Article

Not peer-reviewed version

Toward Smart SCADA Systems in the Hydropower Plants through Integrating Data Mining-based Knowledge Discovery Modules

Gheorghe Grigoras * , Razvan Garbea , Bogdan-Constantin Neagu

Posted Date: 3 September 2024

doi: 10.20944/preprints202409.0264.v1

Keywords: smart SCADA; Knowledge Discovery; Data Mining; Clustering; Hydropower plants

Article

Toward Smart SCADA Systems in the Hydropower Plants through Integrating Data Mining-Based Knowledge Discovery Modules

Gheorghe Grigoras [1,*], Răzvan Gârbea [2] and Bogdan-Constantin Neagu [3]

[1] Department of Power Engineering, "Gheorghe Asachi" Technical University of Iasi, 700050 Iasi, Romania
[2] Department of Power Engineering, "Gheorghe Asachi" Technical University of Iasi, 700050 Iasi, Romania; razvan-vasile.garbea@student.tuiasi.ro
[3] Department of Power Engineering, "Gheorghe Asachi" Technical University of Iasi, 700050 Iasi, Romania; bogdan.neagu@tuiasi.ro
* Correspondence: ggrigor@tuiasi.ro; Tel.: +40232278683/1162

Abstract: The increasing importance of hydropower generation has led to the development of new smart technologies and the need for reliable and efficient equipment in this field. As long as hydropower power plants are more complex to build up than other power plants, the operation regimes and maintenance activities become essential for the hydropower companies to optimize their performance, such that including the data-driven approaches in the decision-making process represents a challenge. In the paper, a comprehensive and multi-task framework integrated into a Knowledge Discovery module based on Data Mining to support the decisions of the operators from the control rooms and facilitate the transition from the classical to smart Supervisory Control and Data Acquisition (SCADA) system in hydropower plants has been designed, developed, and tested. It integrates tasks related to detecting the outliers through advanced statistical procedures, identifying the operating regimes through the patterns associated with typical operating profiles, and developing strategies for loading the generation units that consider the number of operating hours and minimize the water amount used to satisfy the power required by the system. The proposed framework has been tested using the SCADA system's database of a hydropower plant belonging to the Romanian HydroPower Company. The framework can offer the operators from the control room comparative information for a time horizon longer than one year. The tests demonstrated the utility of a Knowledge Discovery module to ensure the transition toward smart SCADA systems that will help the decision-makers improve the management of the hydropower plants.

Keywords: smart SCADA; Knowledge Discovery; Data Mining; Clustering; Hydropower plants

---

1. Introduction

Hydropower is one of the most important sources of electricity globally, along with wind, solar, and biomass. It helps cut down on greenhouse gas emissions, which are a major issue of global warming. The electrical power sector is playing its part in mitigating the effects of climate change by utilizing clean and renewable energy sources [1]. Hydropower plants currently account for over 75% of the world's renewable energy sources. Despite their widespread use, they have a huge potential to expand globally. [5]. A hydropower reservoir's power generation is affected by various factors, such as the water quantity yearly discharge, and load rate. Efficient planning and computational enhancements can increase the energy output by using the avail-able water [1].

The optimal operation of a hydroelectric power plant involves the gathering and processing of vast amounts of input data. Unfortunately, the techniques used in the design and implementation of the plant's operations are not always able to extract the most value from the data. This paper proposes a clustering-based method that can help the Decision-Makers (DEMA) identify the optimal hourly load patterns of the generators. The daily generation scheduling method is an integral part of the decision-making process in a power plant. It helps reduce the time it takes to make critical decisions by implementing effective maintenance plans and energy production techniques [2].

The increasing importance of hydropower generation has led to the development of new technologies and the need for reliable and efficient equipment. It has become a vital factor in the operations and maintenance of these machines. Hydropower plants are typically more expensive to set up than other energy sources. Nonetheless, they have a longer lifespan than other power plants, and their average maintenance and operations costs are typically around 2% of the investment. Although routine maintenance is carried out on the equipment, predictive maintenance can improve their efficiency [3], helping to maximize the life of a plant's resources and assets. It can also prevent costly repairs by identifying potential issues early. One of the most significant decisions to implement this approach is to monitor continuously the equipment's condition through a smart SCADA system where the Knowledge Discovery modules are to be integrated. The core step used in the Knowledge Discovery modules is Data Mining. It involves extracting information and transforming it into a more understandable format (containing the operation patterns) by the DEMA from the operators in the control room of the hydropower plant. The information provided through such Knowledge Discovery modules can help improve the lifetime of the power units by reducing their downtime and enhancing their production. It can also minimize the costs of operations and maintenance [4]. The challenge is associated with the following two questions:

1. How can a deeper analysis of the data from various processes and equipment within the hydropower plant be performed?
2. Which is the best approach to perform the analysis?

This process must be carried out efficiently to maximize the information from the SCADA database. The literature presents different Knowledge Discovery applications in the hydropower industry. Parvez et al. [1] proposed a linear regression procedure used to determine the energy production relationship between upstream and downstream hydro plants. A cluster analysis has been performed to find the typical generation curves. The goal of this project is to develop a class-based extreme learning machine that can determine the optimal operation rule for a hydropower reservoir. Through a k-means clustering algorithm, the cluster analysis is performed to split the influence factors into several sub-regions. The extreme learning machine is then optimized by particle swarm analysis to identify the complex relationship between the cluster's input and output. Feng et al. developed in [6] a class-based extreme learning machine that can determine the optimal operation rule for a hydropower reservoir. Through a k-means clustering algorithm, the cluster analysis is performed to split the influence factors into several sub-regions. The extreme learning machine is then optimized by particle swarm analysis to identify the complex relationship between the cluster's input and output. Zhang et al. propose in [7] an approach that can improve the quality of the monitoring data collected from hydropower units by implementing a clustering algorithm. This approach can be used to solve various problems related to the condition monitoring. A standard system to classify the huge amount of information that is collected and stored has been proposed by researchers in the study performed [8]. The system can meet the needs of the DEAM and provide them with the necessary services. Ahmed et al. [9] used three approaches, Local Outlier Factor as a density-based method, Feature Bagging for Outlier Detection as an ensemble method, and Subspace Outlier Degree, to analyze the anomalous data collected from a hydropower plant and compare their performance. The outliers were then verified by the expert utilizing a feature selection process and a decision tree to identify the critical variables that could be associated with the anomalies. Valencia et al. presented in [10] a procedure that uses Knowledge Discovery to analyze a data set and extract structured information related to a hydroelectric power plant. This method can be utilized to train systems focused on identifying faults. Zhang et al. proposed in [11] a decision tree-based clustering scheme that can be used to determine the various operating regimes in hydropower plants. The method uses k-means++ clustering to classify the data. The decision tree is then constructed using the group labels and other features. The decision tree is then analyzed and pruned according to the classification accuracy and complexity requirements. The reference [12] introduced the data mining concept integrated into a SCADA system to help the hydropower plant's operators make informed decisions. The data collected by the data mining process can determine the typical loading profiles of each generation unit. Sahin and Karakus presented in [13] a study on the energy generation forecasting of a hydroelectric plant based on Machine Learning and a hybrid Genetic Grey Wolf Optimizer-based Convolutional Neural Network/Recurrent Neural Network-Long Short-Term Memory regression approach. The findings can help improve the efficiency of resource management and energy generation.

Two aspects draw attention concerning the current applications of the KD in hydropower plants:

1. The vast majority of applications aim for a single task implemented at the level of data analysis regarding the energy production relationship between upstream and downstream hydro plants, energy production forecasting, identifying the operating regimes, improving the data quality, outliers' detection, identifying faults, or determining the typical operating profiles.
2. The analysis time horizon corresponds with a day, season, or year.

Using as a starting point these two remarks, the main contribution of the paper is associated with designing, developing, and testing an original clustering-based data mining framework integrated into a Knowledge Discovery module from the SCADA software of a hydropower plant which fulfil more tasks regarding:

- Performing an advanced statistical analysis and outliers' detection,
- Identifying the operating regimes and hourly typical operating profiles,
- Developing the strategies for loading the generation units that consider the number of hours of operation and the minimization of the amount of water used to satisfy the power required by the system.

The framework can offer comparative information for a time horizon longer than one year, allowing the quick identification of key performance indicators that characterize the operation of the hydropower plant.

The remainder of the paper includes four sections. Section 2 presents the theoretical aspects regarding the integration of the Knowledge Discovery and Data Mining in the smart SCADA, Section 3 integrates the details on the multi-task framework integrated into the Data Mining-based Knowledge Discovery, Section 4 covers the case study where the proposed framework has been tested using the SCADA system's database of a hydropower plant belonging to the Romanian HydroPower Company, and Section 5 highlights the conclusions and future work.

2. Knowledge Discovery and Data Mining in Smart SCADA

2.1. Knowledge Discovery vs. Data Mining

Knowledge discovery (KD) and Data Mining (DM) have transformed the power engineering research. To carry out effective and meaningful research, a deep understanding of various aspects of data mining and knowledge discovery is necessary [14]. It includes the expertise of specialists who work in all components associated with the energy generation, transmission, and distribution of the chain representing the power system.

There are some misunderstandings about the terms data mining and knowledge discovery defined in databases. Although many specialists and researchers use DM as a synonym for knowledge discovery, DM is not the entire knowledge discovery process. In addition to being defined as data mining, it comes with other names, such as information discovery or knowledge extraction. The KD is a process that aims to identify relationships and patterns in large datasets. It is defined typically as a non-trivial process that involves identifying novel, useful, and understandable patterns. In a narrow sense, KD refers to extracting information from a data source. While it can be done through various methods, the term refers to obtaining knowledge from textual or database data. The combined process is referred to as the KD process.

Figure 1 shows the steps of a KD process, integrating the Data Mining and the details are introduced in the following [12,14,15].

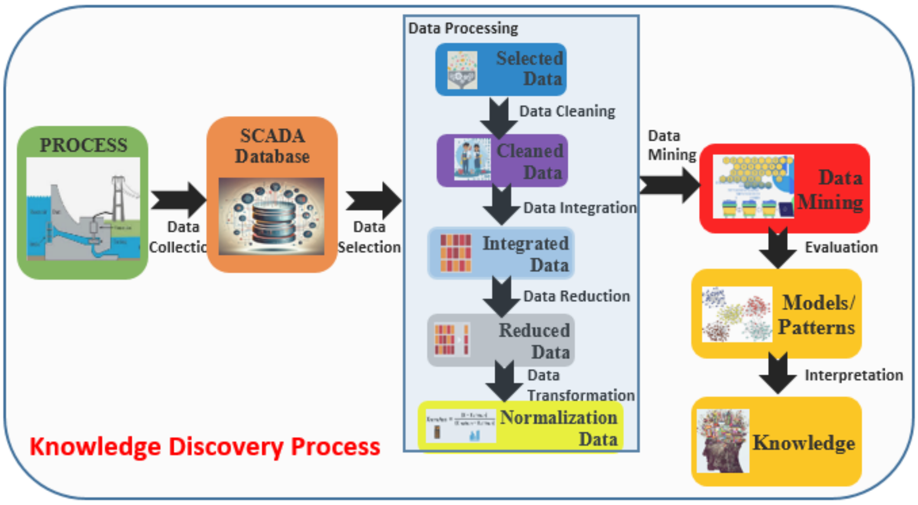


Figure 1. The steps of a KD process.

