Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

Clustering the Impacts of The Russia-Ukraine War on Personnel and Equipment

Wargijono Utomo

Sistem Informasi
Universitas Krisnadwipayana
Jakarta, Indonesia
https://unkris.ac.id/
wargiono@unkris.ac.id

Abstract

In post-pandemic recovery efforts, uncertainty arose due to the unresolved conflict between the Russia-Ukraine war. This conflict impacts world security stability and affects the economic, energy, and food sectors. This conflict also impacts humanity by causing death to civilians and military personnel, including children in Ukraine. The clustering analysis results of the impact of the Russian-Ukrainian war show losses and losses in personnel and war equipment, with three cluster optimization methods used through k-means. Of the two methods that can be recommended, namely elbow and Silhouette, both produce K=3. The profiling results show that losses or losses in Ukrainian personnel and war equipment are categorized into three clusters, with cluster one being the lowest category, cluster two being the very high category, and cluster three being the moderate category. This research is helpful for state agencies, international organizations (NGOs), and other stakeholders.

Keywords: Clustering; K-Means; Elbow; Silhouette; Gap Statistics

Abstrak

Dalam upaya pemulihan pascapandemi, ketidakpastian muncul akibat konflik yang belum terselesaikan antara perang Rusia-Ukraina. Konflik ini berdampak pada stabilitas keamanan dunia, dan juga mempengaruhi sektor ekonomi, energi dan pangan. Konflik ini juga berdampak pada kemanusiaan dengan menyebabkan kematian warga sipil dan personel militer, termasuk anak-anak di Ukraina. Hasil analisis clustering dampak perang Rusia-Ukraina menunjukkan kerugian dan kerugian personel dan peralatan perang, dengan tiga metode optimasi cluster yang digunakan melalui k-means. Dari dua metode yang dapat direkomendasikan yaitu elbow dan Silhouette, keduanya menghasilkan K=3. Hasil profiling menunjukkan bahwa kerugian atau kehilangan personel dan peralatan perang Ukraina dikategorikan menjadi tiga klaster, dengan klaster satu kategori paling rendah, klaster dua kategori sangat tinggi, dan klaster tiga kategori sedang. Penelitian ini bermanfaat bagi lembaga negara, organisasi internasional (LSM), dan pemangku kepentingan lainnya.

Kata kunci: Clustering; K-Means; Elbow; Silhouette; Gap Statistics

INTRODUCTION

In post-pandemic joint recovery efforts, the world is uncertain due to the implications of the conflict between Russia and Ukraine, which has not been resolved to date. The conflict that is still happening has an impact on the world, security stability, and its impact on the economy, energy, and food which one day will have an indirect impact on defense and security(Darmayadi & Megits, 2023; Nerlinger & Utz, 2022; Paul, 2015). In addition, the outbreak of the Russian-Ukrainian military conflict had implications for humanity, resulting in the deaths of civilians, military personnel, and even children in Ukraine(Haque et al., 2022; Osokina et al., 2022). Statista.com, quoted from the Office of the UN High Commissioner for Human Rights (OHCHR), verified 6,952 civilian deaths during the Russian invasion of Ukraine as of January 9, 2023. Of these, 431 were children. Subsequently, 11,144 people were reported injured.

Various studies that have been conducted on the impact of the Russian and Ukrainian wars include The EU in the South Caucasus and the Impact of the Russia-UkraineWar with a qualitative approach(Paul, 2015), The Impact of the Russia-Ukraine War on the Cryptocurrency Market with the IV-GMM method, The impact of the Russia-Ukraine conflict on energy firms: A capital market perspective using the average abnormal returns (AAR) method(Nerlinger & Utz, 2022), The human toll and humanitarian crisis of the Russia-Ukraine war: the first 162 days using the descriptive statistics method(Haque et al., 2022).

