Development of a cluster analysis method for solving the problem of identifying static voltage load characteristics

Mikhail A Kondrashov1,2, Maksim G Popov1, Dmitry S Bukharov2 and Denis E Petrushin1,2,3

1 Peter the Great St. Petersburg Polytechnic University, Saint Petersburg, Russia
2 System Operator of the United Power System, JSC, Moscow, Russia
3 Petrozavodsk State University, Petrozavodsk, Russia

Abstract. Identification of static voltage load characteristics is not a trivial task. It is extremely difficult to develop a single methodology that will allow achieving the required accuracy results for any input data. The authors made an attempt to solve the problem, developed a method based on the use of methods for processing large data sets – BigData. One of the most important stages of this method is the stage of cluster analysis. The problem of choosing a method of cluster analysis also has nuances: it is necessary to maintain a balance between time and accuracy of calculations, while performing calculations with sufficient accuracy on various data. A number of domestic and foreign authors have investigated this issue, but the proposed options cannot be used in the developed methodology. The purpose of this research is to develop a cluster analysis method that will satisfy the above requirements. The author's method is based on the k-means method. The main feature of the author's method is the mechanism of initial generation of cluster centroids, which allows to exclude the influence of random generation of two-dimensional coordinates imperfection and to achieve uniform generation of initial clusters centers. The developed method was tested on a model problem with all known solutions, which can be easily verified analytically. Also the method was tested on actual industrial load data. The results obtained the requirements of accuracy and calculation speed and allow us to conclude that the developed method of cluster analysis can be used in the task of identifying static voltage load characteristics.

1. Introduction
One of the priority tasks of improving tools and activities for calculating, analysing and planning current and future electric power regime is to update the actual static voltage load characteristics (SVLC) [1]. SVLC is a model for representing the electric load of the consumer, showing the change in the amount of consumed electrical power depending on the change in the voltage of the power supply network. Due to the lack of actual data, as a rule, generalized SVLC characteristics are used. It obtained by calculation on the basis of reference data on the complex load of a certain voltage class. The use of generalized SVLC gives an erroneous representation of the behaviour of the electric load and leads to significant errors in the calculations of steady-state regime.

The main method of SVLC identification is the experimental method. The experimental method is the research of the load node using the accumulation of information about the operating regime or conducting an experiment with a change in the supply voltage. This method involves processing large
data sets using software for identification SLVC that will allow you to calculate the SVLC most effectively.

The detailed information about algorithm, architecture, and functionality of the software for identification SLVC was presented in [2], [3]. In a simplified form, the process of identifying static load characteristics for an abstract consumer of electric energy can be represented in the form of a diagram in figure 1.

**Figure 1.** Diagram of the process for identifying static voltage load characteristics.

Initial data for process is received from telemetry, SCADA, or vector measurement systems. The choice of initial data representation is based on the source, but in any case data can be represented as an array of active, reactive power and voltage values with time stamps.

From the initial data, it is necessary to select stable condition in which the change in the consumption of active and reactive power is caused by a change in the supply network voltage. The authors cluster analysis methods is proposed to solve this problem. Cluster analysis is the division of a set of objects into groups using selected metrics. Also, cluster analysis is supposed to be used for splitting points on the $P(U)$ and $Q(U)$ planes into separate "clouds" and using the results of clustering in regression analysis.

The task of selecting a suitable cluster analysis method is not straightforward. In papers [4-6], the authors tried to solve this problem, but the options they proposed did not show the required accuracy when used in specialized software. Therefore, the purpose of this research is to develop the author's method for cluster analysis.

The developed author's method for cluster analysis is a modification of the $k$-means method. More information about the developed method is provided below.

2. **The author's method**

A key feature of the author's method is the generation of initial positions $n$ of cluster centers $W_j$ ($j = 1, n$). The initial position $W_j$ is set so that the minimum distance $L(W_i,W_j)$ between the centers $W_i$ and $W_j$ ($j = 1, n$) does not exceed the value. The need to use this generation is explained by the imperfection of pseudorandom number generator that generate random values from a given range in a one-dimensional space and necessity to generate the initial positions of centers $W_j$ ($j = 1, n$) as, a point of two-dimensional space.

The author's method can be divided into four logically sequential modules, each of them implements a certain functionality (see figure 2).