Data Selection. KD's initial step is data selection, which involves gathering information from the SCADA database. This process is carried out to create a raw dataset.

Data Cleaning (Preprocessing). Ensuring that the data collected is of good quality is done through preprocessing. This process involves handling noise, missing values, and inconsistencies.

Data Transformation. After cleaning the data, it's usually necessary to transform it into something suitable for mining. This can be done through various methods such as feature engineering and scaling. To make it easier for the machine learning tools, the employed label encoding to convert categorical data into a more readable format can be done.

Data Mining. The DM step involves uncovering patterns, anomalies, or relationships between the data. The DM process is composed of numerous steps, each of which is related to a specific discovery task. The extraction of knowledge involves the process of gathering and storing information. It also involves analyzing and visualizing the data, designing models for machine and human interaction, and learning how to use efficient methods. One of the most frequently used techniques is clustering, which enables us to group the data into distinct groups based on similarity.

Interpretation. After Data Mining, the next step is to interpret the results. This involves understanding the clusters (patterns) and their features.

2.2. Data Mining Techniques

The Data Mining techniques are divide in two main categories: descriptive and predictive, see Figure 2 [16–18].

doi:10.20944/preprints202409.0264.v1

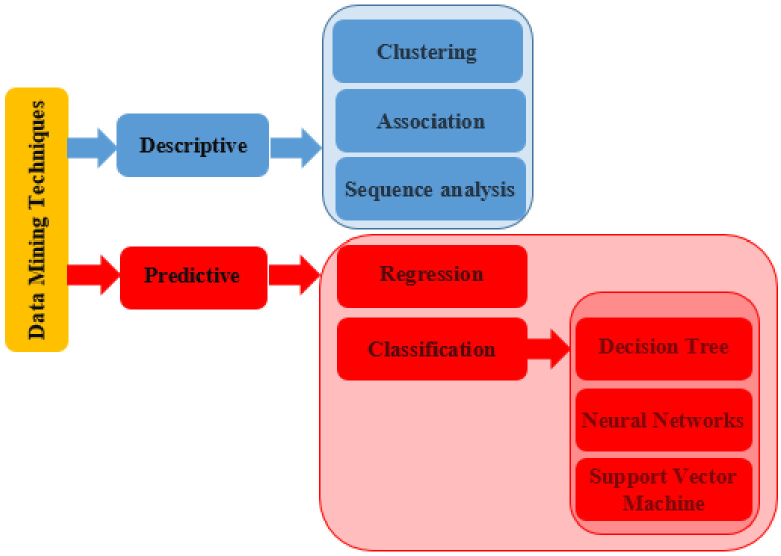


Figure 2. Data Mining techniques.

Cross-tabulation, correlation, and frequency are some of the characteristics used in the production of descriptive data mining. This process is utilized to identify the similarities between the data and the existing patterns. Another characteristic used in this type of analysis is associated with developing captivating subgroups. This is done by analyzing the data and transforming it into meaningful information. Descriptive data mining involves techniques, such as clustering, association rule mining, and sequence discovery analysis.

Clustering is commonly used in data mining to organize information by grouping related data points. It helps the DEMA identify patterns and similarities between different sets of information. It can be used to classify and extract patterns in the data, identify anomalies, or analyze spatial data.

Association Rules. The goal is to find the correlations in the data sets from the database. Even if the data sets come from different sources, these correlations can identify patterns that can help reveal process trends or explain the operation characteristics of the equipment/installations.

Sequence discovery analysis. The goal of sequence discovery analysis is to find interesting data sets that contain sequential patterns. This process usually involves identifying frequent patterns about a certain frequency support measure.

The second category aims to predict the future results of a given variable with a high degree of accuracy. This data mining is carried out using supervised learning techniques. There are three categories of methods that are used in this type of mining: regression, classification, and time-series analysis. The latter two are utilized in predictive analysis to model the data.

Regression It is similar to the clustering technique in that it focuses on the relationship between a target and an independent variable. The target variables can be influenced by different predictors or independent factors, which is why a regression analysis is utilized. It predicts the outcomes based on the input fields relevant to the target.

Classification. It is carried out by identifying the various features of the data. This process helps to identify patterns and extract meaningful insights. It also helps in improving the quality of the data. The appropriate features are selected to classify it after the data has been collected and analyzed. A suitable algorithm is then chosen to implement.

- *Support Vector Machine*. This algorithm creates a boundary between the different classes/patterns. It identifies the features that are most important to the classification process.
- *Decision Tree*. he classification process is carried out using a tree-based structure. This algorithm uses a set of conditions to categorize the data. The root nodes of the structure are set for the test conditions, while the leaf nodes are for the outcome.
- *Neural Network*. A neural network model is a computational resource that can recognize the relationships between various data sets. These units, which act like neurons, are formed by

connecting the inputs and outputs. The model considers the connection strength and outputs the information in a hidden layer. The neural network is similar to the human brain in that it requires training to be effective. Although it can be hard to interpret, the models are reliable and can even classify past training procedures.

The Smart SCADA can improve the efficiency of the hydropower plant operation by combining Knowledge Discovery and statistical analysis. Figure 3 displays the interdependencies between the Knowledge Discovery and Smart SCADA (adapted after [19]).

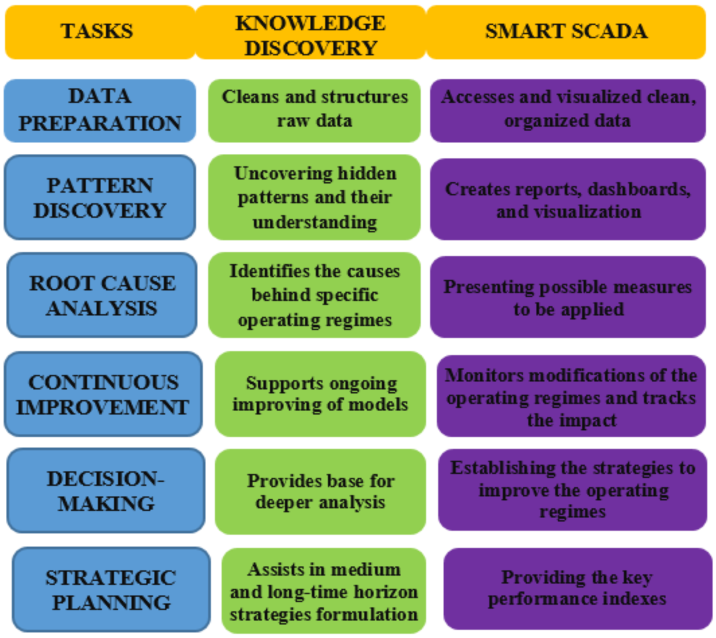


Figure 3. Interdependencies between the Knowledge Discovery and Smart SCADA.

3. Multi-task Framework Integrated into the Knowledge Discovery Module

The main steps in the multi-task framework related to developing the Knowledge Discovery module based on Data Mining to facilitate the transition from the classical to smart SCADA system in the hydropower plants are discussed in this section.

Figure 4 presents the basic structure of an automation architecture, including the SCADA system implemented in a power plant.

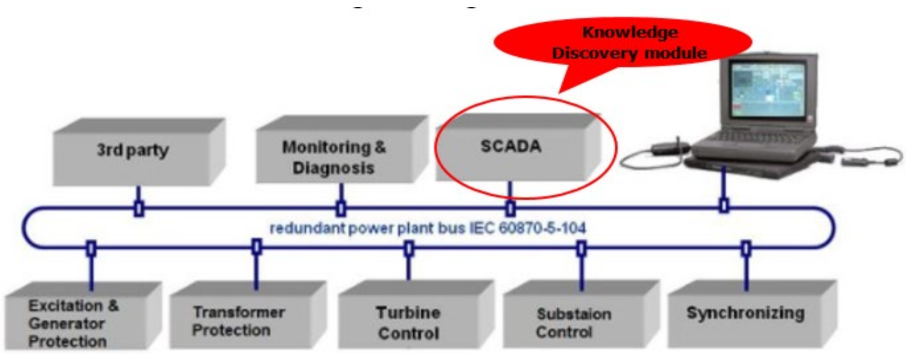


Figure 4. The basic structure of an automation architecture including the SCADA system.

The data acquisition is performed on main hydro components (turbine, generator, power transformer, substation level), recorded in the SCADA system and made available to the operator from the control room through the Human-Machine Interface. The Data Mining-based Knowledge Discovery module will be implemented at the SCADA level to ensure data-driven decision-making, ensuring the transition toward smart SCADA.

Figure 5 shows the flow chart of the multi-task framework integrated into the Knowledge Discovery module. Details regarding each task are provided above.

The SCADA system collects the water flows in the pipes (WF_{pipe}), in [m³], that sup-ply the turbines and the requested power of the system to the hydropower plant (P^{sys}_{req}), in [MW]. It also contains information about the various technical parameters of the generation units, such as the active and reactive powers (P_{HPP}^{GU} and Q_{HPP}^{GU}), in [MW] and [MVAr], the stator voltage and current ($V_{HPP}^{s,GU}$ and $I_{HPP}^{s,GU}$), in [kV] and [kA], the excitation voltage or current of each generation unit ($V_{HPP}^{ex,GU}$ and $I_{HPP}^{ex,GU}$), in [kV] and [kA], frequency (f_{HPP}), in [Hz], water levels of the reservoir upstream ($WL_{HPP}^{rd,u}$ and downstream and $WL_{HPP}^{rd,d}$). The information is recorded in the database with the time details in the format *year-month-day-hour*, as seen in Figure 6.

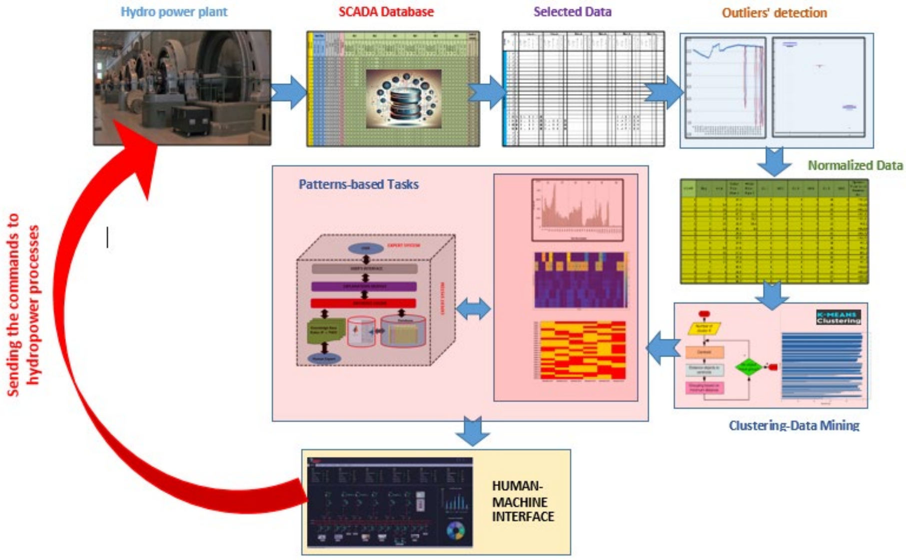
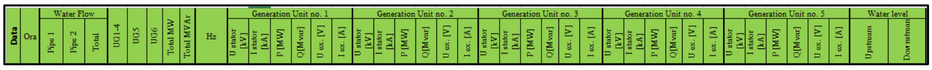


Figure 5. The multi-task framework integrated in the Knowledge Discovery module.



