is used in the learning process in order to build the structure of the tree, while the second set (the testing set) has the role to assess the generalization capability of the tree using new input data. The sizes of the two sets are different and depend on the number of elements from the database. The

Fig. 20.5 The flow-chart of decision tree generation [30]



ratio between the two sets can vary between 90/10 and 60/40%. In the literature, the most used method used to generate a decision tree is Classification and Regression Trees (CART) method [32, 33].

This method uses the data recorded in a database during a period of time to construct the structure of the decision tree. Also, a number of patterns must be introduced from the beginning of the learning process. These patterns can be obtained using various techniques (for example, clustering techniques).

Finally, the information resulted from the learning and classification processes can be arranged in a matrix named confusion matrix [34, 35]. Based on this matrix, the performance of the classification process can be assessed.

20.2.6 Clustering Techniques

Clustering techniques are based on the arranging/organization of an objects/elements set into groups (clusters) with similar characteristics. With the help of these techniques, information can be extracted easily from large databases, such that the decision maker understands the existing relationships between recorded data.

In the clustering process, there are some important steps which the decision maker should take into account. The steps are presented in Fig. 20.6 [25, 36, 37].

Feature selection In this step, the relevant characteristics corresponding to the elements/objects from the database will be chosen for the clustering process.

Clustering algorithm According to the choice of the clustering algorithm, the elements/objects from the database can be separated either into clearly defined groups, either into fuzzy groups, where each element has a varying degree of membership to each resulted group.

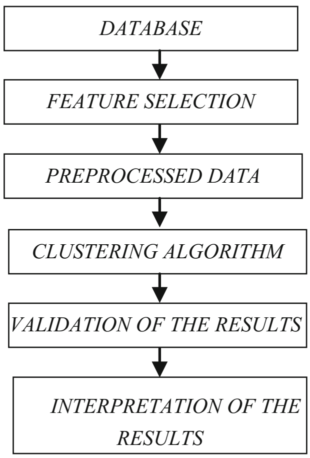


Fig. 20.6 The steps of a clustering process [37]

Validation of the results In this step, an evaluation of the clustering process is made. Usually, an optimization criterion is used. This analysis must be objective and it is performed to determine if the results are grouped as accurately as possible.

Interpretation of the results The decision maker will compare the obtained results with other methods. In accordance with the benchmarking, it will reach a conclusion and it will take a decision.

In literature, two clustering methods categories were developed: hierarchical clustering and *K*-means clustering.

Hierarchical clustering In the hierarchical clustering, two type of methods can be identified: agglomerative methods, which perform a succession of unifications of the n elements from the initial database into clusters, and divisive methods, which separate the n elements gradually into finer clusters. The agglomerative techniques are easier to implement. The clustering process can be interpreted through a two dimensional diagram (dendrogram). This explains the unifications/splits which take place in each step.

The hierarchical clustering can be used for databases with a number up to several hundred elements. The methods from this category are differentiated by how the similarity between clusters is evaluated. The similarity is determined by computing the Euclidean distance. Single linkage clustering (connectedness or minimum method), complete linkage clustering (diameter or maximum method), average linkage clustering, and the centroid method can be used.

K-means clustering In this method, the elements are grouped in K clusters (K is a positive integer number, introduced a priori) based on the similar characteristics of the elements from the database. Initially, the c_k centroids of each cluster $C_k \subset S$ (the set of clusters), $k = 1, \ldots, K,$ are chosen randomly. An element x_i will be assigned to a centroid c_k if the distance between them $d(x_i, c_k)$ is the minimum of all distances between this element, x_i, and the other centroids. The algorithm has as objective the minimization of a squared error function.

The objective function E will minimize the distance between each element and the centroid of the cluster to which the element will be assigned.

$$\min(E) = \min\left(\sum_{k=1}^{K}\sum_{x\in C_k} d\left(x^i,\, c_k\right)\right) \tag{20.1}$$

where:

c_k—represents the centroid of cluster C_k ($c = \{c_1, \ldots, c_k\}$, $C = \{C_1, \ldots, C_K\}$, $C_k \subset S$, $k = \{1, \ldots, K\}$);
$d(x_i, c_k)$ is the Euclidean distance computed between the element $x_i \in$ S, $i = 1, \ldots,$ N, and the centroid c_k

In the first step, a set of K centroids will be initiated randomly. Then, the algorithm tries to assign each element from the dataset to a cluster when distance between it and the centroid is the shortest. Further, the new positions for centroids will be recomputed. The iterative process will stop when the positions of centroids will not change. After that, the decision maker will verify the validation of the obtained results.

The validation and evaluation of the results must highlight the quality of the clustering process. This quality will be evaluated through an analysis that will consider the clusters' density, sizes and, separation degrees. In the literature, the main approaches used for the cluster validation are [38]:

- *external tests*—the results obtained in the clustering process will be compared with the results obtained using other databases.
- *internal tests*—the quality of clustering process is evaluated based on the input data to validate the division mode of the clusters.
- *relative tests*—the final results are compared using the same clustering algorithm, but with different parameters.

The most used tests in clustering analysis are the internal ones. Many of these are based on a Silhouette Global Index (*SGI*). In the determination of this index, the following silhouette widths will be evaluated successively: for each sample, for each cluster, and for the entire database. The last width represents the average silhouette that will be used for validation. Also, with its help, the optimal number of clusters will be determined.

$$SGI = \frac{1}{K}\sum_{k=1}^{K} S_{C_k} \tag{20.2}$$

where:

S_{Ck}—the silhouette width for cluster C_k. This is evaluated with the formula:

$$S_{C_k} = \frac{1}{n_j}\sum_{i=1}^{n_j} s_i \tag{20.3}$$

s_i—the silhouette width for i-sample. It can be calculated with the formula:

$$s_i = \frac{b_i - a_i}{\max\{b_i, a_i\}} \tag{20.4}$$

a_i—the average distance between the element i and elements of the same sample j;
b_i—the minimum average distance between the element i and elements in class closest to class j.