2.1. Module 1 – Generating initial locations of clusters centers
All initial positions of cluster centers $W_j$ ($j = 1, n$) are generated randomly. For each currently generated $W_g$ ($g = p + 1, n$), the condition is checked that it is outside the restricted area with radius for previously generated points $W_h$ ($h = 1, p$). The coordinates of the center $W_g$ are re-generated, if the distance is $L(W_g, W_h) \leq d$.

**Figure 2.** Algorithm and principle of operation of the cluster method.

The radius $d$ is determined empirically. The results of computational experiments have shown that the best result of generating the initial positions of the centers $W_i$ ($i = 1, n$) can be achieved with the value $d = 0.1$. The dimension of the value is a percentage of relative units of the minimal dispersion of values of the original data on either coordinate axes.

### 2.2. Module 2 – Building a distance matrix

This module calculates the distance between each point $S_i$ ($i = 1, m$) and the cluster center $W_j$ ($j = 1, n$) within the selected metric and builds a distance matrix. Distance metrics are defined as function that allows you to determine the distance between any points in the metric space $R^p$.

The choice of a specific distance metric is based on the application problem being solved and the source data. When solving model problems, the Euclidean distance metric (1) is used, since it allows you to repeat the calculations analytically without much difficulty.

$$
E(x, y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_p - y_p)^2} = \sqrt{(x - y)^T(x - y)},
$$

(1) where $E(x, y)$ – Euclidean distance, $x = (x_1, \ldots, x_p)^T$ and $y = (y_1, \ldots, y_p)^T$ – points in the metric space $R^p$.

The calculated distance matrix is used in module 3.

### 2.3. Module 3 – Clustering of the original data
This module selects the smallest distance $r_{ij}$ to the center $W_j$ ($j = 1, n$) for each point $S_i$ ($i = 1, m$) and assigns the cluster number $j$ to the point $S_i$, thus forming a cluster $K_j = \{ S_i | r_{ij} \rightarrow \text{min} \}$.

If a situation occurs where the smallest distance $r_{ij}$ for a point $S_i$ ($i = 1, m$) is equal to two or more centers $W_j$ ($j = 1, n$), then the point $S_i$ is assigned a cluster number $K_j$ that with this point $S_i$ will have the minimum average clusterwide distance $F_j$:

$$F_j = \frac{\sum_{i=1}^{m_j} r_{ij}}{m_j} \rightarrow \text{min}$$

where $m_j$ – number of points in the cluster $K_j$.

2.4. Module 4 – Overriding the coordinates of cluster centers

To solve the problem of determining the optimal location of the centroids of each cluster, the mass center method is used [7, 8]. This method is based on determining the coordinates of the location centers of each cluster $K_j$:

$$W'_x = \frac{\sum_{i=1}^{m_j} x_i}{m_j} \cdot W'_y = \frac{\sum_{i=1}^{m_j} y_i}{m_j}$$

where $W'_x$, $W'_y$ – the overridden coordinates of the cluster center, $x_i$, $y_i$ – coordinates of points included in the $j$ cluster.

The final step is to check for the stability center $W'_j$ ($j = 1, n$) location within the selected metric: if the center $W'_j$ is located inside the $\varepsilon$-neighborhood of the center $W_j$, the final cluster $K_j$ is found, otherwise calculations for modules 2-3 are performed again.

The presented author’s method is used for solving the clustering problem many times-$q$ times. Each iteration of applying the author’s method on the source data ends with the calculation of some stable solution. From all $q$ stable solutions, the best one is chosen, in which the total average clusterwide distance is minimal:

$$\sum_{j=1}^{n} F_j \rightarrow \text{min}$$

Two computational experiments were conducted: the first computational experiment is performed on a model problem for which all possible solutions are known and easily verified analytically. The experiment shows the need to enter the initial positions of the cluster centers of the restricted area with a radius into the generation and the correctness of finding the optimal solution.

The second computational experiment is based on the actual industrial load data.

3. Computational experiments

3.1. Model task

Given data: the quantity of clusters $n = 2$, the quantity of clusterized points $m = 4$, the quantity of iterations $q = 1000$. Clusterized points are located to coordinates $(1;1)$, $(1;20)$, $(20;1)$, $(20;20)$, which
form the square with sides by 20 units, that makes it possible to visualize results of algorithm action and to check the correctness of calculations.