Data	Ora	Water Flow			UG1-4	UG5	UG6	Total MW	Total MVAr	Hz	Generation Unit no. 1						Generation Unit no. 2						Generation Unit no. 3						Generation Unit no. 4						Generation Unit no. 5						Water level	
		Pipe 1	Pipe 2	Total							U stator [kV]	I stator [kA]	P [MW]	Q[Mvar]	U ex. [V]	I ex. [A]	U stator [kV]	I stator [kA]	P [MW]	Q[Mvar]	U ex. [V]	I ex. [A]	U stator [kV]	I stator [kA]	P [MW]	Q[Mvar]	U ex. [V]	I ex. [A]	U stator [kV]	I stator [kA]	P [MW]	Q[Mvar]	U ex. [V]	I ex. [A]	U stator [kV]	I stator [kA]	P [MW]	Q[Mvar]	U ex. [V]	I ex. [A]	Upstream	Downstream

Figure 6. The fields of the SCADA database.

Task 1 – Statistical analysis and outliers' detection.

The statistical analysis is associated with exploring and presenting large amounts of data on the technical parameters based on the parameters such as mean, standard deviation, confidence degree, quintiles - Q0 (minimum value), Q1 (25th), Q2 (50th), Q3 (75th), and Q4 (maximum value). Also, the boxplot is used to show the spread and skewness of the variables from the database through their quintiles. It can also include lines known as whiskers, which extend from the box to indicate variability outside the lower and upper limits of the dataset. Outliers with significant differences from the rest of the data can be plotted on the box-and-whisker diagram [20]. Thus, the DEMA (represented by the operator from the control room) can select the fields containing the values of the monitoring parameters. Regarding the outliers' detection, a rules-based algorithm based on the values of the quintiles has been integrated [21,22]. These two rules refer to Q0 and Q4:

(1) If < X^h < $Q0^x$ > then < X^h is outlier >

(2) If < X^h > Q4x > then < X^h is outlier>

where: X^h is hourly value of the analysed technical parameter and x refers to the name of the analysed technical parameter (identically with the name of the field).

However, there are cases when the outliers identified for certain technical parameters can be associated with the different operating regimes compared to most regimes but which do not lead to the violation of the allowable limits. In these cases, an attention message will alert the DEMA, who will verify only those regimes. For all other cases, when the outliers are identified, the least squares

 doi:10.20944/preprints202409.0264.v1

method is applied to estimate the "true" value. The approach is based on a regression model determined for that parameter.

A comparison between more years can be done by choosing the analysis period.

Task 2 – Determining the operating regimes.

This task is based on the clustering-based data mining process. The input data are associated with a matrix structure built with the values of technical parameters selected by the DEMA. The regimes are identified through the typical operating profiles divided into the intervals associated with the generating units (GU1, ..., GUn) or hourly loading patterns, including the hourly average power of each generation unit. The DEMA can choose any algorithm from the hierarchical clustering category (complete-linkage clustering, single-linkage clustering, average linkage clustering, centroid linkage clustering, median linkage clustering, or Ward linkage clustering) or K-means clustering algorithm. In hierarchical clustering, a similarity measure between the sets of observations is necessary to determine which patterns should be grouped or separated. Usually, in hierarchical clustering, the set's similarity is determined by using a distance between the observations. It is done through a linkage criterion that specifies the set's similarity.

When the DEMA requests characterization of the operating regimes through the typical operating profiles, then the matrix structure of the input data corresponding to the clustering process is identified with the fields of the active power produced by each generation unit in a certain time horizon (usually a year to cover all operating regimes).

The matrix size is H_{year} x (N_{HPP}^{GU}x24 + 2), where H_{year} represents the number of hours when at least one generation unit worked and N_{HPP}^{GU} – the number of the generation units from the hydropower plant. The additional columns correspond to the water levels of the reservoir - upstream and downstream.

Each obtained pattern is associated with a typical operating profile which characterizes the operating regime of the hydropower plant in certain periods (days inside a year). The DEMA can identify the operating regimes of the hydropower plant through an evolution along the time axis of the degree of hourly loading of all generating units and the water volume used in each operating regime. He can establish an operating strategy of the plant for the next day depending on the forecasting of the requested powers by the system.

The input data matrix is different when the DEMA's requests are as the operating regimes to be characterized through hourly loading patterns. The structure is identified with the fields of the active power produced by each generation unit in a certain time horizon (usually a year to cover all operating regimes). The size of the matrix is H_{year} x (N_{HPP}^{GU}+ 2), where H_{year} and N_{HPP}^{GU} have the same signification as above. The last two columns correspond to the water levels of the reservoir.

Task 3 – Developing the strategies to load the generation units.

The task is associated with an expert system which uses the operating regimes to be characterized by hourly loading patterns determined above and the number of hours of operation. This last information is recorded in a database and considered in the decision-making process to avoid overloading the generation units over a long period, which leads to minimizing the number of maintenance operations. Using a water amount to satisfy the power required by the system represents the main objective. The main components of the expert system are presented synthetically in the following [23,24].

- The knowledge base is composed of two main elements: the rules base (which contains the knowledge required to solve problems) and the facts base (the patterns obtained in the clustering-based data mining are recorded in this base).
- The inference engine can determine the mode in which knowledge derived from the rules base is utilized to interpret the data from the information base. It can perform various tasks, such as confirming or rejecting a hypothesis or the solution of a problem.
- The editor of knowledge base provides the DEMA with the ability to update and inspect the information base's content, particularly its rules base's content.
- The explanation system can provide explanations for the stages in the Expert System's reasoning.

4. Case Study

The proposed framework has been tested using the SCADA system's database of a hydropower plant belonging to the Romanian HydroPower Company.

The plant, identified through the red circle in Figure 7, is the first from a hydro arrangement located on an important river in eastern Romania. The plant has six Francis-type units. The first water pipe supplies the last two units, and the second piper the first four units. The first four units have a total installed power of 27.5 MW, while units five and six have 50 MW.

The SCADA database includes fields completed at every hour with the following technical parameters: the individual water flows of each pipe, the total water flow of the two pipes, the total produced power of the first four-generation units (GU1 – GU4), the produced power by the GU5, the produced power by the GU6, the total produced active and reactive power by the plant, the frequency, the stator voltage, the stator current, the produced active and reactive powers, the excitation voltage, the excitation current of each generation unit and water levels of the reservoir (upstream and downstream).

The SCADA file associated with a day from the database containing the fields highlighted above is shown in Figure 8. The signification of the blank cells corresponds with the case of the non-loading of the generation unit. The obtained results at the year level for each task integrated into the proposed framework are presented in the following.

The DEMA can see after data processing the summary information on the operation of the plant regarding the number of hours and the total energy produced by each generation unit (GU1 – GU2), see Figure 9.

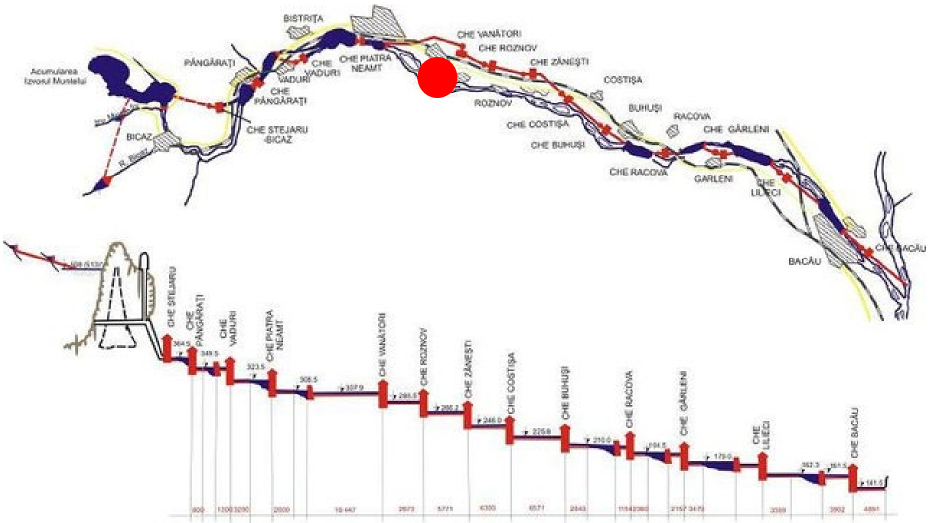
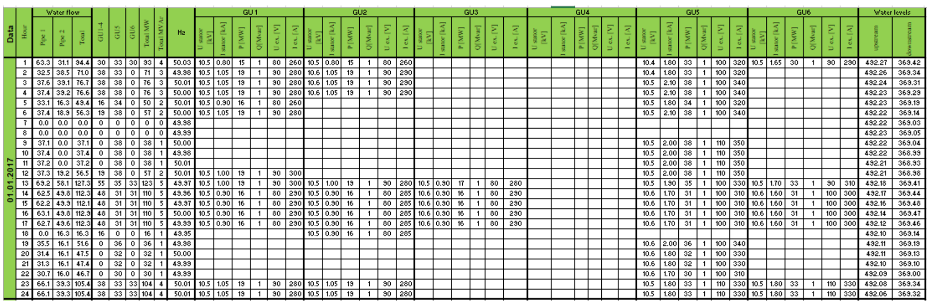


Figure 7. The hydro arrangement of which the analyzed plant is a part.



