Vol. 05, No. 1 (2023) 39-48, doi: 10.24874/PES05.01.004

Proceedings on Engineering Sciences

www.pesjournal.net

STUDY OF FAILURE STATISTICS OF CAVITATORS IN THE FUEL OIL FACILITIES THROUGH THE APPLICATION OF REGRESSION AND CLUSTER ANALYSIS

Pavel Shcherban[1]
Andrey Sokolov
Reda Abu Hamdi

Received 05.10.2022.
Accepted 18.12.2022.
UDC – [658.588:665.6/.7]:519.237

Keywords:

Regression analysis, Cluster analysis, least squares method, K-means clustering, Cavitation equipment, Oil and gas equipment.

A B S T R A C T

An essential, and at the same time poorly covered in scientific publications topic is the failures of cavitation systems. Practice of using cavitators at fuel oil facilities at thermal power plants in Kaliningrad region shows that these technical devices can fail rather often. In this study, cavitators failures problem was analyzed by survey of accumulated statistical data concerning the parameters of cavitators operation in fuel oil farms and the moments of failure occurrence. The regression and cluster analysis, which was used for data processing, allowed to determine the relationship between failure types and influencing factors and to rank the factors by the degree of their impact on the cavitation equipment. Based on the results of mathematical and data analysis, proposals for ensuring better technical reliability of cavitators, reorganization of its maintenance system and reduction of the number of failures were developed.

1. INTRODUCTION

Regression analysis (F. Galton, C. Pearson, E. Sluchsky, etc.) developed in the second half of the 19th century and cluster analysis (H. Steingauz, S. Lloyd, J. Hodges, R. Sokel, etc.) developed in the middle of the 20th century have found wide application in the assessment of the technical condition of machines and mechanisms, determination of causes of failure and establishment of close relationships between failure frequencies and influencing factors. These mathematical tools are often used in the analysis of statistical data concerning technological processes of conveyor production, equipment repairs, analysis of deviations in welding processes accuracy, installation, assessment of interconnection of product quality deviations with influencing factors. Arrays of statistical data obtained in various mechanical, thermobaric, chemical, complex processes are processed with similar tools and serve as a basis for engineering staff decision-making.

The data obtained during operation of fuel oil farms, including data on cavitators failures, allowed to assess the degree of influence of various factors on the frequency of failures, as well as to identify the cause-effect relationship between the factors affecting the cavitators and equipment shutdown / accidents, using regression and cluster analyses.

[1] Corresponding author: Pavel Shcherban
Email: ursa-maior@yandex.ru

It is worth noting that the use of cavitators for fuel reserve (fuel oil) treatment allows:

- To increase the energy value of fuel oils and to maintain their qualitative indicators.
- -To decrease the content of substances that reduce the service life of furnaces and boilers, as well as to reduce the toxicity of the products of combustion (Avvaru et. al. 2018). When discharged bubbles collapse in fuel oil, a shock wave occurs, which breaks chains of molecules and substances and destroys its physical and chemical structure. After such treatment the number of low molecular weight compounds increases and a new structure is formed. As a result, after cavitation treatment and subsequent chemical reactions, the amount of impurities containing free sulfur and phosphorus decreases, which leads to a decrease in the corrosion rate (Carpenter et. al. 2018). The use of cavitators is complicated due to failures, which is a consequence of a number of factors: production, organizational, quality of materials used, accuracy of equipment diagnostics, frequency and completeness of maintenance and repairs (Carpenter et. al. 2017). As a result of using regression and cluster analysis, the causes of premature failure of cavitators, the factors affecting the frequency of failure of this equipment have been established. The statistical data obtained as a result of mathematical processing allowed to develop organizational, managerial and technical recommendations for the operators of cavitation systems in order to reduce the number of accidents and to ensure the continuity of the equipment operation.

2. THE SPECIFICS OF THE APPLICATION OF REGRESSION AND CLUSTER ANALYSIS FOR STATISTICAL DATA PROCESSING ON EQUIPMENT FAILURES

Regression analysis is a research method that allows to get a functional empirical dependence of one random variable on another on the basis of statistical processing of experimental data array. The main task of regression analysis is to make a mathematical model (regression curve) describing the relationship between variables: dependent, random and one or more independent random or non-random parameters.

In the simplest form of regression analysis for two variables, there are n pairs of experimentally found or observed values (Xi; Yi) and the dependence y = f(x) must be obtained in analytical form (Hagiwara et. al. 2008). First of all, for visual analysis, these pairs of values are plotted on an X-Y diagram. Then, based on the visual analysis and/or a priori information, one or another type of function (polynomial, exponential, sinusoidal, etc.) is chosen, which, from the researcher's point of view, most likely describes the dependence of y(x). After choosing the type of function, we need to find the numerical coefficients at which the chosen function best corresponds to the observation results and evaluate the correctness of the chosen mathematical model.

The calculation of the numerical values of the coefficients, at which the chosen function in the best way corresponds to the results of observations can be carried out in several ways. The most common and easiest method is the method of least squares (MLS) (Draper and Smith 1998). Alternative methods are the maximum likelihood estimation (Rossi 2018), the method of moments (Bowman and Shenton 2004), and some others. In this study, the method of least squares is used, according to which the coefficients of the approximating function y=f(x) are chosen so that the sum of the squares of deviations $(Yi – f (Xi))^2$ would be minimal (Denham 1995).