Figure 3 shows the uniformity of generating initial locations of clusters centers is determined of variable values $d$.

![Figure 3. The distribution of initial centroid locations of clusters by various values $d$:
(a) $d = 0$,
(b) $d = 0.08$,
(c) $d = 0.1$,
(d) $d = 0.12$,
(e) $d = 0.2$.](image)

Ideally when generation of initial centroid locations is uniform, every nods on the plane should be five hits. Due to generation process the distribution of initial points is nonuniform (see figure 3), in other words if the variable $d = 0$ (absence of restricted area), then the spread in values can be seen in the range [1:10].

For the purpose of obtaining solution of explored model task variable $d > 0$ should be used, representing radius of restricted area for initial centroid location of clusters. Increasing value to 0.1 units rises distribution of initial points locations on the plane uniformity. Further increase value to 0.2 units will decrease distribution uniformity.
For estimating the uniformity of initial points location the uniformity coefficient will be used:

\[ K_{\text{uniformity}} = \frac{k_4 + k_5 + k_6}{k_{0-11}} \]  

where \( k_{0-11} = 400 \) – summarized quantity of nods with integer coordinates, \( k_4, k_5, k_6 \) – the quantity of nods, which contains the number of initial locations \( W_j \ (j = 1, n) \), for which values are alike 4, 5 and 6 respectively.

According to table 1 the best value \( K_{\text{uniformity}} = 0.74 \) by \( d = 0.1 \), that will be seen in figure 3. In this case we can see increasing computation time, caused by extra re-calculations of initial location of centroids \( W_j \ (j = 1, n) \), which occasionally falls into restricted area. As we increase value of the variable , the more re-calculations are carried out.

**Table 1. The dependence on the variable .**

| \( d \), units | 0   | 0,08 | 0,1  | 0,12 | 0,2  |
|----------------|-----|------|------|------|------|
| \( K_{\text{uniformity}} \) | 0.66| 0.705| 0.74 | 0.635| 0.61 |
| Computation time, ms | 1.327| 7.716| 8.493| 9.359| 11.265 |

So, the most preferable value of restricted area radius of restricted area (for explored model task) for initial location of centroids is 0.1. For further calculations we will use the radius \( d = 0.1 \).

Through the execution of \( q \) iterations of computations we obtain six stable solutions (see figure 4), which are checked analytically. Figure 4 shows that author’s method find these solutions nonuniformly: with the most probability (75%) – solutions №3 and №6. The value of integrated average of clusterwide distance (2) is the most (38 units).

![Figure 4. Coordinates and frequency of cluster analysis solutions finding.](image)

The best solution (2) is the optimum for explored model task and is 24,854 units. Integrated probability finding of this solution, with used of author’s method, is 25% (see figure 4, solutions №1-2, №4-5) after doing \( q \) iterations of computations.

A low probability (6–7%) of optimum finding is explained by most attraction domain of local minimum for solutions №3 and №6.

Figure 5 (a) shows the diagram of attraction domain of local minimums for solutions №3 and №6. On the inside of the cluster for these solutions \( F_1 = F_2 = \text{const} \) without distinction of centroid location on the red line. With the scale of using of the mass center method for computing of cluster center point, stable solution (as defined above in module 4) is on the inside of \( \varepsilon \)-neighborhood for points (1;10) and (20;10) (gray ball) of solution №3 and for points (10;1) and (10;20) (green ball) of solution №6.

Figure 5 (b) shows the diagram of attraction domain of local minimums for solution №2. Since attraction domain for optimums №1–2 and №4–5 is less than for solutions №3 and №6, finding solution for these optimums is very complicated.

---

\( K_{\text{uniformity}} = \frac{k_4 + k_5 + k_6}{k_{0-11}} \)
Figure 5 (c) shows summary attraction domain diagram for all local minimums, which is truncated for illustrative purposes.

**Figure 5.** Attraction domain of local minimums for:
(a) – solutions №3 and №6;
(b) – solutions №2;
(c) – summary;

For computation of the optimum (one of №1–2, №4–5) it is necessary to locate centroids as close as possible to vortex center and to lying opposite vertex (see figure 5 (b)). In other cases centroid relocate to one of attraction domains for solutions №3 and №6 (see figure 6), after that author’s method restructure clusters, turned thoroughly into one of stable conditions: (1;10)–(20;10) or (10;1)–(10;20).