Based on various previous studies, this paper aims to cluster the impact of the Russian war on Ukraine on the human side, such as the number of deaths and losses in armament using the k-means algorithm and cluster optimization using three methods, including elbow, Silhouette, and Gap Statistics while clustering data processing using R programming(Sinaga & Yang, 2020). Clustering is an unsupervised learning method that allows the grouping of objects based on different characteristics(Yuan & Yang, 2019). The purpose of cluster analysis is to discover the structure in forming groups of similar cluster objects. Clustering is needed to identify the structure of the data.

This research can contribute to state agencies, international organizations (NGOs), and stakeholders. This paper is organized into four parts to be systematic, including the first part of the research background, the second part of the research methodology, the third part of the results and discussion, and the fourth part of the conclusions.

RESEARCH METHODS

Studies on the impact of the Russian-Ukrainian war on personnel and war equipment are carried out using structured, planned, and systematic quantitative methods to make it better. The research process involves several stages, as shown in Figure 1.

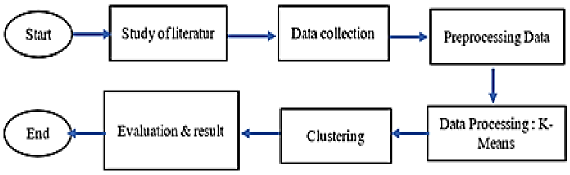


Figure 1. Research Methodology

Clustering

Clustering is a method in data mining that aims to group (or classify) items in data into several groups (or clusters) based on similarities in their features. The goal of clustering is to discover hidden structures in data and understand how items are related to one another. Clustering is an unsupervised machine-learning technique that involves grouping similar data points based on similarity or distance metrics. The goal of clustering is to identify natural groupings within a dataset that can be used for further analysis or to gain insight into the underlying structure of the data. Clustering algorithms typically require no prior knowledge of the data or its structure and instead attempt to partition the data into distinct clusters based on their similarity or dissimilarity. There are many different clustering algorithms, including k-means, hierarchical, and density-based clustering, each with strengths and weaknesses. Clustering is widely used in many applications, such as image and text processing, marketing segmentation, customer profiling, bioinformatics, and anomaly detection, to name a few

The K-Means Algorithm

K-Means is one of the most popular clustering algorithms. This algorithm divides data into K adjacent groups based on the distance between data points. This process is done by determining K central points or centroids representing each group and then placing each data point into the closest group based on the shortest distance to the nearest centroid. Given a set of objects, the primary aim of the k-means clustering is to optimize the following objective function(Cohn & Holm, 2021; Govender & Sivakumar, 2020):

$$J = \sum_{j=1}^{k} \sum_{i \,\epsilon\, c\, j} \left\| x_i - c_j \right\|^2 \text{..} \quad (1)$$

The formula involves a criterion function (represented by "j") and various variables, including the i-th observation (represented by "xi"), the j-th cluster center (represented by "cj"), the set of objects in the j-th cluster (represented by "cj"), and the number of clusters (represented by "k"). The distance between the data object and the cluster's center is represented by a norm denoted by "‖∗‖." The goal of the criterion function is to minimize the distance between each data point and the cluster center it is located in. The k-means iterative clustering method is commonly executed in the following manner:

1. Please select a value for k and use it to establish the initial set of k centroids.
2. Group each object with the centroid nearest to it.
3. Calculate the mean of the cluster members to determine the new centroids for each k cluster.
4. Iterate steps 2 and 3 until there is no modification in the criterion function after an iteration.

The Elbow, Silhouette, Gap Statistic

Elbow Method: The Elbow Method is a clustering evaluation method that plots the number

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v5i2.215

Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

of clusters against the Within-Cluster-Sum-of-Squared-Errors (WCSS) values (Sinaga & Yang, 2020). The basic idea of this method is to choose the number of clusters that give the best results in terms of WCSS. In the WCSS plot versus the number of clusters, the "elbow" point on the graph indicates the optimal number of clusters to choose from.

Silhouette Score: Silhouette Score is a clustering evaluation metric that considers the distance between data points and centroids of other clusters. Silhouette scores range between -1 and 1, with a value of 1 indicating that the data point is strongly related to the current cluster and not related to other clusters, while a value of -1 indicates that the data point is better moved to another cluster.