Application of regression analysis in this case is complicated due to the difficulty of choosing a value that could be interpreted as Y, the variable against which the values of the approximating function would be compared. Usually Y is a continuous variable. In case of cavitators, the failure event is binary by its nature: either there is a failure, or there is no failure, and we trace only the values of the influencing parameters, at which this failure occurred. Besides, we had information about the circumstances of only about two dozens cavitator failures, so we cannot speak about any correct probability of failure, which could serve as a continuous variable Y. Due to the mentioned circumstances, the following approach was applied. For each of the influencing parameters X a series containing information about the values of the parameters in case of failure was sorted in ascending order of X and the number of the member in such an ordered series played the role of the variable Y. i.e. yi is the number of failures at which the value of the analyzed parameter is more or equal to xi. Application of cluster analysis for processing of statistical data on machinery failures is also an effective tool, which allows to stratify data, to determine mutual influence of parameters. This type of analysis has a number of features. Thus, it is possible to apply different approaches of cluster analysis to data processing. Hierarchical approach allows to analyze multiple objects with some connectivity (for example, types of equipment defects, types of repair materials, types of chemical reagents for maintenance) and either separate them (divisive approach) or combine them (agglomerative approach), depending on the problem to be solved (Wang et. al. 2019).

The approach based on the Kohonen neural network (Kohonen maps) - allows us to process information on a set of elements, grouping the elements into relatively homogeneous groups. After that the indicator matrix is formed. The Kohonen map is then built based on that matrix. During the preparation of the Kohonen map the vector of input array is compared with vectors of neurons of the active layer. This comparison is done using the proximity function. The neuron of the active layer, for which the value of the proximity function between vector of input array and vector of neurons of the active layer is maximal, is considered a "winner". Thus it becomes possible to stratify data. For example, it is possible to process data on failures and wear of pipeline networks, on failures of instrumentation (sensors systems) - to select elements, which require overhaul / replacement, routine repair, additional maintenance, and those which do not require operations beyond current procedures. At the same time there are some limitations for using Kohonen maps. For example, the final result of the neural network in the map depends on the initial settings of the network, i.e., the outliers in the initial processed data set can significantly affect the accuracy of the results. DBSCAN (Density-based spatial clustering of applications with noise) approach allows to effectively filter noises and outliers in the stream of processed data. Since this algorithm is focused on finding "neighbors" in each of the analyzed data elements, if there is no neighborhood, the element is discarded as "noise". The method is effective when processing telemetry data, remote control of technical condition of the sensor system, processing the results of vibrodiagnostic control of operating equipment. Removal of apparent "noise" elements allows to clear the sample and form clear clusters, which improves analysis results without loss of data quality. However, there is a peculiarity of its application. This approach can be applied to data obtained from technical devices functioning in conditions of increased generation of non-standard and random "noise" data - running turbines, ship propellers, engines, otherwise at stable constant operation of equipment element (linear parts of pipelines, tanks, reservoirs) there is a risk to take the outlier single data as "noise", while they can signal an accident or failure in these devices.

Approach based on k-means clustering (one of the most common types of cluster analysis). The idea of the method is that, having a set of statistical data and initially assuming the number of clusters in the processed array, it is possible by consecutive iterations to calculate the center of mass for each of the clusters. The vectors for each of the data elements are partitioned into clusters again according to which of the new centers turns out to be closer according to the chosen metric. The algorithm terminates when there is no change in the cluster center of mass at a certain iteration (Putra et. al. 2021).

As a result, it is possible to establish the frequency and probability of realization of a particular event, its weight and the relationship with other factors (depending on the characteristics of the formed clusters, the position of the center of mass and the density of event distribution). This method is convenient for the analysis of failures and malfunctions of the equipment operating under standard conditions, without sharp loads and changes in the pattern of acting external forces. In this case, it is not required to reduce the number of noises and surges, besides, rather definite correlations between the causes of failures and accidents and the acting factors or a group of factors are established (Protalinskii et. al. 2013). For the considered problem of failures and stoppages of cavitators operation in fuel oil facilities we will apply this method.

Next, we will consistently present the problems of cavitation equipment operation, data on failures, and apply regression and cluster analysis for processing the available statistical information.

3. THE CAVITATION SYSTEM FAILURES AND THE ANALYSIS OF THE REASONS OF THEIR OCCURRENCE. FORMATION OF THE STATISTICAL DATABASE

Cavitation systems used in tank farms include a mechanical part consisting of a cavitating body, which creates an artificial vacuum of liquid flow, as well as an electrical part, which is responsible for control and management of the process. During 2018-2021 the statistical data on failures of cavitation equipment has been collected at a number of heat power plants of Kaliningrad region. As a result, it was determined that the most frequent failures in mechanical part of the system are fuel oil leakages in flange nipple connections, and in electric part - activation of emergency protection in one of electric consumers (Figure 1).