Data	Hour	Water flow Pipe 1	Water flow Pipe 2	Water flow Total	GU1-4	GU5	GU6	Total MW	Total MVAr	Hz
01.01.2017	1	63.3	31.1	94.4	30	33	30	93	4	50.03
	2	32.5	38.5	71.0	38	33	0	71	3	49.98
	3	37.6	39.1	76.7	38	38	0	76	3	50.01
	4	37.4	39.2	76.6	38	38	0	76	3	50.00
	5	33.1	16.3	49.4	16	34	0	50	2	50.01
	6	37.4	18.9	56.3	19	38	0	57	2	50.00
	7	0.0	0.0	0.0	0	0	0	0	0	49.98
	8	0.0	0.0	0.0	0	0	0	0	0	49.99
	9	37.1	0.0	37.1	0	38	0	38	1	50.00
	10	37.4	0.0	37.4	0	38	0	38	1	49.98
	11	37.2	0.0	37.2	0	38	0	38	1	50.01
	12	37.3	19.2	56.5	19	38	0	57	2	50.01
	13	69.2	58.1	127.3	55	35	33	123	5	49.97
	14	62.5	49.8	112.3	48	31	31	110	5	49.96
	15	62.2	49.9	112.1	48	31	31	110	5	49.97
	16	63.1	49.8	112.9	48	31	31	110	5	50.00
	17	62.7	49.6	112.3	48	31	31	110	5	49.99
	18	0.0	16.3	16.3	16	0	0	16	1	49.95
	19	35.5	16.1	51.6	0	36	0	36	1	49.98
	20	31.4	16.1	47.5	0	32	0	32	1	50.00
	21	31.3	16.1	47.4	0	32	0	32	1	49.99
	22	30.7	16.0	46.7	0	30	0	30	1	49.99
	23	66.1	39.3	105.4	38	33	33	104	4	50.01
	24	66.1	39.3	105.4	38	33	33	104	4	50.01

Hour	GU1 U stator [kV]	GU1 I stator [kA]	GU1 P [MW]	GU1 Q[Mvar]	GU1 U ex. [V]	GU1 I ex. [A]	GU2 U stator [kV]	GU2 I stator [kA]	GU2 P [MW]	GU2 Q[Mvar]	GU2 U ex. [V]	GU2 I ex. [A]
1	10.5	0.80	15	1	80	260	10.5	0.80	15	1	80	260
2	10.5	1.05	19	1	90	280	10.6	1.05	19	1	90	290
3	10.5	1.05	19	1	90	280	10.6	1.05	19	1	90	290
4	10.5	1.05	19	1	90	280	10.6	1.05	19	1	90	290
5	10.5	0.90	16	1	80	260						
6	10.5	1.05	19	1	90	280						
7												
8												
9												
10												
11												
12	10.5	1.00	19	1	90	300						
13	10.5	1.00	19	1	90	300	10.5	1.00	19	1	90	280
14	10.5	0.90	16	1	80	290	10.5	0.90	16	1	80	285
15	10.5	0.90	16	1	80	290	10.5	0.90	16	1	80	285
16	10.5	0.90	16	1	80	290	10.5	0.90	16	1	80	285
17	10.5	0.90	16	1	80	290	10.5	0.90	16	1	80	285
18							10.5	0.90	16	1	80	285
19												
20												
21												
22												
23	10.5	1.05	19	1	90	280	10.5	1.05	19	1	90	280
24	10.5	1.05	19	1	90	280	10.5	1.05	19	1	90	280

Hour	GU3 U stator [kV]	GU3 I stator [kA]	GU3 P [MW]	GU3 Q[Mvar]	GU3 U ex. [V]	GU3 I ex. [A]	GU4 U stator [kV]	GU4 I stator [kA]	GU4 P [MW]	GU4 Q[Mvar]	GU4 U ex. [V]	GU4 I ex. [A]
1												
2												
3												
4												
5												
6												
7												
8												
9												
10												
11												
12												
13	10.5	0.90	17	1	80	280						
14	10.6	0.90	16	1	80	290						
15	10.6	0.90	16	1	80	290						
16	10.6	0.90	16	1	80	290						
17	10.6	0.90	16	1	80	290						
18												
19												
20												
21												
22												
23												
24												

Hour	GU5 U stator [kV]	GU5 I stator [kA]	GU5 P [MW]	GU5 Q[Mvar]	GU5 U ex. [V]	GU5 I ex. [A]	GU6 U stator [kV]	GU6 I stator [kA]	GU6 P [MW]	GU6 Q[Mvar]	GU6 U ex. [V]	GU6 I ex. [A]	Water levels upstream	Water levels downstream
1	10.4	1.80	33	1	100	320	10.5	1.65	30	1	90	290	492.27	369.42
2	10.4	1.80	33	1	100	320							492.26	369.34
3	10.5	2.10	38	1	100	340							492.24	369.31
4	10.5	2.10	38	1	100	340							492.23	369.29
5	10.5	1.80	34	1	100	320							492.23	369.19
6	10.5	2.10	38	1	100	340							492.22	369.14
7													492.22	369.03
8													492.23	369.05
9	10.5	2.00	38	1	110	350							492.22	369.04
10	10.5	2.00	38	1	110	350							492.22	368.99
11	10.5	2.00	38	1	110	350							492.21	368.93
12	10.5	2.00	38	1	110	350							492.21	368.98
13	10.5	1.90	35	1	100	330	10.5	1.70	33	1	90	310	492.18	369.41
14	10.6	1.70	31	1	100	310	10.6	1.60	31	1	100	300	492.17	369.44
15	10.6	1.70	31	1	100	310	10.6	1.60	31	1	100	300	492.16	369.48
16	10.6	1.70	31	1	100	310	10.6	1.60	31	1	100	300	492.14	369.47
17	10.6	1.70	31	1	100	310	10.6	1.60	31	1	100	300	492.12	369.46
18													492.10	369.14
19	10.6	2.00	36	1	100	340							492.11	369.19
20	10.6	1.80	32	1	100	330							492.11	369.13
21	10.6	1.80	32	1	100	330							492.10	369.10
22	10.6	1.70	30	1	100	310							492.09	369.00
23	10.5	1.80	33	1	110	330	10.5	1.80	33	1	110	330	492.08	369.34
24	10.5	1.80	33	1	110	330	10.5	1.80	33	1	110	330	492.06	369.32

Figure 8. SCADA file associated with a day from the database.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 September 2024 doi:10.20944/preprints202409.0264.v1

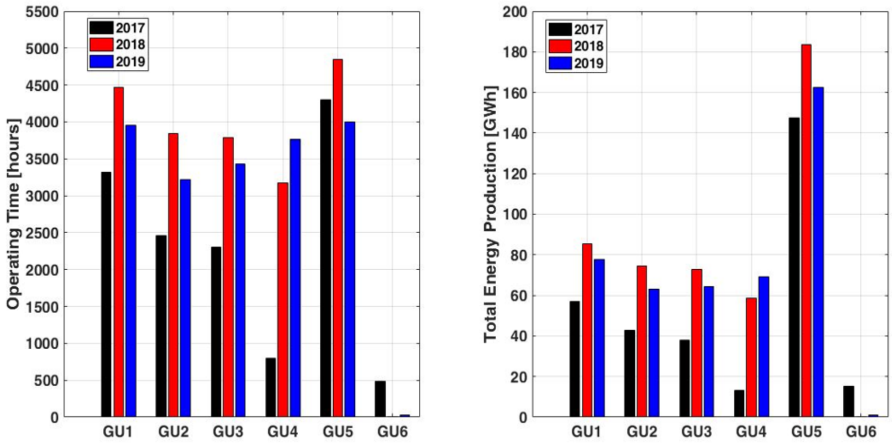


Figure 9. The summary information regarding the operating of the plant regarding the number of hours and the total energy produced by each generation unit.

Table 1 presents the extracted results from the advanced statistical analysis containing the mean (m), standard deviation (σ), and quintiles (Q0, Q1, Q2, Q3, and Q4) for each technical parameter from the database. The values have been calculated only for the hours when at least one unit was in operation. Also, the DEMA has available boxplots from which outliers are identified quickly.

Table 1. The extracted results from the statistical analysis.

Statistical parameters		m	σ	Q0	Q1	Q2	Q3	Q4
Water flow	First Pipe [m3/s]	36.42	9.75	18.70	31.50	34.50	36.30	74.40
	Second Pipe [m3/s]	29.92	14.02	13.40	17.30	30.80	36.60	78.20
	Total [m3/s]	56.14	21.35	10.50	44.50	54.10	70.80	133.20
Active and Reactive Powers	GU1-GU4 [MW]	29.60	13.54	0.90	18.00	30.00	37.00	78.00
	GU5 [MW]	34.33	3.09	29.00	31.00	35.00	37.00	40.00
	GU6 [MW]	31.51	2.01	1.00	30.00	31.00	33.00	40.00
	Total [MW]	56.34	20.39	13.00	45.00	56.00	72.00	126.00
	Total [MVAr]	6.39	5.90	1.00	2.00	3.00	10.00	30.00
Frequency	[Hz]	49.99	0.02	49.10	49.98	50.00	50.01	50.40
GU 1	U stator [kV]	10.43	0.71	1.40	10.40	10.40	10.50	50.00
	I stator [kA]	0.96	0.11	0.08	0.90	0.95	1.00	1.90
	P [MW]	17.20	1.89	0.90	15.00	17.00	18.00	22.00
	Q[Mvar]	2.60	1.99	1.00	1.00	1.00	5.00	11.00
	U ex. [V]	88.28	8.80	8.00	80.00	90.00	95.00	110.00
	I ex. [A]	287.91	17.29	100.00	280.00	290.00	300.00	360.00
GU 2	U stator [kV]	10.48	0.38	1.10	10.50	10.50	10.50	10.70
	I stator [kA]	0.96	0.10	0.70	0.90	0.95	1.05	1.20
	P [MW]	17.35	1.86	14.00	16.00	17.00	19.00	21.00
	Q[Mvar]	2.58	1.99	1.00	1.00	1.00	5.00	21.00
	U ex. [V]	89.84	7.80	70.00	85.00	90.00	95.00	110.00
	I ex. [A]	289.65	15.01	245.00	280.00	290.00	300.00	320.00
GU 3	U stator [kV]	10.41	0.42	1.60	10.40	10.50	10.50	10.70
	I stator [kA]	0.91	0.08	0.70	0.85	0.90	0.95	1.15
	P [MW]	16.48	1.41	13.00	15.00	16.00	17.00	20.00
	Q[Mvar]	2.76	2.00	0.50	1.00	1.00	5.00	10.00
	U ex. [V]	89.60	40.80	65.00	80.00	90.00	95.00	870.00
	I ex. [A]	289.21	58.37	115.00	280.00	290.00	300.00	2980.00
GU 4	U stator [kV]	10.45	0.09	10.10	10.40	10.50	10.50	10.60
	I stator [kA]	0.92	0.08	0.75	0.85	0.90	1.00	1.10

	P [MW]	16.59	1.36	14.00	15.00	17.00	18.00	20.00
	Q[Mvar]	2.78	1.99	1.00	1.00	1.00	5.00	5.00
	U ex. [V]	87.38	8.87	60.00	80.00	90.00	95.00	105.00
	I ex. [A]	290.13	15.75	260.00	280.00	290.00	300.00	330.00
GU 5	**U stator [kV]**	10.41	0.52	1.40	10.40	10.50	10.50	10.70
	I stator [kA]	1.89	0.17	0.85	1.70	1.90	2.05	2.70
	P [MW]	34.33	3.09	29.00	31.00	35.00	37.00	40.00
	Q[Mvar]	2.56	1.96	1.00	1.00	1.00	5.00	10.00
	U ex. [V]	104.78	17.89	10.00	100.00	100.00	110.00	1110.00
	I ex. [A]	332.53	108.88	235.00	315.00	330.00	340.00	3210.00
GU 6	**U stator [kV]**	10.49	0.07	10.30	10.50	10.50	10.50	10.60
	I stator [kA]	1.70	0.07	1.50	1.65	1.70	1.75	1.85
	P [MW]	31.52	1.33	29.00	30.00	31.00	33.00	33.00
	Q[Mvar]	2.07	1.77	1.00	1.00	1.00	5.00	5.00
	U ex. [V]	94.47	6.15	80.00	90.00	95.00	100.00	110.00
	I ex. [A]	303.34	10.77	280.00	300.00	300.00	310.00	340.00
Water Level	**Upstream [mdMB]**	492.66	5.86	479.28	489.80	493.40	497.99	500.80
	Downstream [mdMB]	368.21	5.90	356.43	368.90	369.06	369.20	497.48

Two situations are presented in the following. The first refers to the loading of the generation units when the outliers identified associated with the different operating regimes compared to most regimes did not lead to the violation of the allowable limits, see Figure 10. In these cases, an attention message is launched to the DEMA, who must verify the regimes and establish if any disturbance appeared.