An interpretation of the Silhouette Global Index is given in [39, 40]:

- 0.71–1.0: The formed structure is strong. The results can be trusted.
- 0.5–0.7: An acceptable structure of the clusters was obtained.

- 0.26–0.5: The obtained structure is very poor. Low confidence degree.
- <0.25: The degree of confidence in results is zero. No structure exists.

Data validity represents the most important step in clustering analysis. Based on its results, the decision maker can assess the confidence degree in the relationships established between the input data.

20.3 Electrical Energy Consumption Forecasting in Water Distribution Systems

20.3.1 Clustering-Based Approach

In this section, an extension of the clustering based profiling techniques in the area of water distribution systems for electricity consumption forecasting is presented. The *K*-means clustering algorithm will be used for determining the representative loading profiles (RLP) of hydrophore stations, and these, together with a statistical approach will forecast the electrical energy consumption from the water distribution systems [41, 42].

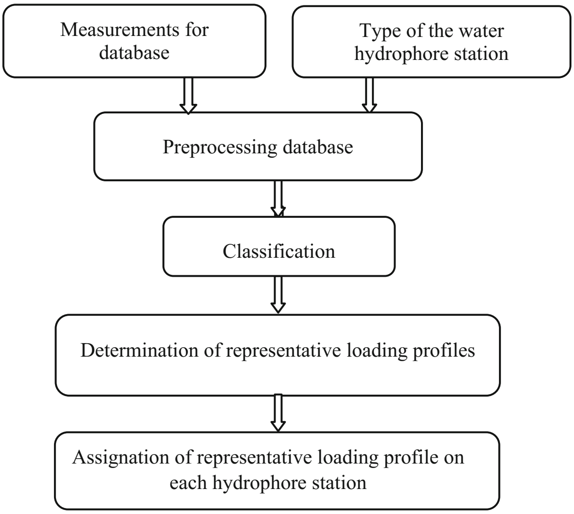
By knowing the loading profiles of the hydrophore stations, the distribution companies can forecast easily the water demand for a certain zone. Based on this forecast, these can establish the best energy efficiency strategies. The load profiling can represent a more realistic approach for energy consumption forecasting because the loads are not monitored in all hydrophore stations at each hour. Thus, a representative loading profile will be attached for each hydrophore station, determined based on the different operation conditions (season, type of day, types of supplied consumers, energy consumption, water consumption etc.). The sampling interval can be 10, 15, 30 or 60 min. In these conditions, the loading profiles can have 144, 96, 48 or 24 load values. The forecasted profile of a hydrophore station is obtained using a representative loading profile and its daily (monthly, yearly, depending the case) electric energy consumption.

20.3.1.1 Load Profiling for Water Hydrophore Stations

In the profiling process, the clustering techniques can be used for grouping the loading profiles recorded in hydrohpore stations (in various measurement campaigns) in coherent clusters based on their the similarities. Finally, a representative loading profile will be attached to each cluster. This profile is achieved if all loading profiles belonging to the respective cluster are averaged.

Each RLP is represented by a vector $x_i = \{x_{ih}, h = 1, \ldots, T\}$ for $i = 1, \ldots, K$, and the comprehensive set of RLPs is contained in the set $P = \{x_i, \mathrm{i} = 1, \ldots, K\}$. The time scale along the day is partitioned into T time intervals of duration Δt_h, for $h = 1, \ldots, T$. Hourly values are used in this chapter. The variables used in the

Fig. 20.7 The flow-chart for obtaining of the representative load profiles [41]



calculations are assumed to be represented as constant (average) values within each time interval. The clustering process generates K clusters corresponding the hydrophore stations. In the final step, each hydrophore station will have a RLP attached, based on its loading characteristics. All steps of the algorithm are presented in Fig. 20.7 [41]. Each step will be detailed in the following.

Measurements After the measurement campaigns performed in each hydrophore station from the water distribution system during a certain time period (day, month, season, or year), a set of loading profiles will be obtained. The profiles can have a certain number of values, according to the used sampling step. After that, all loading profiles will be collected into a database.

Data cleaning and pre-processing Due to the following factors: measurements carried out over a long period of time, high number of recorded data, and a spread of the hydrophore stations on a large geographical area, the recorded data can be affected by communication errors and failures of the measurement equipment. These problems will impact the quality of recorded data. In order to remove the missing or irregular values from the database of, data pre-processing is required.

Classification In the classification process, the clustering methods will be used. Because the database can have a high number of profiles, the best suited algorithm is K-means. In order to use this algorithm, an important aspect is represented by the normalization of the loading profiles before starting the clustering process. A suitable factor is the energy consumption over the analysed time. This factor is used because each hydrophore station has an energy meter with classical measurement technology.

Determination of RLPs For each cluster, a representative loading profile will be determined by averaging all its assigned profiles. This profile will characterize the operation regime of all hydrophore stations from a certain cluster from hourly load and energy consumption viewpoints.

Assignation For the characteristic regime of a hydrophore station, a representative loading profile will be assigned.

20.3.1.2 Electrical Load and Energy Consumption Forecasting

The electrical load and energy consumption forecasting of a water distribution system can be made with an improved simulation method which uses the representative loading profiles of the hydrophore stations. The method can be used if the following assumptions are considered [41, 42]:

- the hourly average load is approximately proportional to the electrical energy consumption. This load represents a cluster of hydrophore stations from a water distribution system during the analysed time period.
- the hourly average loads have a normal statistical distribution.