Following the Pareto principle, we decided to study the causes of the two most frequent types of cavitator failures. To understand the causes of leakages in the cavitators flange nipple connectors, as well as the causes of activation of emergency protection, we made diagrams of the main factors, the implementation of which could lead to the occurrence of such failures (Figure 2 and Figure 3).

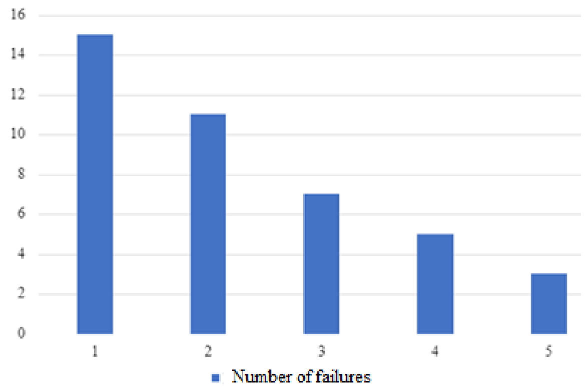


Figure 1. Number of failures (by type) registered during cavitation systems operation in 2018-2021.

1-Liquid leakages in the flange nipple connectors; 2-An emergency protection activation on one of the electric consumers; 3-Lack of readings, incorrect readings of pressure gauges; 4-Lack of readings of flow meters; 5-Fan not working in the control room

Based on the records of maintenance and repair logs of cavitation systems, acts of technical examination of the equipment performance, recorded indicators of monitoring systems, a pool of statistical data was formed. Namely, for each fact of leakage, the depth of metal corrosion, gasket density and tightening torque corresponding to this event were recorded. For each fact of activation of the protection system we recorded the corresponding transformer element temperature, current and overvoltage of the transformer.

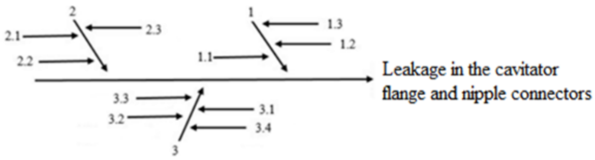


Figure 2. Factors influencing the occurrence of leakage in the cavitator flange and nipple connectors (failure due to mechanical indicators - leakage)

1. Loose fasteners; 1.1. Personnel non-compliance with installation technology; 1.2 Backlash exceeding permissible values; 1.3 Wear of connection hole threads; 2 Gasket defects; 2.1 Gasket crumpling; 2.2 Gasket bruising; 2.3 Poor quality of the detail; 3 Leakage caused by detail failure; 3. 1 Chemical degradation of flange joint composite; 3.2 Mechanical damage; 3.3 Deformation due to vibration and bending loads; 3.4 Corrosion failure of metal elements of gasket or flange.

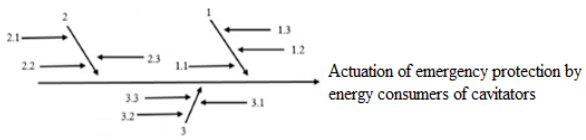


Figure 3. Group of factors affecting the actuation of emergency protection by energy consumers of cavitators (failure by electrical indicators - emergency activation)

1. circuit breaker actuation; 1.1. short circuit; 1.2. circuit breaker failure; 1.3. remote control system failure; 2. electric thermal relay actuation; 2.1. relay failure; 2.2. electric motor overload; 2. 3. Problem with the pump shaft seal sensors (no feedback into the system); 3. Pressure drop at the pump outlet is lower than allowed (pressure drop sensor actuation); 3.1. faulty pressure gauge; 3.2. faulty alarm circuit; 3.3. faulty pump operation sensor

In the course of the study we faced the task of mathematical processing of the obtained data and identification of relationships between the two main recorded types of cavitator failures, on the one hand, and the output of the given factors beyond the allowable values (according to technical regulations), on the other hand.

The formed data pool was stratified, first by the type of cavitator failure, then within each type, by the influencing factors. The cavitator failure rate was chosen as the dependent variable (i.e., "y"), more precisely, it was the fixed number of failures in case the value of one or another independent variable exceeds the boundary allowable value in accordance with the working documentation for the equipment. The recorded value of the factor parameter at the moment of failure detection and the range of admissible values were indicated. Further it was decided to use regression and cluster analysis for data processing. This approach allowed us to determine the relative impact of the factors deviation from the allowable values on the failure of the cavitation equipment. As a result, it is possible to identify the weight of factors that led to the occurrence of cavitator failure and, as a consequence, to develop recommendations for minimizing their impact.

4. APPLICATION OF REGRESSION ANALYSIS FOR DETERMINING THE RELATIONSHIP BETWEEN CAVITATOR FAILURES AND INFLUENCING FACTORS

As mentioned above, the main task of the regression analysis is to establish the functional dependence of one value on another. In our case it is to establish the functional dependence of the number (frequency) of failures on the numerical value of a certain influencing parameter (Mohan and Sivakumar 2020). To carry out the regression analysis, we use the data obtained during fixing of leaks in the flange-nipple connectors of cavitators, as well as in case of activation of emergency protection by energy consumption of cavitators.

Let us compare the data on the frequency of leakage occurrence and the density of used gaskets, the depth of metal corrosion, as well as the degree of tightening of flange-nipple connectors fasteners. For failures caused by activation of emergency power protection, we compare data on frequency of failures, temperature at which the failure occurred, current strength and excess of permissible voltage.