The second situation belongs to the downstream level of the water reservoir recorded, where more outliers have been identified exceeding the upper limit. For these values, the least squares method has been applied to estimate the “true” value.

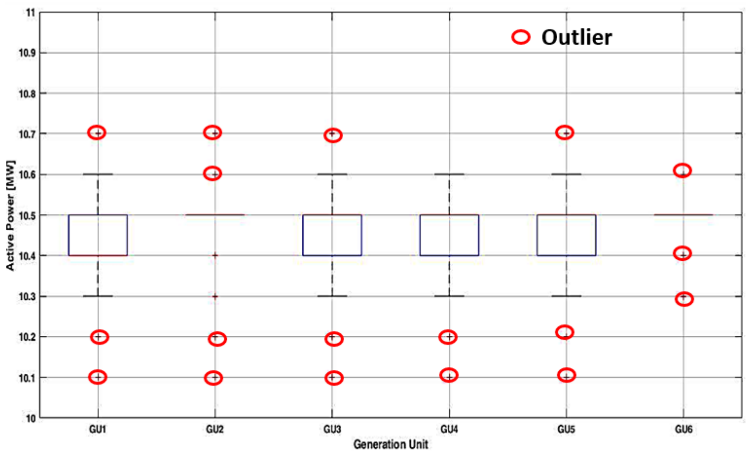
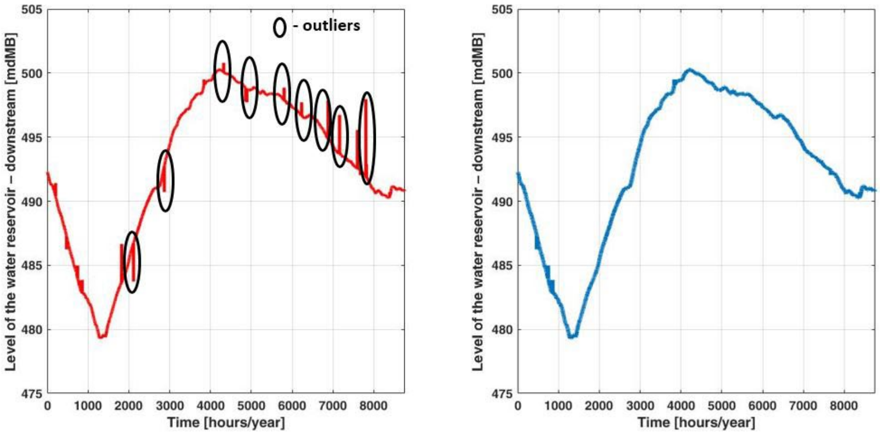


Figure 10. The boxplots corresponding to the loading of the generation units over a period of one year.

In the case of the outliers below Q0, these depend on the upstream level of the water reservoir, which did not lead to the violation of the allowable limits. Figures 11 and 12 present the results obtained in these cases (with and without outliers over the maximum limit).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 September 2024 doi:10.20944/preprints202409.0264.v1



(a) (b)

Figure 11. The values corresponding to the level of the water reservoir – downstream (a - taken from the database containing outliers; b – after data processing, without outliers).

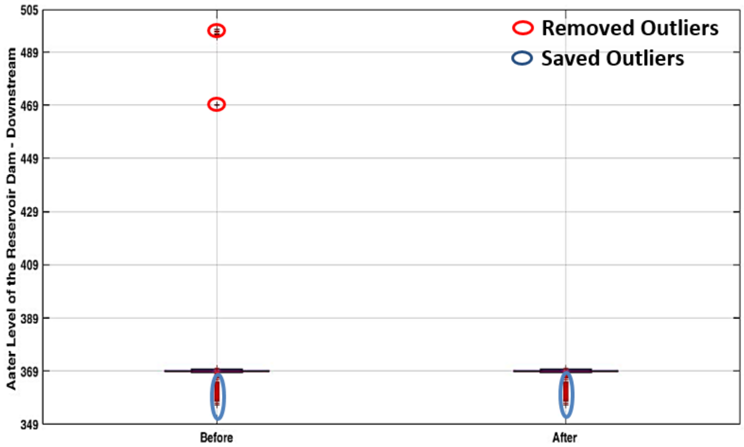


Figure 12. The boxplots corresponding to the level of the water reservoir – downstream.

The second task refers to determining the operating regimes identified through the typical operating profiles divided into the intervals associated with the generating units (GU1, ..., GU6) and hourly loading patterns, including the hourly average power of each generation unit. This task is based on the clustering-based data mining process.

The input data have been associated with a matrix structure built with the values of hourly active powers of all six generation units selected from the database from three successive years (2017 – 2019). The K-means clustering algorithm has been used to obtain the typical operating profiles presented in Figures 13 - 16.

Annex A includes in Tables A1 – A4 the hourly values of each operating regime (identified through the Patterns P1 – P4) containing the typical operating profiles divided into the intervals associated with the generating units (GU1, ..., GU6).

The input data used to obtain hourly loading patterns contains the following fields: water flows on the two pipes, the loading of each generator unit, and the upstream and downstream water level of the reservoir.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 September 2024 doi:10.20944/preprints202409.0264.v1

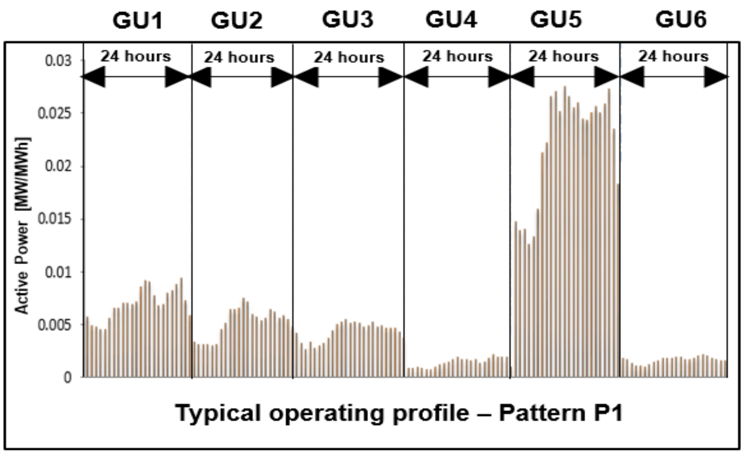


Figure 13. The typical operating profile of the hydropower plant assigned to pattern P1.

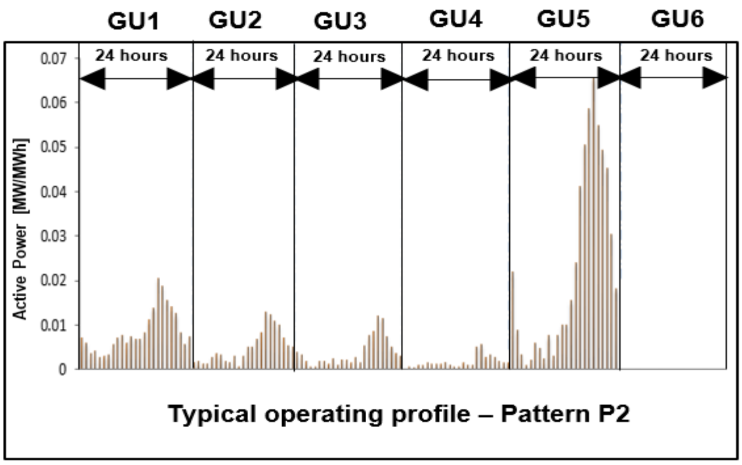


Figure 14. The typical operating profile of the hydropower plant assigned to pattern P2.

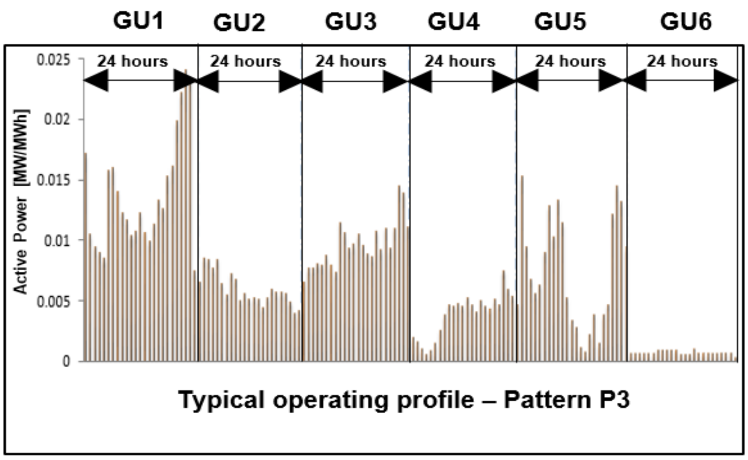


Figure 15. The typical operating profile of the hydropower plant assigned to pattern P3.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 September 2024 doi:10.20944/preprints202409.0264.v1

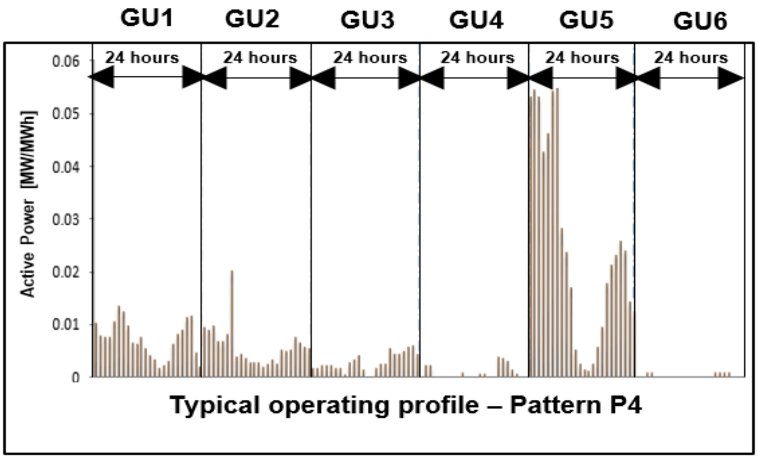


Figure 16. The typical operating profile of the hydropower plant assigned to pattern P4.