If these assumptions are satisfied, the hourly load in a water distribution system can be forecasted with:

$$P_h^S = \sum_{i=1}^{K} n_i W_{av}^{C_i} p_h^{C_i} + \sqrt{\sum_{i=1}^{K} n_i \left(W_{av}^{C_i} \sigma_h^{C_i}\right)^2}, \text{(kW)}, \ h = 1, \ldots, 24 \tag{20.5}$$

where:

P_h^S—the forecasted load at the hour $h = 1, \ldots, 24$, (kW);
K—the number of clusters resulted from the clustering process for the hydrophore stations;
n_i—the number of the hydrophore stations from cluster C_i;
W_{av}^{Ci}—the daily average energy consumption of the hydrophore stations belonging to cluster C_i, $i = 1, \ldots, K$, (kWh);
p_h^{Ci}—the average hourly loading factor of the hydrophore stations belonging to cluster C_i, (kW/kWh);
σ_h^{Ci}—the standard deviation of the load from hydrophore stations belonging to cluster C_i, (kW/kWh).

The hourly loads from a particular hydrophore station can be estimated with the help of the loading factors p_h^{Ci} belonging the cluster C_i which contains the station, and of the real energy consumption recorded in that station. For the electrical energy consumption forecast, the following equation can be used:

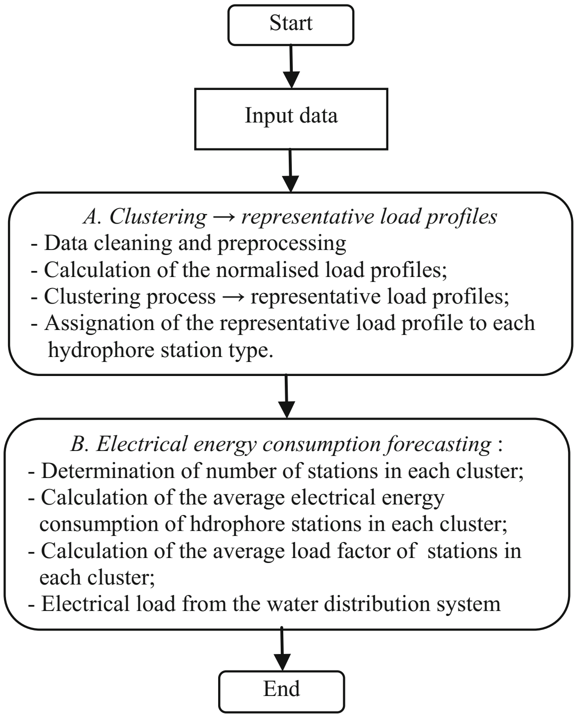


Fig. 20.8 The flow-chart of the proposed algorithm

$$W^S = \sum_{h=1}^{T} P_h^S, \ (kWh) \tag{20.6}$$

20.3.1.3 The Electrical Energy Consumption Forecasting Algorithm

The algorithm adopts a two stage procedure presented in Fig. 20.8:

- the use of a clustering technique (K-means method) in the first stage to deal with the time evolution of (normalized) electrical load patterns and to determine a subset of representative load profiles to be processed in the second stage. At each iteration, the clustering outcomes simplify the process of selecting a relatively small number of profiles corresponding to the hydrophore stations;
- the use of load simulation in the second stage. This approach is based on a statistical method which assumes that the hourly loads have a normal statistical distribution.

A. Determination of Representative Load Profiles

The electrical energy consumption is used in the normalization process of the characteristic loading profiles. This normalization factor is used because each hydrophore station has measurement meters that record continuously the electrical energy consumption. In the next step, the algorithm will determine the optimal numbers of clusters with the *K*-means clustering and an internal test that uses the silhouette global index to validate the classification. Subsequently, the representative loading profiles (RLP) will be obtained through an averaging process of all normalized profiles inside each cluster. In the final step, the representative loading profiles will be assigned to all hydrophore stations based on their type.

B. The Electrical Load Forecasting

Using information from the clustering process (hourly average values, average energy consumption and standard deviation of the distribution of load for each hydrophore station from the cluster C_i, $i = 1, \ldots, K$ and Eq. 20.5, the hourly loads of the analyzed water distribution system will be obtained.

The forecast accuracy depends on the accuracy of information recorded in the database. If the real value of the load is known, the forecasting error can be assessed using the mean absolute percentage error (MAPE):

$$MAPE = \frac{\sum_{h=1}^{T} \frac{\left|P^S_{h,real} - P^S_{h,f}\right|}{P^S_{h,real}}}{T} \cdot 100 \quad (20.7)$$

In Eq. 20.7, the variables $P^S_{h,real}$ and $P^S_{h,f}$ are the real and forecasted values for the load of a water distribution system at the hour h.

MAPE is a dimensionless quantity. This can highlight the comparison with other results obtained using other databases in order to establish the accuracy of the algorithm.

Table 20.1 The technical data of hydrophore stations

Type	Number of stations	Rated power (kW)	Rated water flow (m^3/h)
I	3	3×2.2	24
II	6	4×2.2	32
III	11	4×4	64
IV	13	3×5.5	96
V	33	4×5.5	128
VI	7	4×7.5	128

20.3.1.4 Results

The proposed method was tested to forecast the electrical energy consumption in a real water distribution system with N = 85 hydrophore stations. For each hydrophore station, the technical characteristics and the load patterns are known. The technical data is presented in Table 20.1.