Gap Statistics: Gap Statistics is a clustering evaluation method comparing actual data distribution with the same random distribution. This method measures how well the clustering algorithm separates data into clusters and evaluates the optimal number of clusters. A higher Gap Statistic value indicates that the clustering algorithm better separates data into clusters.

R-Programming Language

R is a programming language for data analysis, visualization, and statistical modeling [16]. Developed in 1993, R has a very active community and has thousands of packages that can assist in performing data analysis tasks, from data cleaning to statistical modeling (Peng, 2015). The advantages of the R programming language are Open-Source, Compatibility with various systems, and Thousands of packages available(Mailund, 2017). Powerful visualization capabilities: R has many packages that allow easy and efficient data visualization. While it has many advantages, R also has some disadvantages, such as R has many features and packages that make it very powerful, but it also makes the learning curve quite steep. Slower performance compared to other programming languages: because it was written in an interpretive language, R's performance is sometimes slower than other compilation-oriented programming languages. In general, R is a compelling and flexible programming language used by data analysts and statisticians to tackle data analysis and statistical modeling tasks

RESULTS AND DISCUSSION

This study used the dataset from kaggle.com, which was accessed on January 8, 2023, and consisted of two excel format files, namely losses equipment, and personnel, each of which was 319 data consisting of 18 for equipment variables and five variables for personnel, which can be seen from table 1. From the dataset tables 1 and 2, cleaning and unifying the data consists of one personnel attribute and eight equipment attributes, then transformed into one dataset consisting of 9 attributes and 319 data, as shown in Figure 2.

Table 1. Personnel Dataset

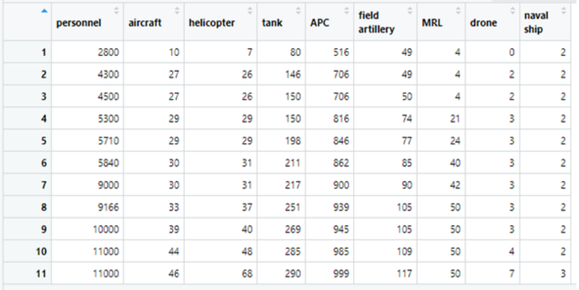
Date	Day	Personnel	POW
2/25/2022	2	2800	0
2/26/2022	3	4300	0
⋮	⋮	⋮	⋮
1/7/2023	318	110740	0
1/8/2023	319	111170	0

Table 2. Dataset of equipment losses

date	day	aircraft	helicopter	...
2/25/2022	2	10	7	...
2/26/2022	3	27	26	...
⋮	⋮	⋮	⋮	⋮
1/7/2023	318	285	272	...
1/8/2023	319	285	272	...

Before processing K-Means clustering data using R-Studio, install the required library packages such as tidyverse, cluster, factoextra, and dplyr to be used (Dmitry & Yerkebulan, 2022; Zhu, Idemudia, & Feng, 2019). In the first stage, create scripts or code to import excel datasets from the R Studio application, which can be used. As seen in code 1, then the results of the import dataset can be seen in Figure 2.

```
Code_1_import dataset excel
library(readxl)
data_ukraina <- read_excel("D:/Dokumen /data
ukraina.xlsx")
View(data_ukraina)
```



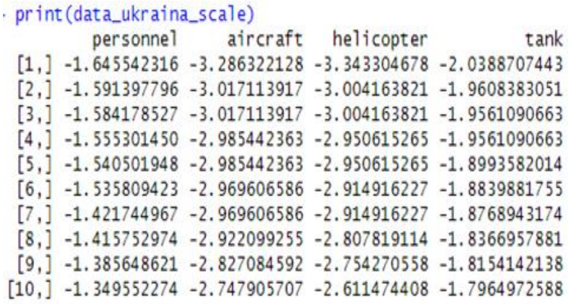
	personnel	aircraft	helicopter	tank	APC	field artillery	MRL	drone	naval ship
1	2800	10	7	80	516	49	4	0	2
2	4300	27	26	146	706	49	4	2	2
3	4500	27	26	150	706	50	4	2	2
4	5300	29	29	150	816	74	21	3	2
5	5710	29	29	198	846	77	24	3	2
6	5840	30	31	211	862	85	40	3	2
7	9000	30	31	217	900	90	42	3	2
8	9166	33	37	251	939	105	50	3	2
9	10000	39	40	269	945	105	50	3	2
10	11000	44	48	285	985	109	50	4	2
11	11000	46	68	290	999	117	50	7	3