After ranking the number of failures for each of the analyzed mechanical parameters, we obtain linear dependences of the frequency of leakages in the cavitators caused by each of the influencing factors (Figure 4).

To assess the accuracy of the model (to determine how well the approximating function corresponds to the experimental data) the following statistical parameters are most frequently used: BIAS, root mean square error (RMSE), correlation coefficient (R) and coefficient of determination (square of correlation coefficient, R2) (Kafle 2019). By ranking the number of failures for each of the analyzed parameters, depending on the energy consumption, we obtain the dependences of the frequency of emergency protection activation caused by each of the influencing factors, shown in (Figure 5).

BIAS is the difference between the average calculated and measured values during the entire observation period:

$$BIAS = \frac{1}{n}\left(\sum_{i=1}^{n} f(x_i) - \sum_{i=1}^{n} y_i\right) \quad (1)$$

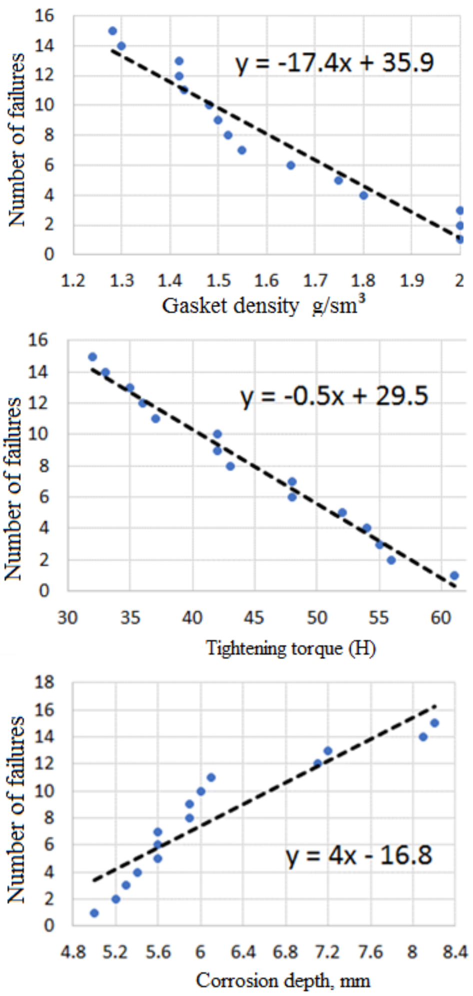


Figure 4. Observed values of leakage frequency in cavitators as a function of: gasket density, tightening torque and metal corrosion depth.

A positive BIAS value means that the model gives on average overestimated values, and a negative value means that it gives on average underestimated values. In other words, BIAS serves as an indicator of systematic error. The root mean square error (RMSE) is the square root of the sum of squares of the differences between the observed and calculated results, divided by the total number of observations:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - f(x_i))^2}{n}}. \quad (2)$$

In fact it is an evaluation of the effectiveness of the method of least squares (Frumin 2020).

In the future, if there is significantly more actual data on equipment failures, more complex models could be used instead of simple linear models. For example, let us consider a piecewise linear model for the dependence of the number of failures on the gasket density. We will assume that if the density of the gasket lies within the normal range (1.5-2 g/cm^3) the approximation is made by one linear function, and if outside these limits - by another one.

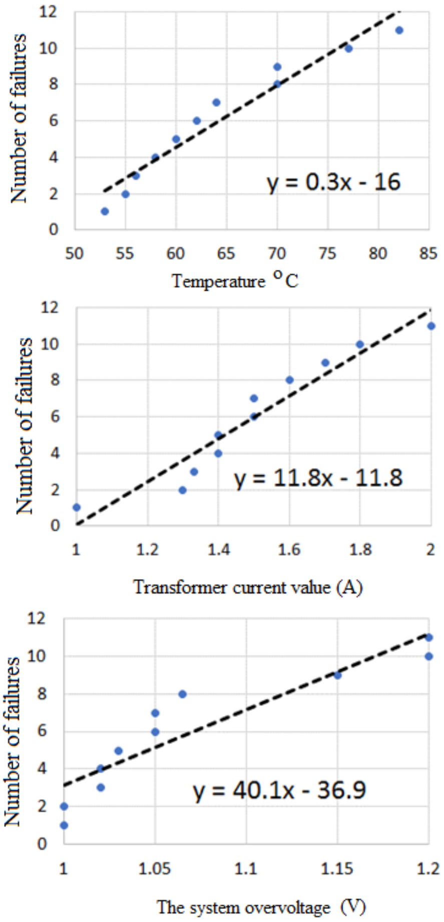


Figure 5. Observed values of the emergency protection activation frequency as a function of: the transformer element temperature, the transformer current value and the system overvoltage.

The least squares method allows us to draw a piecewise linear model, shown in (Figure 6) (Markelov 2015).