The primary characteristics of a water distribution station are: the daily water flow, the pressure growth, the number and nominal power of force pumps located in these distribution substations. The stations are equipped with 3 or 4 pumps working in parallel, with rated powers in the 2.2–7.7 kW range. The input data used in the analysis is based on a measurements campaign performed during a year. The recorded parameters are represented by the consumed active power, the water flow, and the pressure of the water in repression and aspiration zones of the pumps in each hydrophore station, recorded using a 5 min sampling rate.

In the pre-processing step, 12 water hydrophore stations were eliminated due to missing technical characteristics (the daily water flow, pressure growth and rated powers of force pumps located in the stations). Then, the remaining data was recorded in a database for each of the remaining stations (73 stations).

In the analysis, one of the technical characteristic used in the clustering process is the pressure growth of the water. This parameter represents the difference between the pressures of the water in the repression and aspiration zones of the pumps:

$$\Delta p = p_{rep} - p_{asp} \tag{20.8}$$

where:

p_{rep}—the pressure of the water in the repression zone of the pumps;
p_{asp}—the pressure of the water in the aspiration zone of the pumps.

The consumed active powers are represented by loading profiles of the water hydrophore stations. The profiles were analysed only for the day when the maximum load in the water distribution system was registered. The time interval is defined by taking hourly steps within a day, that is, T = 24 and Δt_h = 1 h.

In the first step of the clustering process, normalization is applied to all loading profiles from the database, using the daily electrical energy consumption as normalizing factor.

Table 20.2 Clustering process results

Cluster	Stations		Type of stations						W_{med}
	(no.)	(%)	I	II	III	IV	V	VI	(kWh)
C_1	8	12.33	3	5	1	–	–	–	28.58
C_2	24	32.88	–	–	4	2	11	7	41.42
C_3	41	54.79	–	1	6	11	22	–	32.35
Total	73	100	3	1	11	13	33	7	34.11

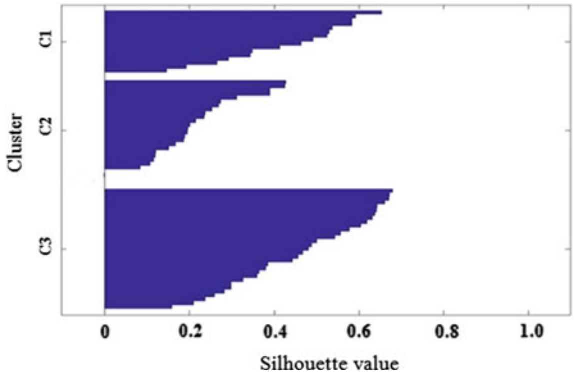


Fig. 20.9 The silhouette plot for $K_{opt} = 3$

In the next step, the maximum number of clusters will be determined using formula:

$$(K_{max} = \sqrt{N} \approx 9) \tag{20.9}$$

Further, the K-means algorithm will be applied successively for K varying in the range between 2 and K_{max} in order to determine the optimal number of clusters (K_{opt}). For each result, an internal test based on the silhouette global index (SGI) will verify the classification quality. The optimal solution will correspond to the classification that will obtain the highest value for the silhouette global coefficient. In our case, K_{opt} is 3 (SGI = 0.52).

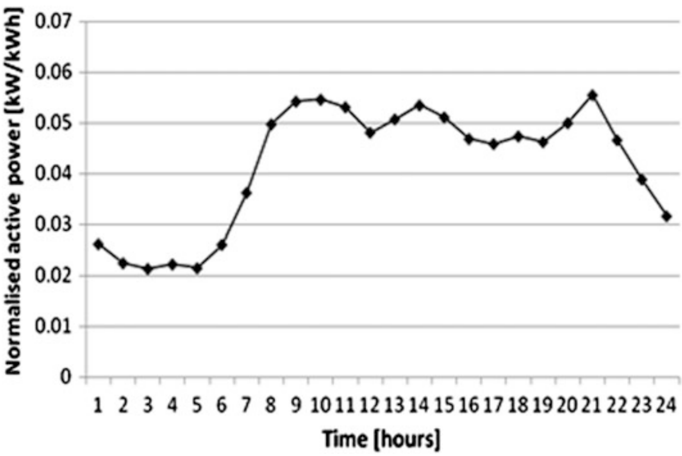


Fig. 20.10 Representative load profile for C_1

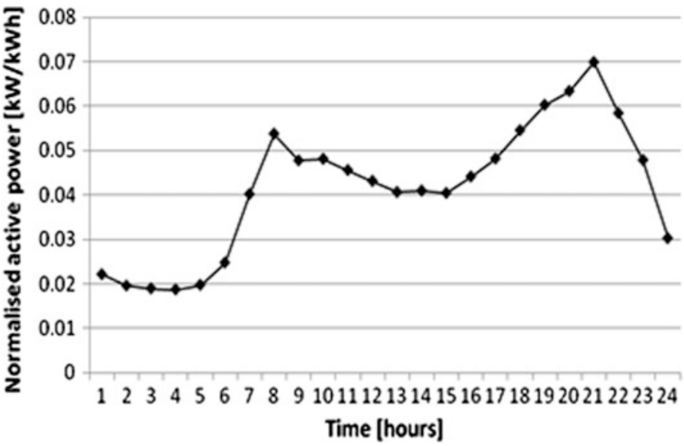


Fig. 20.11 Representative load profile for C_2

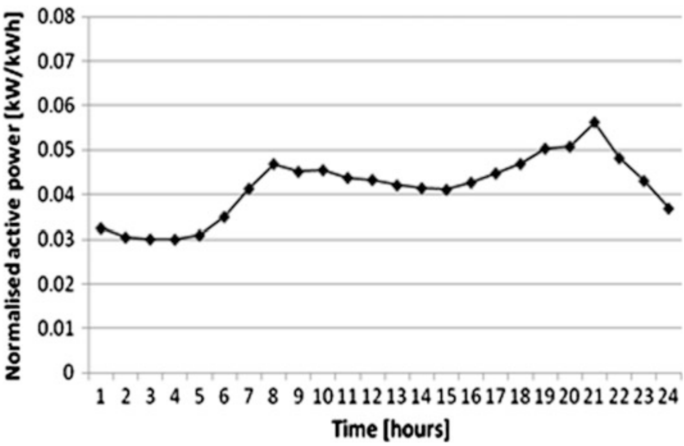


Fig. 20.12 Representative load profile for C_3

The representation of the silhouette coefficient in the optimal variant is given in Fig. 20.9. The results of the first stage of the clustering procedure are presented in Table 20.2.