Figure 2. results of import dataset excel

After importing the dataset is complete, then coding or scripting for standardizing or normalizing data processing, determining the number of clusters, implementing K-Means clustering, determining classes, data visualization, and group profiling. After that, in the second stage, coding is done to standardize data or create scale, as seen in code 2.

```
Code_2_the scala distribution and cluster determination
# normalization of data with scala
>data_ukraina_scale <- scale(data_ukraina)
#data standard
>print(data_ukraina_scale)
# determining the number of clusters
>fviz_nbclust(data_ukraina_scale, kmeans, method = "wss")
# metode elbow
>fviz_nbclust(data_ukraina_scale, kmeans, method =
"silhouette") # metode silhouette
>fviz_nbclust(data_ukraina_scale, kmeans, method =
"gap_stat") # metode gap_stat
```

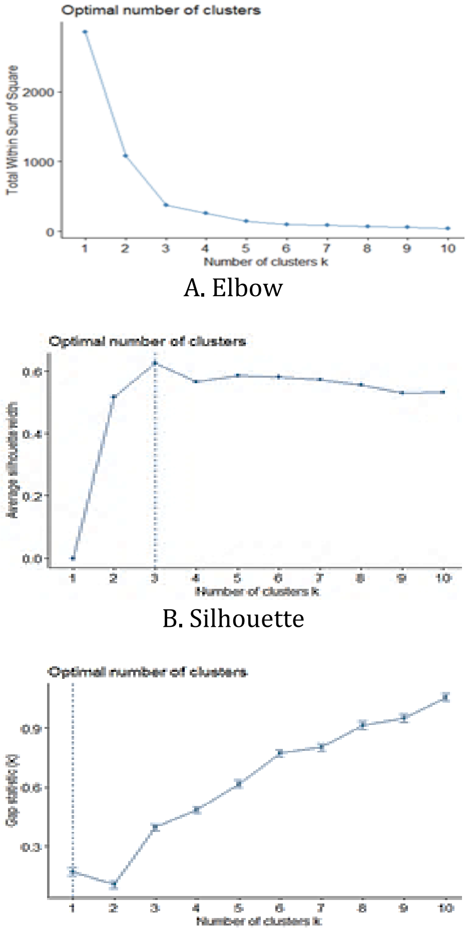
Data normalization in R Studio converts data from different scales to a uniform or the same scale(Kaparang, Moningkey, & Sumual, 2021; Shelly et al., 2020). Normalization is done to correct differences in scale between variables that can affect the statistical analysis and predictive models that are performed. Scale: This function is used to standardize data by converting data values into z-scores, which are an average value of zero and a standard deviation of one, and the result is that not all of the data can be displayed because there are many, as shown in Figure 3.



```
print(data_ukraina_scale)
         personnel     aircraft   helicopter         tank
 [1,] -1.645542316 -3.286322128 -3.343304678 -2.0388707443
 [2,] -1.591397796 -3.017113917 -3.004163821 -1.9608383051
 [3,] -1.584178527 -3.017113917 -3.004163821 -1.9561090663
 [4,] -1.555301450 -2.985442363 -2.950615265 -1.9561090663
 [5,] -1.540501948 -2.985442363 -2.950615265 -1.8993582014
 [6,] -1.535809423 -2.969606586 -2.914916227 -1.8839881755
 [7,] -1.421744967 -2.969606586 -2.914916227 -1.8768943174
 [8,] -1.415752974 -2.922099255 -2.807819114 -1.8366957881
 [9,] -1.385648621 -2.827084592 -2.754270558 -1.8154142138
[10,] -1.349552274 -2.747905707 -2.611474408 -1.7964972588
```