The K-means clustering algorithm has been used to extract the patterns, and the optimal number has been 25. Figures 17 - 19 show how the hydropower plant has been operated based on the obtained patterns of hourly loading of the generation units in three consecutive years, 2017 – 2019. Annex A includes in Tables A5 – A6 details of the patterns regarding the hourly loading of the generation units GU1 – GU6 and their features regarding the operating conditions in an analysed three-year period.

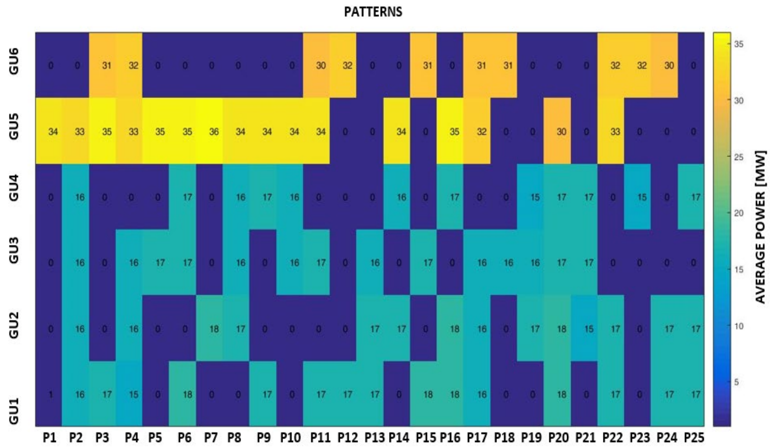


Figure 17. The patterns obtained for the hourly loading of the generation units in 2017.

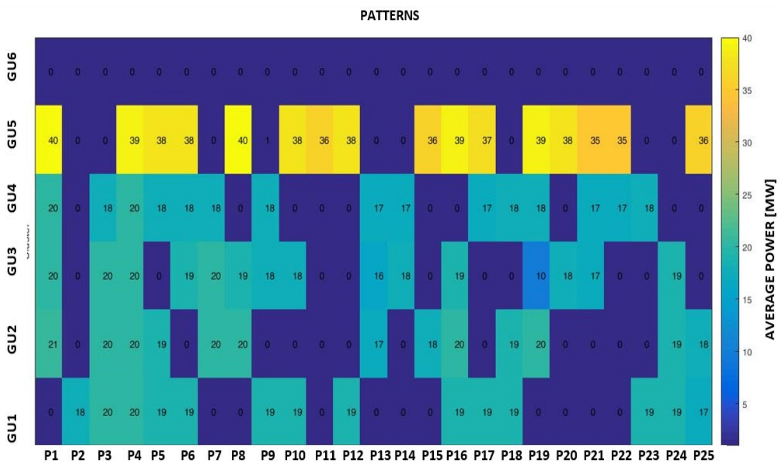


Figure 18. The patterns obtained for the hourly loading of the generation units in 2018.

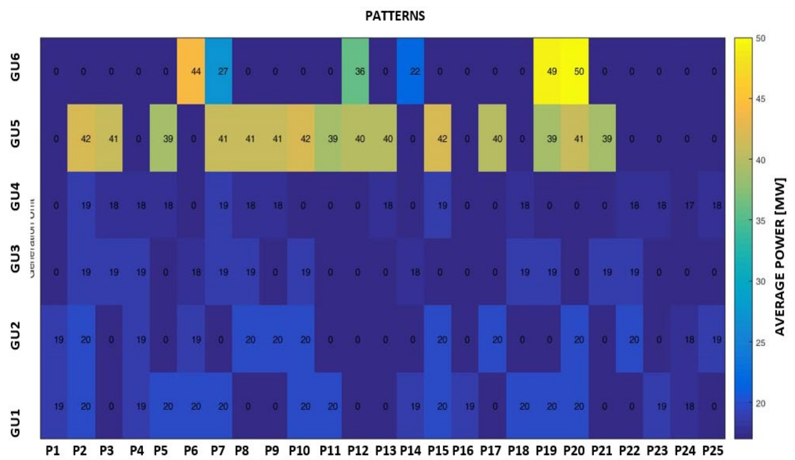


Figure 19. The patterns obtained for the hourly loading of the generation units in 2019.

This task can help improve the quality of the solutions that are obtained and the decision-making processes that are involved in the loading of the generation units.

The third task is based on an expert system involved inside the module that analyses the hourly patterns and operating conditions of the units to determine the optimal loading solution depending on the number of operating hours and the power requested by the system. Figures 20 – 22 presents the obtained results for a representative day.

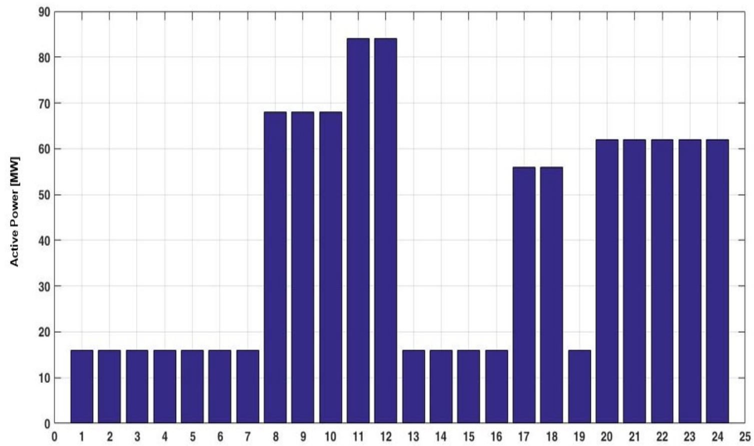


Figure 20. The active power requested by the system.

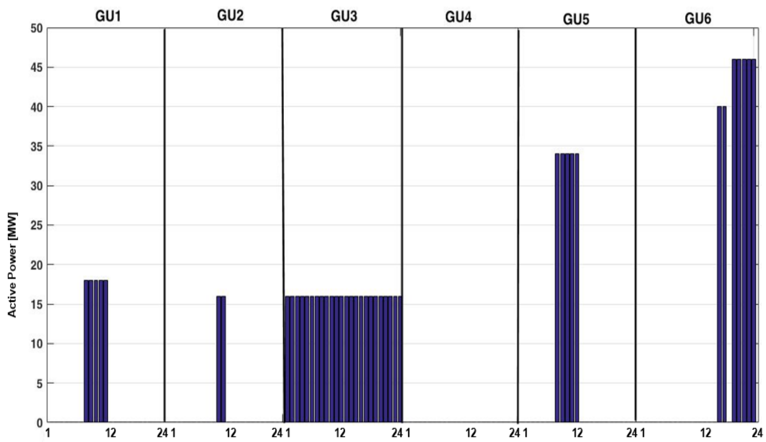


Figure 21. The active power distributed among the six GUs – the strategy adopted by the DM without the Knowledge Discovery module.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 September 2024 doi:10.20944/preprints202409.0264.v1

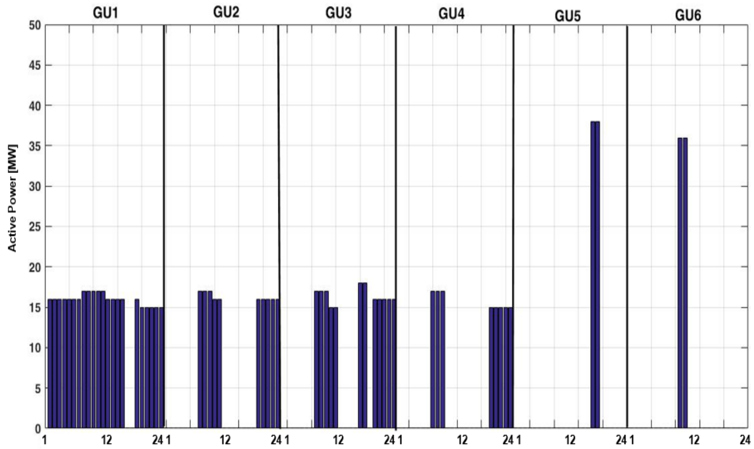


Figure 22. The active power distributed among the six GUs – the strategy adopted by the DM based on the Knowledge Discovery module.

It can be observed that the loading of the six GUs is different in the developed strategy compared with the experience-based strategy adopted by DEMA. Table 2 presents the differences between the experience-based strategy and the expert system-based developed strategy. The signification of the colors is the following:

- red color is associated with the generation units which have been loaded in the experience-based strategy but not considered in the case of the expert system-based strategy (the sign is "-");
- blue color is associated with the generation units which have been loaded in the expert-based strategy but not considered in the case of the experience-based strategy (only with the sign "+");
- green color is associated with the generation units loaded in the expert-based strategy, having the same (value "0") or having another loading in the expert-based strategy (with signs "+" or " – ");
- yellow color is associated with the generation units which have not been loaded in either strategy.

The proposed strategy adopted a loading of the first four GUs relatively uniform (between 60 and 72% of the rated power) with loading of GU5 and GU6 only for powers required by the system with higher values. These GUs will be available for ancillary services.

Table 2. The differences between the experience-based strategy and system expert-based strategy.

Hour	GU1	GU2	GU3	GU4	GU5	GU6
1	+16	0	-16	0	0	0
2	+16	0	-16	0	0	0
3	+16	0	-16	0	0	0
4	+16	0	-16	0	0	0
5	+16	0	-16	0	0	0
6	+16	0	-16	0	0	0
7	+16	0	-16	0	0	0
8	-1	+17	+1	+17	-34	0
9	-1	+17	+1	+17	-34	0
10	-1	+17	+1	+17	-34	0
11	-1	0	-1	0	-34	+36
12	-1	0	-1	0	-34	+36
13	+16	0	-16	0	0	0

14	+16	0	-16	0	0	0
15	+16	0	-16	0	0	0
16	16	0	-16	0	0	0
17	0	0	+2	0	+38	-40
18	0	0	+2	0	+38	-40
19	+16	0	-16	0	0	0
20	+15	+16	0	+15	0	-46
21	+15	+16	0	+15	0	-46
22	+15	+16	0	+15	0	-46
23	+15	+16	0	+15	0	-46
24	+15	+16	0	+15	0	-46

5. Conclusions

As the world moves toward a more sustainable energy future, the need for more reliable and dispatchable sources of electricity is increasing. An important factor that can help improve the optimal operation of hydropower plants is associated with quick data processing and extracting hidden patterns. Knowledge discovery can represent an efficient tool for addressing various challenges, among which is the optimal operation of hydropower plants.

This paper proposes a framework that combines data mining and knowledge discovery to help the transition from a traditional SCADA system to a smart one in hydropower plants. It will allow the DEMA (operators from the control room) to identify the outliers and implement effective strategies to minimize water consumption and maximize power generation. Based on the advanced statistical tools, the framework will also help them identify the optimal operating conditions for the plant. The performance has been tested in a Romanian hydropower plant using the SCADA database. It allowed the control room operators to obtain comparative information on the plant's performance over a longer time horizon (in our case study, three years). The results of the tests revealed the utility of a knowledge discovery module in helping the control room operators improve the efficiency of their operations by transitioning toward smart SCADA systems.