The analysis of the results highlights that the clusters C_2 and C_3 include a high number of loading profiles (approximately 88%). This emphasizes that the majority of the hydrophore stations have a similar operation mode. Concerning the technical characteristics, the stations from cluster C_1 have a rated power lower than 8.8 kW, and the stations from clusters C_2 and C_3 have a higher rated power, between 16 and 30 kW. In the final step of the clustering process, RLPs will be built using the

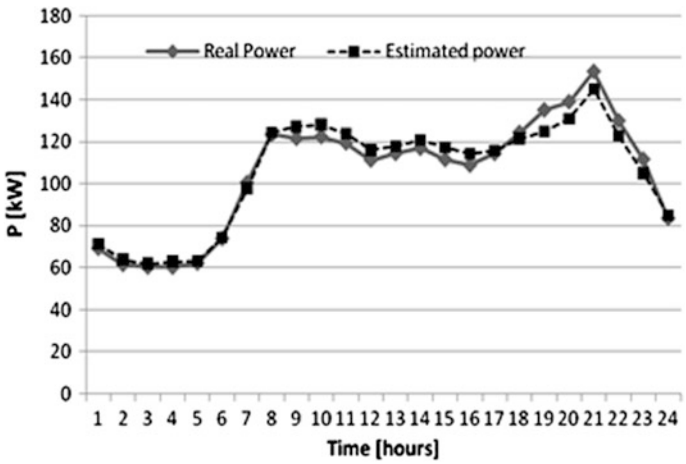


Fig. 20.13 The forecasted and real loads

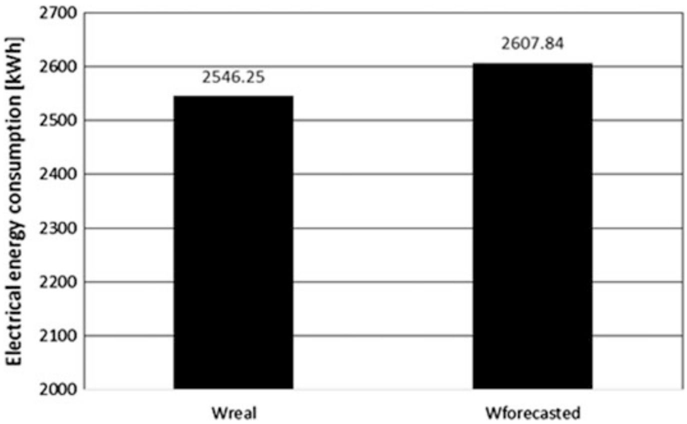


Fig. 20.14 The real and forecasted electrical energy consumption

averaging method of the normalized loading profiles. The representative loading profiles are plotted in Figs. 20.10, 20.11 and 20.12.

In the second stage of the study, the total load in the water distribution system at each hour $h = 1, \ldots, 24$ will be computed considering the results from the clustering process (Table 20.2) and the reasoning presented in Sect. 20.3.1.2. The forecasted and real values are plotted in Fig. 20.13. The forecasting errors have the following values: maximum error—4.79%, minimum error—1.37%, and mean absolute percentage error—3.11%. Finally, the electrical energy consumption was forecasted using Eq. 20.6 and the results are presented in Fig. 20.14. The forecasting error obtained with the proposed approach is 2.42%. This value can be

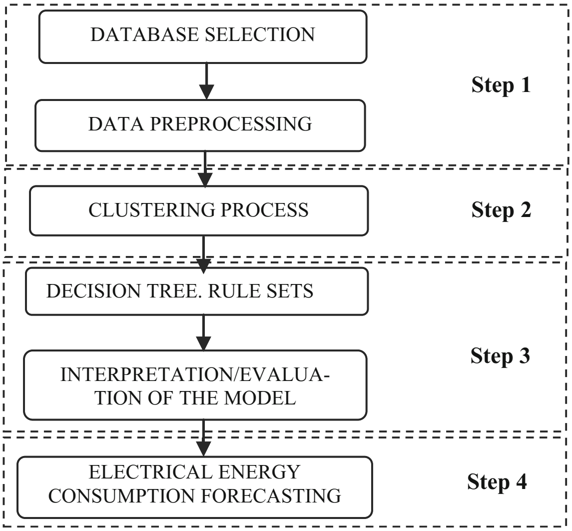


Fig. 20.15 The flow-chart of the proposed method

considered small, if the arbitrary inherent behaviour of the consumers in a water distribution system is taken into account.

20.3.2 *Decision Trees-Based Approach*

20.3.2.1 The Electrical Energy Consumption Forecasting Algorithm

In this section, a decision trees-based methodology for the daily electrical energy consumption forecasting of a water distribution system is presented. The algorithm consists of four steps, as presented in Fig. 20.15.