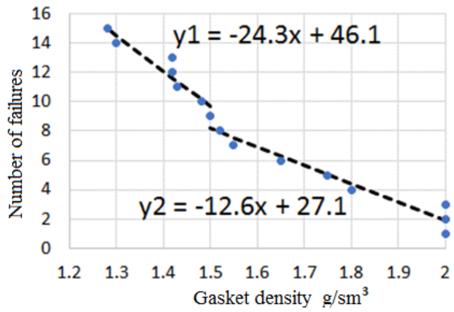


Figure 6. Piecewise linear model of the dependence between the number of failures and the gasket density

It follows from Table 1 that the accuracy of such a model by RMSE parameters and determination coefficient is significantly higher than the simple linear model presented in Figure 4.

Table 1. Statistical parameters for assessing the correctness of linear and piecewise linear models

linear			**piecewise linear**		
BIAS	RMSE	R^2	BIAS	RMSE	R^2
0	1,13	0,93	0	0,58	0,95

Thus, with a gradual accumulation of statistical data on the type and number of cavitator failures at thermal power plants, it is possible to recalculate and move to more accurate piecewise linear regression models (Seraya and Demin 2012). At the same time, from the technical point of view, the results of the current regression modeling involving the linear dependencies show that in the mechanical part, the most clear dependence between the number of failures and deviations from the permissible values of parameters is observed in the tightening torque of the gasket. If the gasket is not tightened, there is free space through which leakage occurs or water and air can enter to form corrosion zones. If the gasket is overtightened, it becomes more stressed and brittle - it wears down and cracks faster, causing even more leakage and damage. So in the mechanical part, it is this factor that plays the greatest role in the occurrence of failure. As for the electrical part, the current strength and the temperature of the electrical elements of the cavitation equipment are of greater importance. These parameters are interrelated, and at the same time the growth of failures due to these two parameters is more intensive, than with increase of resistance. It is connected with the specificity of used materials, for resistor and sensor systems in cavitators. The obtained conclusions indicate the necessity of revising the characteristics of the systems by the manufacturers and making them more resistant to current fluctuations and temperature changes.

5. APPLICATION OF CLUSTER ANALYSIS TO DETERMINE THE RELATIONSHIP BETWEEN CAVITATOR FAILURES AND INFLUENCING FACTORS

The applied regression analysis allowed us to examine the individual influence of deviation of one of the operating parameters on the frequency of cavitator failures. However, undoubtedly, the equipment is affected by a complex of factors, and to study the synergistic effect of such deviations, we use cluster analysis. Cluster analysis allows to reveal the structure of statistical data and establish the presence of atypical objects or specific clusters (Sharma and Chandra 2020). Determination of clustering centers will allow to establish the value of scatter of data on each of the clusters, i.e. to establish the average deviation, and, therefore, to determine what values of working parameters of cavitation equipment are most typical for operable, limit and inoperable states.

When carrying out this study, it is possible to use different methods of clustering, such as Kohonen neural networks, hierarchical method, but the most convenient and visually representative is the method of k-means clustering, developed by the Polish scientist Hugo Steingauz.

K-means clustering method involves the initial determination of the number of clusters. Taking into account statistical data on each of the factors affecting cavitators, three clusters can be distinguished for leakage failures - cluster of data on parameters in the "permissible zone", cluster of data in the "limit state zone", cluster of data when the equipment is in the "failure zone" state. For the electrical part, only two clusters are distinguished - "permissible zone" and " failure zone". After allocation of the number of clusters within them, the cluster centers are randomly selected. Then, we calculate the arithmetic mean values of the points belonging to a certain cluster. These values become the new cluster centers. After a number of iterations, the clustering center point of each cluster stabilizes, taking the optimal value.

Since the beginning of the algorithm relies on a random selection of the initial clustering centers, to ensure the most qualitative partitioning of the data into clusters, the algorithm is repeated several times, and the result with the most qualitative solution to the clustering problem is chosen. According to Kleinberg's impossibility theorem, there is no optimal clustering algorithm, but the quality of the algorithm result can be evaluated by a number of criteria. To assess the quality of clustering of the data obtained we chose the criteria of compactness and separability of clusters, Dunn index and silhouette (Pandey and Singh 2016). Cluster Cohesion measures the degree of similarity between cluster elements. In this case, the degree of similarity is the sum of the squares of the Euclidean distance between points within one cluster and the center of this cluster. The smaller is this value, the better is the quality of clustering, since the same cluster contains conventionally similar points. Compactness is expressed by the following formula:

$$W(c_i) = \sum_{j=1}^{|c_i|} \left(|x_j - \overline{x_i}| \right)^2 \qquad (3)$$

In the above formula c_i is a cluster from the set of clusters C obtained by the algorithm, x_j is the cluster element c_i, x_i is the center of cluster c_i.

Cluster separation estimates the distance by which the resulting clusters are separated from each other. The greater the value of cluster separation, the better similar elements were grouped. In our data analysis, we will consider the separability of each cluster individually, and the average separability of clusters in general (Majhi and Biswal 2018). Individual separability of a cluster can be defined as the minimum distance from the center of this cluster, to the centers of the other clusters:

$$S(c_i) = min_{c_j \in C \setminus c_i}(d(\bar{x_i}, \bar{x_j})) \quad (4)$$

$\bar{x_i}$, $\bar{x_j}$ -centers of clusters c_i and c_j, respectively.