**Figure 6.** The relocation of centroid at another local minimum.

In conclusion, developed modification of the k-means method adequately solve the model task. After that author’s method was tested on actual power data, obtained from an industrial facility.

3.2. Actual power data
Telemetry data of actual industrial load has been obtained from a feeder supplying aluminum smelter. Figure 7 shows the dependence of reactive power on time and the dependence of reactive power on voltage.

Figure 7. Reactive power of actual industrial load.

Figure 7 shows that the load has different states. These transformations are caused by different processes in aluminum producing. Above said the SVLC shows the dependence of load power on supplied voltage. For this purpose it is necessary to select different periods of aluminum smelter activity for further SVLC identification.

The result of clusterization by author’s modification of the k-means method is showed on figure 8.

Figure 8. Result of clusterization by author’s modification of the k-means method.

The usage of author’s modification of the k-means method has allowed to compute quasi-stable in time load states for actual data of industrial load with required accuracy of calculations. As a result we can make the conclusion that the developed method of cluster analysis is usable for solving task of the SVLC identification in the productivity software.

4. Conclusion

The developed method of cluster analysis based on the k-means method. Main feature of this method is the algorithm of initial location generation of cluster centers. This algorithm is used for neutralization imperfection of pseudorandom number generator of two-dimensional coordinates and for uniform generation of initial location of cluster centers.

The developed method is tested on the model task for which is known all possible solutions that easy to check analytically. Knowledge of the all solutions has made possible estimate both of the accuracy and closeness of executed calculations and of correctness of the method. After that the one is tested on actual data, obtained from an industrial facility. This method has made possible achieve the goal with required accuracy of calculations.
Above said allows to conclude that the developed method of cluster analysis is adequate. The method is suited to the requirements of cluster analysis methods for identification of static voltage load characteristics. This method will be used in software for identification SLVC.

References
[1] Petrushin D and Vessart V 2018 Analysis of Voltage Static Load Characteristics Electricity Through the Eyes of Youth – 2018. Materials of the IX International Youth Scientific and Technical Conference - Proc. (Kazan, Russia: KSEU) 1 pp 176–9 (In Russian).
[2] Kondrashov M A and Kondrashova A Yu 2018 Software to identify the actual voltage steady-state load characteristics identification of large consumers Prikl Inform — J Appl Informatics 13(5(77)) pp 44–50 (In Russian).
[3] Kondrashov M A, Pankratov A V and Batseva N L 2019 2019 International Siberian Conference on Control and Communications (SIBCON) - Proc. (Tomsk, Russia: IEEE Inc) doi: 10.1109/SIBCON.2019.8729660
[4] Zhang X, Grijalva S and Reno M J 2014 A time-variant load model based on smart meter data mining IEEE Power and Energy Society General Meeting(National Harbor, USA: IEEE Inc) doi: 10.1109/PESGM.2014.6939365
[5] Tavlintsev A S et al. 2018 Search for the single-type load schedules of the power facility 2018 IEEE 59th International Scientific Conference on Power and Electrical Engineering of Riga Technical University - Proc. (Riga, Latvia: IEEE Inc) doi: 10.1109/RTUCON.2018.8659825
[6] Pankratov A V et al. 2019 Application of expectation maximization algorithm for measurement-based power system load modeling 2019 International Siberian Conference on Control and Communications (SIBCON) - Proc. (Tomsk, Russia: IEEE Inc) doi: 10.1109/SIBCON.2019.8729610
[7] Celebi M E, Kingravi H A and Vela P A 2013 A comparative study of efficient initialization methods for the k-means clustering algorithm Expert Syst. Appl. 40(1) pp 200-10 doi: 10.1016/j.eswa.2012.07.021
[8] Bukharov D 2017 On computer-aided adjustment of automatics for elimination of asynchronous operation with trapezoidal characteristic Prikl Inform — J Appl Informatics 12(2(68)) pp. 44–51 (in Russian).

Acknowledgments
This work was partially funded by the Russian Foundation for Basic Research (RFBR), project number 19-38-90123.