Firstly, a database of possible primary variables which characterize the behaviour of the water hydrophore stations from the analysed system is built. The choice of primary variables is based on the decision maker's experience. The chosen variables are: the daily water flow, pressure growth and rated powers of the pumps located in the stations. In this step, data preprocessing is also performed. This aspect is very important because preprocessing detects/corrects/removes improper data. In the following step, the water hydrophore stations will be classified based on their daily water flow, pressure growth, and electrical energy consumption, using the clustering techniques. The *K*-means algorithm will be used in the clustering process to obtain a finite number of significant patterns with similar behaviour.

The third step refers to building the decision tree. This tree will explain the behaviour of each station with regard to the variables chosen in the second step. Thus, by entering the significant variables that explain the technical and operational characteristics of the water hydrophore stations, based on the rules that characterize the structure of the decision tree, the daily electrical energy consumption will be forecasted. The decision tree will be verified and validated using a test set.

In the fourth step, the daily electrical energy consumption forecasting in the water distribution system is made using a statistical approach which must satisfy the assumptions made in Sect. 20.3.1.2. The used expression in the proposed approach is:

$$W^S = \sum_{i-1}^{K} n_i W_{med}^{C_i} + \sqrt{\sum_{i=1}^{K} n_i \left(\sigma_W^{C_i}\right)^2} \tag{20.10}$$

where:

W^S—the daily electrical energy consumption of the water distribution system, (kWh);
n_i—the number of the hydrophore stations from cluster C_i;
W_{med}^{Ci}—the daily average energy consumption of the hydrophore stations belonging a cluster C_i, (kWh);
σ_w^{Ci}—the standard deviation of daily energy consumption of the hydrophore stations from the cluster C_i, (kWh);
K—the total number of clusters.

20.3.2.2 Results

The proposed methodology was applied on same database with 85 water hydrophore stations presented in Sect. 20.3.1.4. The daily electrical energy consumption is the target variable and the technical characteristics are the input variables. Thus, the technical variables (the daily water flow, pressure growth, and rated powers of force pumps) were recorded for each hydrophore station. In the pre-processing step, 12 water hydrophore stations were removed due to some incorrect data corresponding to the daily water flow or electrical energy consumption. Then, the data for each of the remaining stations (73 stations) were stored in a database

For generating the decision tree, CART method is used. In the first CART stage, the number of patterns must be known. The patterns can be obtained with the clustering techniques. Using these techniques, a more precise model was obtained that, has ability to adapt easily to any other input data set.

The hydrophore stations were classified in five significant clusters (patterns) by their daily water flow (*WF*), pressure growth (*PG*) and daily electrical energy consumption (*W*) on the basis of the *K*-means clustering algorithm. The representation of these clusters (patterns) is given in Fig. 20.16. Table 20.3 presents the statistical parameters (average and standard deviation) for each input variable corresponding to all clusters.

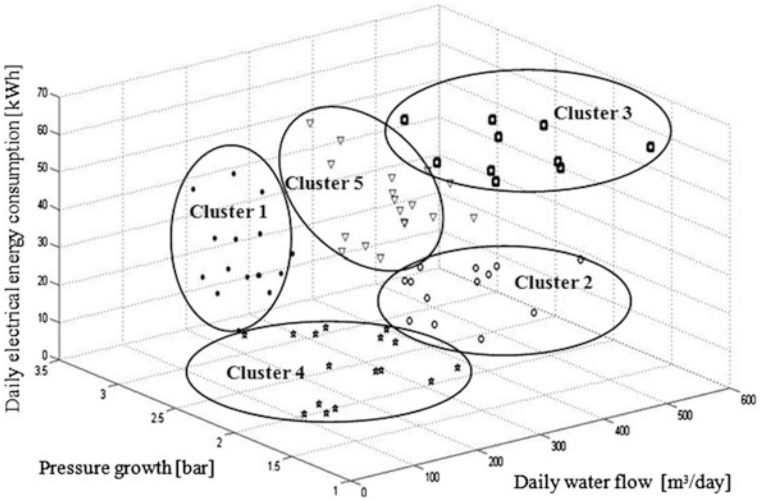


Fig. 20.16 The grouping of the the water hydrophore stations

Table 20.3 The results of the clustering process

Cluster	Number of stations	Daily water flow (m^3/day)		Pressure growth (bar)		Daily electrical energy consumption (kWh)	
		m	σ	m	σ	m	σ
1	16	166.97	39.27	2.83	0.27	29.73	7.19
2	14	314.00	58.68	1.71	0.25	29.20	5.14
3	10	463.37	44.08	2.40	0.22	54.72	5.54
4	15	143.37	59.33	1.78	0.30	14.86	4.52
5	18	329.13	43.07	2.46	0.26	41.90	6.92

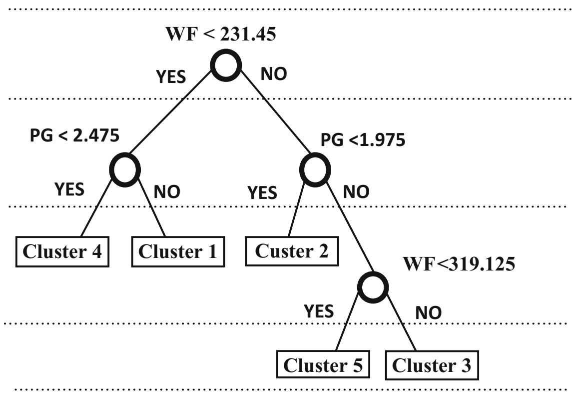


Fig. 20.17 The structure of the decision tree

Table 20.4 Rule set corresponding to decision tree

Rule 1	IF WF < 231.4 and PG < 2.47 THEN W—Cluster 4
Rule 2	IF WF < 231.4 and PG > 2.47 THEN W—Cluster 1
Rule 3	IF WF > 231.4 and PG < 1.97 THEN W—Cluster 2
Rule 4	IF WF > 231.4 and PG > 1.97 and WF < 391.12 THEN W—Cluster 5
Rule 5	IF WF > 231.4 and PG > 1.97 THEN W—Cluster 3