Now, the authors work on uncertainty modelling associated with the hourly powers requested by the system and the upstream level in the water reservoir. If the first variable depends very much on the forecasting method used, rainfall, temperature, and the operating regime of the hydro cascade of which the plant is a part represent the factors which can influence the second variable. This task is essential for the operators to determine quickly the best strategy to load the generation units.

Author Contributions: Conceptualization, G.G.; methodology, G.G. and R.G.; software, G.G.; validation, G.G., R.G. and B.C.N.; formal analysis, R.G. and B.C.N.; investigation, G.G., R.G. and B.C.N.; resources, G.G.; data curation, R.G.; writing—original draft preparation, G.G. R.G. and B.C.N.; writing—review and editing, G.G.; visualization, G.G. and R.G.; supervision, G. G.; project administration, G.G.; funding acquisition, G.G and B.C.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: "The authors declare no conflicts of interest.".

Appendix A

Table A1. The operating regime of Pattern P1 identified through the typical operating profiles divided into the intervals associated with the generating units (GU1, ..., GU6), in [MW/MWh].

Hour	GU1	GU2	GU3	GU4	GU5	GU6
1	0.00575	0.00334	0.00319	0.00091	0.01470	0.00181
2	0.00494	0.00315	0.00259	0.00092	0.01385	0.00169
3	0.00479	0.00311	0.00332	0.00093	0.01399	0.00134

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 September 2024 doi:10.20944/preprints202409.0264.v1

4	0.00458	0.00310	0.00282	0.00091	0.01260	0.00108
5	0.00451	0.00294	0.00295	0.00073	0.01334	0.00107
6	0.00562	0.00317	0.00323	0.00071	0.01596	0.00104
7	0.00656	0.00455	0.00367	0.00097	0.02124	0.00122
8	0.00651	0.00510	0.00448	0.00119	0.02218	0.00145
9	0.00698	0.00638	0.00507	0.00131	0.02658	0.00157
10	0.00697	0.00640	0.00523	0.00142	0.02702	0.00176
11	0.00690	0.00660	0.00548	0.00169	0.02510	0.00176
12	0.00719	0.00751	0.00519	0.00188	0.02749	0.00177
13	0.00854	0.00720	0.00524	0.00174	0.02658	0.00194
14	0.00910	0.00592	0.00516	0.00166	0.02548	0.00192
15	0.00909	0.00573	0.00482	0.00163	0.02595	0.00168
16	0.00778	0.00536	0.00486	0.00171	0.02440	0.00169
17	0.00683	0.00566	0.00527	0.00138	0.02433	0.00182
18	0.00693	0.00648	0.00480	0.00151	0.02508	0.00202
19	0.00792	0.00619	0.00485	0.00186	0.02556	0.00212
20	0.00825	0.00558	0.00470	0.00215	0.02507	0.00200
21	0.00879	0.00586	0.00469	0.00192	0.02590	0.00185
22	0.00936	0.00549	0.00470	0.00193	0.02723	0.00174
23	0.00729	0.00483	0.00428	0.00190	0.02345	0.00160
24	0.00580	0.00418	0.00366	0.00102	0.01833	0.00161

Table A2. The operating regime of Pattern P2 identified through the typical operating profiles divided into the intervals associated with the generating units (GU1, ..., GU6), in [MW/MWh].

Hour	GU1	GU2	GU3	GU4	GU5	GU6
1	0.00714	0.00723	0.00385	0.00000	0.02189	0.00000
2	0.00595	0.00156	0.00325	0.00058	0.00867	0.00000
3	0.00353	0.00180	0.00179	0.00051	0.00342	0.00000
4	0.00401	0.00123	0.00056	0.00099	0.00091	0.00000
5	0.00263	0.00123	0.00056	0.00091	0.00202	0.00000
6	0.00287	0.00274	0.00177	0.00148	0.00597	0.00000
7	0.00334	0.00352	0.00177	0.00119	0.00477	0.00000
8	0.00547	0.00317	0.00116	0.00115	0.00240	0.00000
9	0.00720	0.00184	0.00243	0.00115	0.00764	0.00000
10	0.00775	0.00155	0.00109	0.00167	0.00311	0.00000
11	0.00583	0.00298	0.00225	0.00109	0.00772	0.00000
12	0.00741	0.00053	0.00211	0.00058	0.00998	0.00000
13	0.00683	0.00294	0.00141	0.00058	0.01011	0.00000
14	0.00674	0.00498	0.00262	0.00148	0.01545	0.00000
15	0.00816	0.00514	0.00148	0.00106	0.02402	0.00000
16	0.01109	0.00683	0.00524	0.00102	0.04126	0.00000
17	0.01380	0.00830	0.00751	0.00510	0.05050	0.00000
18	0.02052	0.01275	0.00860	0.00575	0.05867	0.00000
19	0.01872	0.01230	0.01201	0.00280	0.06564	0.00000
20	0.01555	0.01080	0.01157	0.00328	0.05483	0.00000
21	0.01391	0.00992	0.00739	0.00281	0.04916	0.00000
22	0.01258	0.00719	0.00516	0.00181	0.04508	0.00000
23	0.00832	0.00541	0.00366	0.00146	0.03031	0.00000
24	0.00575	0.00512	0.00299	0.00146	0.01821	0.00000

Table A3. The operating regime of Pattern P3 identified through the typical operating profiles divided into the intervals associated with the generating units (GU1, ..., GU6), in [MW/MWh].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 September 2024 doi:10.20944/preprints202409.0264.v1

Hour	GU1	GU2	GU3	GU4	GU5	GU6
1	0.01713	0.00741	0.00655	0.00192	0.01527	0.00062
2	0.01045	0.00646	0.00763	0.00154	0.00945	0.00062
3	0.00950	0.00852	0.00764	0.00099	0.00676	0.00062
4	0.00903	0.00841	0.00805	0.00058	0.00552	0.00062
5	0.00850	0.00765	0.00798	0.00084	0.00623	0.00062
6	0.01580	0.00840	0.00871	0.00144	0.00901	0.00062
7	0.01595	0.00643	0.00792	0.00255	0.01280	0.00095
8	0.01405	0.00548	0.00728	0.00382	0.01031	0.00095
9	0.01227	0.00727	0.01146	0.00465	0.01336	0.00095
10	0.01166	0.00675	0.01059	0.00456	0.01138	0.00095
11	0.01041	0.00494	0.00936	0.00475	0.00521	0.00095
12	0.01072	0.00556	0.00967	0.00458	0.00330	0.00057
13	0.01224	0.00515	0.01046	0.00525	0.00272	0.00058
14	0.01063	0.00528	0.00961	0.00463	0.00108	0.00058
15	0.00989	0.00507	0.00890	0.00402	0.00073	0.00096
16	0.01135	0.00445	0.00857	0.00504	0.00217	0.00066
17	0.01331	0.00520	0.01076	0.00451	0.00385	0.00066
18	0.01266	0.00598	0.00919	0.00431	0.00145	0.00066
19	0.01525	0.00567	0.01098	0.00511	0.00382	0.00064
20	0.01608	0.00567	0.00932	0.00467	0.00467	0.00064
21	0.01984	0.00560	0.01101	0.00747	0.01211	0.00068
22	0.02218	0.00490	0.01453	0.00588	0.01447	0.00068
23	0.02410	0.00390	0.01394	0.00540	0.01314	0.00068
24	0.02336	0.00414	0.01110	0.00468	0.00939	0.00033

Table A4. The operating regime of Pattern P3 identified through the typical operating profiles divided into the intervals associated with the generating units (GU1, ..., GU6), in [MW/MWh].

Hour	GU1	GU2	GU3	GU4	GU5	GU6
1	0.01024	0.00938	0.00177	0.00000	0.05302	0.00000
2	0.00779	0.00898	0.00170	0.00226	0.05455	0.00000
3	0.00762	0.00958	0.00215	0.00226	0.05302	0.00083
4	0.00750	0.00668	0.00215	0.00000	0.04260	0.00083
5	0.01061	0.00667	0.00215	0.00000	0.04609	0.00000
6	0.01355	0.00796	0.00170	0.00000	0.05410	0.00000
7	0.01236	0.02001	0.00177	0.00000	0.06093	0.00000
8	0.00967	0.00388	0.00045	0.00000	0.02815	0.00000
9	0.00640	0.00420	0.00263	0.00000	0.02356	0.00000
10	0.00629	0.00347	0.00319	0.00080	0.01695	0.00000
11	0.00747	0.00262	0.00416	0.00000	0.00504	0.00000
12	0.00540	0.00275	0.00133	0.00000	0.00239	0.00000
13	0.00413	0.00272	0.00000	0.00000	0.00143	0.00000
14	0.00335	0.00181	0.00000	0.00063	0.00120	0.00000
15	0.00171	0.00235	0.00167	0.00063	0.00251	0.00000
16	0.00221	0.00313	0.00250	0.00000	0.00566	0.00000
17	0.00300	0.00236	0.00246	0.00000	0.00937	0.00000
18	0.00631	0.00512	0.00550	0.00372	0.01772	0.00083
19	0.00807	0.00492	0.00445	0.00362	0.02117	0.00083
20	0.00883	0.00515	0.00445	0.00301	0.02311	0.00083
21	0.01128	0.00748	0.00488	0.00131	0.02587	0.00083
22	0.01156	0.00647	0.00575	0.00060	0.02382	0.00000
23	0.00453	0.00568	0.00595	0.00000	0.01412	0.00000

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 September 2024 doi:10.20944/preprints202409.0264.v1

24	0.00193	0.00529	0.00431	0.00000	0.01227	0.00000

Table A5. The patterns regarding the hourly loading of the generation units GU1 – GU6 in analysed three years period.

Pattern	2017						2018						2019					
	GU1	GU2	GU3	GU4	GU5	GU6	GU1	GU2	GU3	GU4	GU5	GU6	GU1	GU2	GU3	GU4	GU5	GU6
P1	0	0	0	0	34	0	0	21	20	20	40	0	20	0	0	0	39	0
P2	16	16	16	16	33	0	18	0	0	0	0	0	19	19	19	18	0	0
P3	17	0	0	0	35	31	20	20	20	18	0	0	20	19	18	0	0	44
P4	15	16	16	0	33	32	20	20	20	20	39	0	0	20	19	18	42	0
P5	0	0	17	0	35	0	19	19	0	18	38	0	20	0	19	0	39	49
P6	18	0	17	17	35	0	19	0	19	18	38	0	19	0	0	18	0	0
P7	0	18	0	0	36	0	0	20	20	18	0	0	20	20	0	19	42	0
P8	0	17	16	16	34	0	0	20	19	0	40	0	0	0	19	18	41	0
P9	17	0	0	17	34	0	19	0	18	18	0	0	18	18	0	17	0	0
P10	0	0	16	16	34	0	19	0	18	0	38	0	20	20	0	0	41	50
P11	17	0	17	0	34	30	0	0	0	0	36	0	20	20	19	19	42	0
P12	17	0	0	0	0	32	19	0	0	0	38	0	0	0	19	0	39	0
P13	17	17	16	0	0	0	0	17	16	17	0	0	0	20	0	18	40	0
P14	0	17	0	16	34	0	0	0	18	17	0	0	20	20	19	0	42	0
P15	18	0	17	0	0	31	0	18	0	0	36	0	20	0	19	19	41	27
P16	18	18	0	17	35	0	19	20	19	0	39	0	19	19	0	0	0	0
P17	16	16	16	0	32	31	19	0	17	17	37	0	0	19	0	18	0	0
P18	0	0	16	0	0	31	19	19	0	18	0	0	20	0	0	18	39	0
P19	0	17	16	15	0	0	0	20	10	18	39	0	0	0	0	0	40	36
P20	18	18	17	17	0	0	0	0	18	0	38	0	19	0	19	18	0	22
P21	0	15	17	17	0	0	0	0	17	17	35	0	0	20	19	18	0	0
P22	17	17	0	0	33	32	0	0	0	17	35	0	0	0	18	18	0	0
P23	0	0	0	15	0	32	19	0	0	18	0	0	19	0	0	0	0	0
P24	17	17	0	0	0	30	19	19	19	0	0	0	0	20	19	0	41	0
P25	17	17	0	17	0	0	17	18	0	0	36	0	0	0	0	18	40	0

Table A6. Comparison between the features of the patterns in analysed three years period.