The average separability can be calculated as the average distance from the center of each cluster to the center of the entire set:

$$S(C) = \frac{1}{|C|} \sum_{c_i \in C} d(\bar{x}, \bar{x_i}) \quad (5)$$

$|C|$ - the number of clusters, x is the center (midpoint) of the set, (x_i) is the center of the cluster c_i.

The Dunn index allows to estimate the intra-cluster compactness and separability of clusters (the distance of clusters from each other). The higher the Dunn index, the higher the quality of clustering. The formula for calculating the Dunn index is as follows:

$$\mathrm{D(C)} = \frac{\min_{c_k \in C}\left(\min_{c_l \in C \setminus c_k}(d(c_k, c_l))\right)}{max_{c_k \in C}(max_{x_i, x_j \in c_k}(d(x_i, x_j))} \quad (6)$$

In this case, C is the set of clusters, d is the distance function. c_k and c_l are clusters from set C, x_i,x_j are elements of cluster c_k. The distance function d in the case of distance between clusters is defined as the minimum distance between elements of clusters:

$$d(c_k, c_l) = min_{x_i \in c_k, x_j \in c_l}(|x_i - x_j|) \quad (7)$$

In the case of distance between elements d(x_i,x_j) we calculate the Euclidean distance. Evaluation of the silhouette of the cluster also evaluates how close the point is to its cluster. The silhouette of a point is determined by the degree to which it is similar to other points in the same cluster, and the degree to which it differs from points in other clusters. For the point $x_i \in c_j$ the silhouette is expressed as:

$$s(x_i \in c_j) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))} \quad (8)$$

The functions a(x_i) and b(x_i) are the average distance from point x_i to other points in it's cluster and the minimum average distance from point x_i to points in other clusters, respectively.

$$a(x_i) = \frac{1}{(|c_j| - 1)} \sum_{x_k \in c_j, x_k \neq x_i} d(x_i, x_k) \quad (9)$$

$$b(x_i) = \min_{c_l \in C \setminus c_j}\left(\frac{1}{|c_l|} \sum_{x_k \in c_l} d(x_i, x_k)\right) \quad (10)$$

A The cluster modulus $|c_j|$ and $|c_l|$ in this case means the number of elements in a given cluster [18]. The estimate of the silhouette of the cluster can be defined as the average value of the silhouette of its elements.

$$s(c_j) = \frac{1}{|c_j|} \sum_{x_i \in c_j} s(x_i) \quad (11)$$

To calculate the presented results, we wrote a script in python using the numpy library for efficient handling of multidimensional data, and matplotlib for data visualization. The script consists of an implementation of the k-means algorithm, functions for calculating the previously described quality estimates, and for visualizing the results obtained. To create a shell for each cluster, we drew an ellipsoid with the center in the central point of the cluster, and axes equal to the cluster diameter (the difference between the maximum and minimum values) for the corresponding coordinate (Lund and Ma 2021). Implementing the clustering algorithm, and calculating the above-described characteristics on the resulting clusters, we obtain the following tables (Tables 2 and 3).

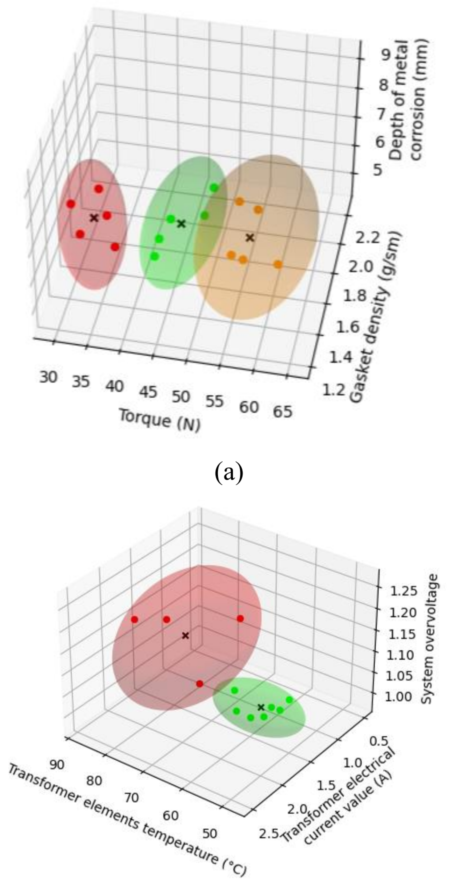
Table 2. The clustering indicators in the analysis of failures of the mechanical part of cavitators

Cluster/Characteristic	**Permissible zone (green)**	**Failure zone (red)**	**Limit state zone (orange)**
Center point	(55.6, 1.7, 5.6)	(34.6, 1.37, 7.34)	(44.6, 1.75, 5.5)
Compactness	46.35	20.14	39.4
Cluster separability	3.68	3.38	3.38
Silhouette	0.6	0.71	0.52
Dunn index	0.45		
Clustering separability	7.28		

Table 3. Clustering indicators in the analysis of failures of the electric part of cavitators

Cluster/Characteristic	Permissible zone (green)	Failure zone (red)
Center point	(58.29, 1.42, 1.01)	(74.75, 1.63, 1.14)
Compactness	93.5	103.33
Cluster separability	8.22	8.22
Silhouette	0.67	0.52
Dunn index	0.5	
Clustering separability	8.22	

Analyzing the results of clustering, we can note that the Dunn index has a rather low value, which is associated with a high value of compactness - that is, dispersion in each particular cluster, and low value of separability, both of each particular cluster and the set of clusters in general, but the silhouette of each cluster exceeds the average value of 0.5, which indicates sufficient similarity of data, grouped into clusters. We can conclude that in spite of the scatter of points in clusters, each cluster contains sufficiently similar points (Figure 7).