Table 20.5 The Confusion Matrix for the classification process (learning database)

Actual		Classified					Correctness rate
		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	
Cluster 1	10	10	0	0	0	0	1.000
Cluster 2	13	0	11	0	1	1	0.846
Cluster 3	2	0	0	2	0	0	1.000
Cluster 4	15	0	0	0	15	0	1.000
Cluster 5	10	0	0	1	0	9	0.900

In the next stage, the algorithm uses 2/3 of the database for learning (50 water hydrophore stations) and of the 1/3 of database (23 water hydrophore stations) for testing. The structure of the decision tree corresponding to the learning database is shown in Fig. 20.17. The internal nodes have test rules for one of the variables (PG or WF) and the leaf nodes represent the final results of the decisions taken.

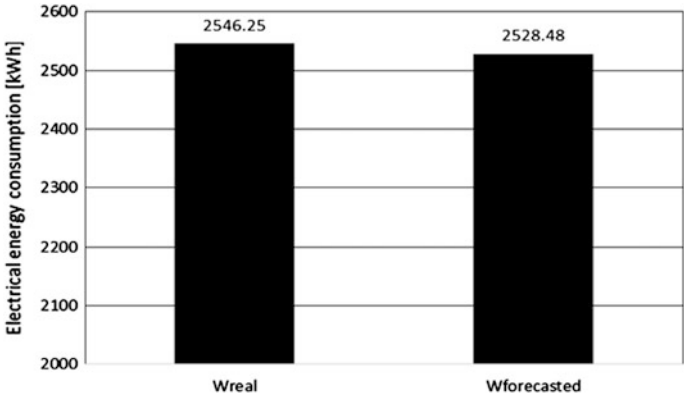


Fig. 20.18 The real and forecasted electrical energy consumption

Table 20.6 The Confusion Matrix for the classification process (test database)

Actual		Classified					Correctness rate
		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	
Cluster 1	5	5	0	0	0	1	1.000
Cluster 2	2	0	2	0	0	0	0.846
Cluster 3	7	0	0	7	0	0	1.000
Cluster 4	1	0	0	0	1	0	1.000
Cluster 5	8	0	0	0	0	8	1.000

The rule set that characterizes the decision tree obtained is presented in Table 20.4. The confusion matrix corresponding to the classification process, for the learning database, is presented in Table 20.5.

The analysis of data from Table 20.5 indicated that all stations from Clusters 1, 3, and 4 were classified correctly, and only 4 stations (3 stations from Cluster 2 and 1 station from Cluster 5) were misclassified. Thus, it can be seen that for the learning database, the accuracy rate is 0.94.

This decision tree structure was verified on the test database. The results are presented in Table 20.6, corresponding to the confusion matrix. Only one station from Cluster 1 was misclassified, so that the accuracy rate is 0.96.

In the final stage, using information from the classification process (Table 20.3 and Eq. 20.10) the daily electrical energy consumption in the analysed water distribution system was obtained, as in Fig. 20.18. The forecasting error is 0.7% (17.77 kWh), which is better than the expected one by water companies (2–3%) [43].

20.4 Conclusions

A proper evaluation of the existing linkage between the two important components from human life (water and energy) can lead to saving both money and resources by water distribution companies. The investigation methods, which take in account the aspects of the two components, can lead to increasing energy efficiency in order to minimize electricity consumption in water distribution systems.

In this study, two approaches based on AI techniques (clustering and decision trees) for forecasting the amount of energy consumed in water use are presented. These approaches are tested on a real water distribution system from Romania.

The first approach is based on two stages in which the *K*-means clustering algorithm and a load simulation technique are exploited the electrical energy consumption forecasting. The *K*-means clustering algorithm was used in the profiling process to obtain the representative loading profiles of the hydrophore stations. The electrical energy consumption was forecasted using statistical techniques. The mean absolute percentage error in the forecasting process was 2.42%.

The second approach is based on a set of input data processed through the *K*-means clustering algorithm, and then on a decision tree through with which the daily electrical energy consumption for one or more hydrophore stations, depending on some input variables (the water flow and the pressure of the water in repression and aspiration zones of the pumps) was estimated with a very good accuracy. The forecasting error for electrical energy consumption was 0.7%.

Using these approaches, the forecasting errors of the electrical energy consumption are very close or even better than the ones expected by water companies (between 2 and 3%).

In both case studies, the results obtained demonstrate the utility of the AI techniques in resolving problems in water distribution systems operation. The solutions proposed can improve decision making, in order to ensure an efficient management in the water distribution systems.