Patterns	Operating Time									Powers required by the system [MW]								
	Hours			Days			Hours/Day			Minimum value			Average value			Maximum value		
	2017	2018	2019	2017	2018	2019	2017	2018	2019	2017	2018	2019	2017	2018	2019	2017	2018	2019
P1	432	471	323	128	35	49	3	`13	7	30	75	42	34	102	63	40	107	59
P2	360	235	539	54	81	55	7	3	10	71	15	60	82	18	79	110	22	76
P3	810	908	276	152	66	42	5	14	7	44	64	47	53	78	108	88	85	58
P4	165	424	467	23	36	45	7	12	10	60	90	89	86	119	100	108	129	99
P5	267	166	276	60	27	29	4	6	10	45	75	67	52	94	129	60	105	79
P6	116	288	47	30	40	16	4	7	3	42	75	34	65	95	40	101	107	37
P7	488	137	372	99	11	34	5	12	11	43	45	88	54	58	103	60	64	101
P8	235	313	429	52	40	52	5	8	8	57	60	63	69	78	80	100	84	78
P9	115	165	209	29	45	27	4	4	8	60	30	48	67	46	59	81	64	54
P10	95	397	129	16	59	30	6	7	4	60	60	73	66	75	132	73	84	82
P11	452	425	516	101	112	38	4	4	14	58	30	103	69	36	122	100	44	120
P12	287	806	101	103	118	16	3	7	6	13	45	47	18	57	62	51	64	58
P13	151	119	131	30	46	19	5	3	7	35	0	55	49	19	81	59	39	71
P14	156	91	149	25	33	21	6	3	7	45	15	92	56	23	102	76	42	100
P15	183	418	243	49	67	22	4	6	11	26	45	86	36	55	128	68	64	99
P16	485	159	186	91	34	42	5	5	4	58	78	28	72	96	42	97	106	38
P17	92	216	131	15	39	42	6	6	3	62	60	15	89	73	40	126	84	23

 doi:10.20944/preprints202409.0264.v1

P18	71	198	355	30	48	40	2	4	9	14	30	58	19	45	82	50	61	77
P19	102	71	280	44	17	64	2	4	4	15	60	30	18	76	90	34	90	40
P20	119	201	191	15	41	32	8	5	6	50	45	33	68	56	93	94	63	49
P21	23	88	117	11	16	20	2	6	6	28	60	35	35	69	58	47	84	56
P22	161	101	58	20	31	21	8	3	3	46	45	15	82	53	38	106	63	27
P23	55	52	121	25	16	48	2	3	3	0	30	15	10	37	21	33	41	19
P24	92	172	34	40	33	8	2	5	4	30	39	75	36	55	82	62	65	80
P25	76	293	179	22	58	44	3	5	4	30	60	46	40	71	61	58	85	58

References

1. Parvez, I.; Shen, J.; Hassan, I.; Zhang, N. Generation of Hydro Energy by Using Data Mining Algorithm for Cascaded Hydropower Plant. *Energies* **2021**, *14*, 298.
2. Garbea, R.; Scarlatache, F., Grigoras, G., Neagu, B.C. Extracting the Operating Characteristics of Hydropower Plants Using a Clustering-based Efficient Methodology, In Proceedings of IEEE 9th International Conference on Modern Power Systems (MPS), Cluj-Napoca, Romania, 16-17 June 2021.
3. Eker, O. F. Data Science for Industry: Hydropower Condition Monitoring and Predictive Maintenance, 2022, Available on-line: https://medium.com/@omerfarukeker/data-science-for-industry-hydropower-condition-monitoring-and-predictive-maintenance-49952215fdd7, last accessed 2024/07/31.
4. Betti, A.; Crisostomi, E.; Paolinelli, G.; Piazzi, A.; Ruffini, F.; Tucci, F. Condition Monitoring and Predictive Maintenance Methodologies for Hydropower Plants Equipment. *Renewable Energy* **2021**, *171*, 246-253.
5. Essenfelder, A. H.; Larosa, F.; Broccoli, D.; Mazzoli, P.; Bagli, S.; Luzzi, V.; Mysiak, J.; dalla Vallw, F. Smart Climate Hydropower Tool: A Machine-Learning Seasonal Forecast-ing Climate Service to Support Cost–Benefit Analysis of Reservoir Management. *Atmosphere* **2020**, *11*, 1305.
6. Feng, Z.K.; Niu, W.J.; Zhang, R.; Wang, S.; Cheng, C.-T. Operation Rule Derivation of Hydropower Reservoir by K-Means Clustering Method and Extreme Learning Machine Based on Particle Swarm Optimization. *Journal of Hydrology* **2019**, 576, 229-238.
7. Zhang, F.; Guo, J.; Yuan F.; Qiu Y.; Wang P.; Cheng F.; Gu Y. Enhancement Methods of Hydropower Unit Monitoring Data Quality Based on the Hierarchical Density-Based Spatial Clustering of Applications with a Noise–Wasserstein Slim Generative Adversarial Imputation Network with a Gradient Penalty. *Sensors* **2024**, *24*, 118.
8. Luo, W.; Xu, J.; Zhou, Z. Mobile Information Systems, Retracted: Design of Data Classification and Classification Management System for Big Data of Hydropower Enterprises Based on Data Standards, Mobile Information Systems, 2022, Available on-line: https://onlinelibrary.wiley.com/doi/10.1155/2022/8103897, last accessed 2024/07/31.
9. Ahmed, I.; Dagnino, A.; Bongiovi, A.; Ding, Y. Outlier Detection for Hydropower Generation Plant, In Proceedings of the IEEE 14th International Conference on Automation Science and Engineering (CASE), Munich, Germany, 20 – 24 August 2018.
10. Valencia, A.M.; Caratar, J.; Caicedo, G.; Chamorro, C. Proposal for a KDD-Based Procedure to Obtain a Set of Intelligent Systems Training Applied to the Identification of Failures in Hydroelectric Power Plants. *Journal of Applied Research and Technology* **2021**, *18*, 376–389.
11. Zhang, W.; Ge, Y.; Liu, G.; Qi, W.; Xu, S.; Peng, Z.; Li, Y. Clustering and Decision Tree Based Analysis of Typical Operation Modes of Power Systems. *Energy Reports* **2023**, *9*, 60-69.
12. Garbea, R.; Scarlatache, F.; Grigoras, G.; Neagu, B.C. Integration of Data Mining Techniques in SCADA System for Optimal Operation of Hydropower Plants, In Proceedings of the IEEE 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Pitesti, Romania, 1 – 3 July 2021.
13. Sahin, M.E.; Ozbay Karakus, M. Smart Hydropower Management: Utilizing Machine Learning and Deep Learning Method to Enhance Dam's Energy Generation Efficiency. *Neural Comput & Applic* **2024**, *36*, 11195–11211.
14. Shu, X.; Ye, Y. Knowledge Discovery: Methods from Data Mining and Machine Learning. *Social Science Research* **2023**, *110*, 102817.
15. Monika; Shauib, M. Implementation Platforms and Strategy for the Knowledge Discovery from the Data, In Proceedings of the International Conference on Computational Modelling, Simulation and Optimization (ICCMSO), Pathum Thani, Thailand, 23 – 25 December 2022.
16. Ghongade, T. G.; Khobragade, R. N. Evaluation on Utilization and Emaciation of Data Mining Techniques in Information System, In Proceedings of the OPJU International Technology Conference on Emerging Technologies for Sustainable Development (OTCON), Raigarh, Chhattisgarh, India 8 – 10 February 2023.
17. Järvinen, P.; Siltanen, P.; Kirschenbaum, A. Data Analytics and Machine Learning. In *Big Data in Bioeconomy*; Södergård, C., Mildorf, T., Habyarimana, E., Berre, A.J., Fernandes, J.A., Zinke-Wehlmann, C., Eds.; Publisher: Springer Nature, Cham. Cham, Switzerland, 2021; pp. 129–146.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 September 2024 doi:10.20944/preprints202409.0264.v1

18. Garbea, R., Grigoras, G. Clustering-Using Data Mining-based Application to Identify the Hourly Loading Patterns of the Generation Units from the Hydropower Plants, In Proceedings of the IEEE International Conference and Exposition on Electrical and Power Engineering (EPE), Iasi, Romania 20 – 22 October 2022.
19. Odrynska A. What is Data Mining: Definition, Process, Techniques and Role in Business Intelligence, 2023, Available on-line: https://www.alphaservesp.com/blog/what-is-data-mining-definition-process-techniques-and-business-intelligence, last accessed 2024/07/31.
20. Mirzargar, M.; Whitaker, R. T.; Kirby, R. M. Curve Boxplot: Generalization of Boxplot for Ensembles of Curves. *IEEE Transactions on Visualization and Computer Graphics* **2023**, *20*, 2654-2663.
21. Chelaru, E.; Grigoras, G. Decision Support System to Determine the Replacement Ranking of the Aged Transformers in Electric Distribution Networks, In Proceedings of IEEE 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI) Proceedings, Bucharest, Romania 25 – 27 June 2020.
22. Neagu, B.C.; Grigoras, G., Scarlatache, F. Outliers Discovery from Smart Meters Data Using a Statistical Based Data Mining Approach, In Proceedings of IEEE 10th International Symposium on Advanced Topics in Electrical Engineering (ATEE), Bucharest, Romania, 23 – 25 April 2017.
23. Wang, Z.; Wang, S.; Zhang, S.; Zhan, J. An Expert System Based on Data Mining for a Trend Diagnosis of Process Parameters. *Processes* **2023**, *11*, 3311.
24. Dandea, V; Grigoras, G. Expert System Integrating Rule-Based Reasoning to Voltage Control in Photovoltaic-Systems-Rich Low Voltage Electric Distribution Networks: A Review and Results of a Case Study. *Applied Sciences* **2023,** *13*, 6158.