(a)

(b)

Figure 7. Results of cluster analysis on failures of mechanical (a) and electrical (b) cavitator systems. Clusters with equipment failures are highlighted in red, clusters with equipment in the limit state are highlighted in orange, clusters with equipment operation in the permissible zone are highlighted in green

As a result of the cluster analysis we have found that deviations in all considered parameters both in mechanical and electrical parts take part in transition of the cavitation equipment from the operable state to the limit and inoperable states. This is proved by the cluster separability.

The greater size of clusters in the inoperable and limit states, as compared to the cluster of the permissible zone, indicates that, stochastic processes play a greater role in these clusters (which is typical for equipment failures), and also indirectly indicates that other factors, not considered in this study, affect the failure rate of cavitation equipment (which makes clusters of limit state and "failure" state of the system less homogeneous).

In most clusters, the distance of points from the clustering centers is uniform, which shows the interrelation between all the considered parameters and the deviations between them (Lopes and Gosling 2020). The "failure" cluster in the electrical part has a slightly larger and fuzzy character, which is obviously associated with an insufficient number of measurements or with large values of deviations of the considered parameters from the limit values.

The obtained results of cluster analysis allow us to develop organizational and technological measures to ensure greater stability of cavitators functioning in the tank farms of thermal power plants. This can be achieved both through reorganization of the system of maintenance of these mechanisms, and through work with suppliers of equipment and consumables.

6. CONCLUSION

The study revealed that the largest number of leakages in the cavitators is due to the poor quality of the gaskets used and the uneven tightening of threaded connections. Failures of the electrical part are generally caused by exceeding the temperature of current-carrying elements of the systems and the growth of current strength, which leads to an emergency stop of the cavitator.

The regression and cluster analyses applied made it possible to determine the mutual influence of the considered factors on the transition of the system from the operable state to the limit state and then to the inoperable state. In the regression analysis, this transition can be associated with the point of change in the function graph (in the piecewise linear model), and in the cluster analysis with the empty zone between clusters, characterized by the separability of one from another. As a result, using both mathematical methods we have established that the greatest influence on the cavitators performance is caused by the deviation from the allowable values of parameters of the gasket torque and density in the mechanical part, and the current strength and temperature in the electrical part. These parameters are most closely related. The used method of least squares, as well as the k-means clustering method showed its effectiveness for processing a small statistical data set (that, unfortunately, is the specifics of using this equipment).

In general, for further research of cavitator failures when additional information is accumulated, it is rational to use not linear regression models, but piecewise linear models (Gruzdev 2019). In addition, it is necessary to consider the possibility of measuring the other parameters to study their mutual influence and increase the accuracy of clustering (Chu et. al. 2022). On the basis of the obtained mathematical results, a number of technical solutions for the identified problems can be formed.

So, the decrease in the number of cavitator failures can be achieved by replacing the types of used gaskets with gaskets of higher strength, as well as by using torque wrenches during the cavitator element assembly. In addition, a change in equipment maintenance and repair schedules is required. It is also necessary to consider the possibility of introducing systems of remote control of unit parameters. It is also reasonable to change the materials used in electrical systems or make changes to the electrical circuit in order to make it less sensitive to current surges or temperature rise (work with the manufacturer of the electrical part). For example, this can be achieved by implementing Hall-effect current sensors in the electrical circuit design.

Acknowledgment: Regression analysis of the data was performed as part of the state assignment of IO RAS (theme No. FMWE-2021-0012)

References:

Avvaru, B., Venkateswaran, N., Uppara, P., Iyengar, S. B., & Katti, S. S. (2018). Current knowledge and potential applications of cavitation technologies for the petroleum industry. *Ultrasonics sonochemistry, 42*, 493-507. https://doi.org/10.1016/j.ultsonch.2017.12.010

Bowman, K. O., & Shenton, L. R. (2004). Estimation: Method of moments. Encyclopedia of statistical sciences, 3.

Carpenter, J., Badve, M., Rajoriya, S., George, S., Saharan, V. K., & Pandit, A. B. (2017). Hydrodynamic cavitation: an emerging technology for the intensification of various chemical and physical processes in a chemical process industry. *Reviews in Chemical Engineering*, *33*(5), 433-468. https://doi.org/10.1515/revce-2016-0032

Carpenter, J., George, S., & Saharan, V. K. (2017). Low pressure hydrodynamic cavitating device for producing highly stable oil in water emulsion: Effect of geometry and cavitation number. *Chemical Engineering and Processing: Process Intensification, 116*, 97-104. https://doi.org/10.1016/j.cep.2017.02.013

Chu, F., Dai, B., Lu, N., Wang, F., & Ma, X. (2022). A Multiprocess Joint Modeling Method for Performance Prediction of Nonlinear Industrial Processes Based on Multitask Least Squares Support Vector Machine. *Industrial & Engineering Chemistry Research*, *61*(3), 1443-1452. https://doi.org/10.1021/acs.iecr.1c04075

Denham, M. C. (1995). Implementing partial least squares. *Statistics and Computing*, *5*(3), 191-202. https://doi.org/10.1007/BF00142661

Draper, N. R., & Smith, H. (1998). Applied regression analysis (Vol. 326). John Wiley & Sons.