References

1. Olsson G (2012) Water and energy nexus. In: Encyclopaedia of sustainability science and technology. Springer, New York
2. World Economic Forum. Energy vision update. Thirsty energy: water and energy in the 21st century. http://www3.weforum.org/docs/WEF_WaterAndEnergy21stCentury_Report.pdf
3. Stillwell A, King C, Webber M, Duncan I, Hardberger A. Energy-water nexus in Texas. Environmental defense fund. http://www.circleofblue.org/waternews/wp-content/uploads/2010/08/9479_Energy-WaterNexusinTexasApr20091.pdf
4. World Energy Council. Water for energy. https://www.worldenergy.org/wp-content/uploads/2012/10/PUB_Water_For_Energy_2010_WEC.pdf
5. Martin Delgad A (2012) Water footprint of electric power generation: modeling its use and analyzing options for a water-scarce future. Master of Science in Technology and Policy Thesis, Massachusetts Institute of Technology
6. Farley J, Gaddis E (2007) An ecological economic assessment. In: Restoring natural capital: science, business and practice. Island Press, Washington D.C.
7. European Commission. Directorate General for Environment. http://ec.europa.eu/environment/water/index_en.htm
8. Alliance to Save Energy. What watergy involves. https://www.ase.org/resources/what-watergy-involves
9. Boulos PF, Wu Z, Orr, CH, Moore M, Hsiung P, Thomas D (2001) Optimal pump operation of water distribution systems using genetic algorithms. In: American water works association distribution system symposium
10. Barry JA. Watergy: energy and water efficiency in municipal water supply and wastewater treatment cost-effective savings of water and energy.
11. Anton A, Perju S (2004) Monitoring the main parameters of a water supply pumping station over ten years. In: 6th International conference on hydraulic machinery and hydrodynamics
12. Goldstein R, Smith W. Water and sustainability: U.S. electricity consumption for water supply & treatment—the next half century. http://www.epri.com/abstracts/Pages/Product-Abstract.aspx?ProductId=000000000001006787
13. House L. Water supply related electricity demand in California. http://drrc.lbl.gov/publications/water-supply-related-electricity

14. Kenway SJ, Priestley A, Cook S, Seo S, Inman M, Gregory A, Hall M. Energy use in the provision and consumption of urban water in Australia and New Zealand. https://publications.csiro.au/rpr/download?pid=csiro:EP116078&dsid=DS1
15. Pulido-Calvo I, Gutiérrez-Estrada JC (2011) Selection and operation of pumping stations of water distribution systems. Environ Res J 5(3):1–20
16. Cutore P, Campisano A, Kapelan Z, Modica C, Savic D (2008) Probabilistic prediction of urban water consumption using the SCEM-UA algorithm. Urban Water J 5(2):125–132
17. Feldman K (2009) Aspects of energy efficiency in water distribution systems. In: 5th IWA Water Loss Reduction Specialist Conference, pp 85–89
18. Ionescu GC, Ionescu DS (2005) The optimization of energy consumption in water supply systems. Acta Electrotehnica 46(4):191–194
19. Grigoras G, Istrate M (2011) An efficient clustering method in evaluation of the electric energy consumption from water hydrophore stations. Int Rev Model Simul 4(2):813–818
20. Sârbu I (1997) Energetic optimization of the water distribution systems. Publishing House of the Romanian Academy, Bucharest
21. Szychta L (2006) Energy consumption of water pumping for selected control systems. Electr Power Qual Util J 12(1):67–73
22. Wagner J, Shamir U, Marks D (1988) Water distribution reliability: analytical methods. J Water Resourc Plann Manage 114(3):253–275
23. Zessler U, Shamir U (1989) Optimal operation of water distribution systems. J Water Resourc Plann Manage 115(6):735–751
24. Magadan M, Rivas J (2011) The impact of energy consumption on the benefit functions of enterprises: proposal of an energy efficiency indicator. Environ Eng Manage J 10(12): 1831–1834
25. Cartina G, Grigoras G, Bobric EC (2005) Clustering techniques in fuzzy modelling, Application in power systems. Venus Publishing House, Iasi
26. Bobric EC, Cartina G, Grigoras G (2009) Fuzzy technique used for energy loss determination in medium and low voltage networks. Electron Electr Eng 90(2):95–98
27. Cartina G (2000) Artificial intelligence modelling techniques in power system control. Model Optim Mach Build Field 1(7):254–257
28. Steitz T, Haubrich HJ, Bovy A (1993) Reliability evaluation of power distribution systems with local generation using fuzzy sets. In: 11th power systems computation conference, Avignon, France
29. Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106
30. Yu Z, Haghighat F, Fung BCM, Hiroshi Y (2010) A decision tree method for building energy demand modelling. Energy Build 42(10):1637–1646
31. Olaru C, Wehenkel L (2003) A complete fuzzy decision tree technique. Fuzzy Sets Syst 138(2):221–254
32. Timofeev R (2004) Classification and regression trees (CART): theory and applications. Master's thesis, Humboldt University, Berlin
33. Loh WY (2008) Classification and regression tree methods. In: Encyclopaedia of statistics in quality and reliability. Wiley, pp 315–323
34. Lu X, Yang Z, Dong A (2005) Electricity market price spike forecast with data mining techniques. Electr Power Syst Res 73(1):19–29
35. Reston JC, Affonso CM, Oliveira RCL (2009) Pricing analysis in the Brazilian energy market: a decision tree approach. In: 2009 IEEE Power Tech, Bucharest
36. Razaee MR, Lelieveldt BPF, Reiber JHC (1998) A new cluster validity index for the fuzzy C-means. Pattern Recogn Lett 19(3/4):237–246
37. Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. J Intell Inf Syst 17(2/3):107–145
38. Gordon AD (1999) Classification, 2nd edn. Chapman & Hall, NewYork
39. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65

40. Istrate M, Grigoras G (2010) Energy consumption estimation in water distribution systems using fuzzy techniques. Environ Eng Manage J 9(2):249–256
41. Grigoras G, Istrate M, Scarlatache F (2013) Electrical energy consumption estimation in water distribution systems using a clustering based method. In: 6th international conference on electronics, computers and artificial intelligence
42. Grigoras G (2014) Assessment of electrical load in water distribution systems using representative load profiles-based method. Adv Elect Eng 2014:1–10
43. Jentgen L, Kidder H, Hill R, Conrad S. Water consumption forecasting to improve energy efficiency. http://www.waterrf.org/PublicReportLibrary/91189.pdf