Frumin, L. L. (2020). Linear least squares method in nonlinear parametric inverse problems. *Journal of Inverse and Ill-posed Problems*, *28*(2), 307-312. https://doi.org/10.1515/jiip-2019-0009

Gruzdev, A. N. (2019, February). Accounting for long-term serial correlation in a linear regression problem. In IOP *Conference Series: Earth and Environmental Science* (Vol. 231, No. 1, p. 012020). IOP Publishing. https://doi.org/10.1088/1755-1315/231/1/012020

Hagiwara, S., Uezono, T., Sato, T., & Masu, K. (2008). Application of correlation-based regression analysis for improvement of power distribution network. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *91*(4), 951-956. https://doi.org/10.1093/ietfec/e91-a.4.951

Kafle, S. C. (2019). Correlation and regression analysis using SPSS. *OCEM J Manag Tech Soc Sci*, 1(1), 126-132.

Lopes, H. E. G., & Gosling, M. D. S. (2020). Cluster analysis in practice: Dealing with outliers in managerial research. *Revista de Administração Contemporânea*, 25. https://doi.org/10.1590/1982-7849rac2021200081

Lund, B., & Ma, J. (2021). A review of cluster analysis techniques and their uses in library and information science research: k-means and k-medoids clustering. *Performance Measurement and Metrics*. https://doi.org/10.1108/PMM-05-2021-0026

Majhi, S. K., & Biswal, S. (2018). Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer. Karbala *International Journal of Modern Science*, *4*(4), 347-360. https://doi.org/10.1016/j.kijoms.2018.09.001

Markelov, G. E. (2015). Constructing a Working Mathematical Model. *Theoretical & Applied Science*, (8), 44-46. http://dx.doi.org/10.15863/TAS.2015.08.28.6

Mohan, R., & Sivakumar, V. (2022). Analysis and correlation of ultrasound cavitation energy in ultrasound tank with coloration of fibrous materials: leather dyeing. *Brazilian Journal of Chemical Engineering*, *1-23*. https://doi.org/10.1007/s43153-022-00241-7

Rossi, R. J. (2018). Mathematical statistics: an introduction to likelihood based inference. John Wiley & Sons.

Pandey, P., & Singh, I. (2016). Comparison between Standard K-Mean Clustering and Improved K-Mean Clustering. *International Journal of Computer Applications*, *146*(13). https://doi.org/10.5120/ijca2016910868

Protalinskii, O. M., Shcherbatov, I. A., & Esaulenko, V. N. (2013). Analysis and modelling of complex engineering systems based on the component approach. *World Applied Sciences Journal*, *24*(2), 276-283. https://doi.org/10.5829/idosi.wasj.2013.24.itmies.80033

Putra, A. B. W., Gaffar, A. F. O., & Suprapty, B. (2021). A Performance of the Scattered Averaging Technique based on the Dataset for the Cluster Center Initialization. *International Journal of Modern Education & Computer Science, 13*(2). https://doi.org/10.5815/ijmecs.2021.02.05

Seraya, O. V., & Demin, D. A. (2012). Linear regression analysis of a small sample of fuzzy input data. *Journal of Automation and Information Sciences*, *44*(7). https://doi.org/10.1615/JAutomatInfScien.v44.i7.40

Sharma, D., & Chandra, P. (2020). Linear regression with factor analysis in fault prediction of software. *Journal of Interdisciplinary Mathematics*, *23*(1), 11-19. https://doi.org/10.1080/09720502.2020.1721641

Wang, Z., Gao, D., Diao, B., Tan, L., Zhang, W., & Liu, K. (2019). Comparative performance of electric heater vs. RF heating for heavy oil recovery. *Applied Thermal Engineering*, *160*, 114105. https://doi.org/10.1016/j.applthermaleng.2019.114105

Shcherban Pavel Sergeyevich
Immanuel Kant Baltic Federal University, Branch Scientific Cluster Institute of High Technologies
Kaliningrad,
Russia
ursa-maior@yandex.ru
ORCID 0000-0001-5106-7852

Sokolov Andrey Nikolaevich
Immanuel Kant Baltic Federal University, Branch Scientific Cluster Institute of High Technologies
Kaliningrad,
Russia

Shirshov Institute of Oceanology, Russian Academy of Sciences.
Moscow,
Russia
tengritag@gmail.com
ORCID 0000-0002-7593-9739

Abu Hamdi Reda Validovich
Immanuel Kant Baltic Federal University, Branch Scientific Cluster Institute of High Technologies
Kaliningrad,
Russia
rabouhamdi@gmail.com
ORCID 0000-0002-2275-0529