Figure 3. Data normalization with scale

Next, in the third stage, determine the number of clusters using the elbow, Silhouette, and statistical gaps to determine the most optimal number of clusters. Then coding can be seen in code 2. significantly towards four and then bends or forms an elbow, so it can be concluded that the optimal number of clusters k = 3 can be seen in Figure 4a. Furthermore, the Silhouette method with an average value approach is used to estimate the quality of the clusters formed, and the higher the average value, the better the quality. From the results of this analysis, several clusters are considered optimal, namely k = 3 and k = 5, which can be seen in Figure 4b, because they have the highest average Silhouette value compared to the number of other clusters



A. Elbow

B. Silhouette

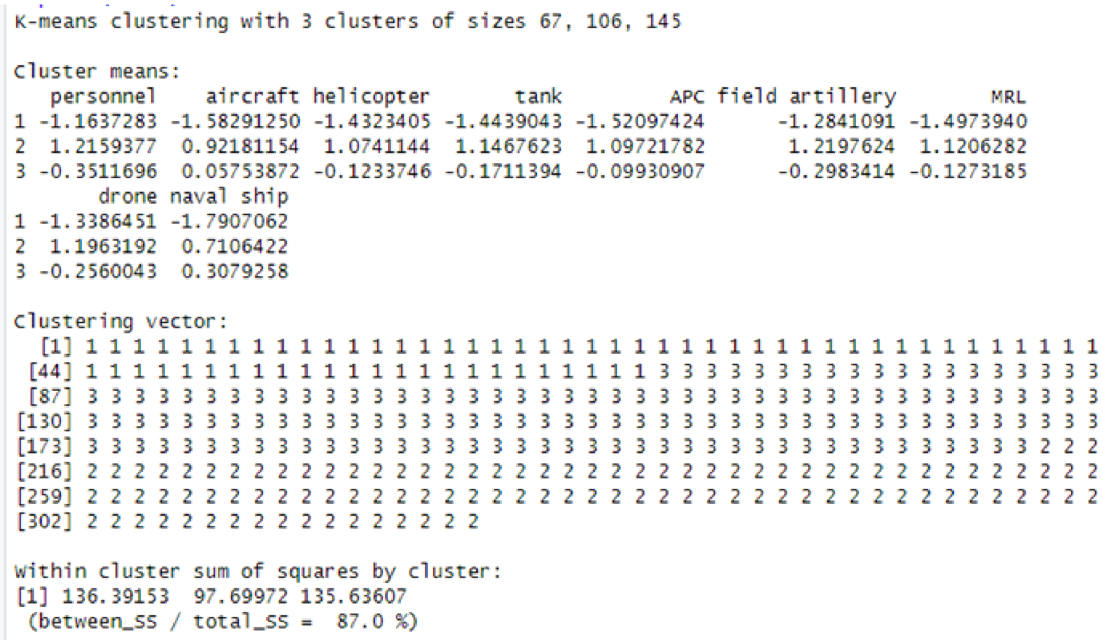
C. Gap Statistic

Figure 4. Determining the number of clusters: A. Elbow, B. Silhouette, and C. Gap Statistics.

The Statistical Gap Method is a cluster quality evaluation method used to determine the optimal number of clusters in cluster analysis. This method compares the Statistical Gap value between the actual data and the data generated randomly, which is at K=1. It can be seen in graph 4c. Based on the test of the three methods, two methods can be used to determine the optimal cluster, including elbow and Silhouette.

```
Code_3_the application of K-Means, visualization, and
profiling
# Application of K-Means clustering
>final <- kmeans(data_ukraina_scale, centers= 3, nstart =
25)
print(final)
# Cluster visualization
>fviz_cluster(final, data_ukraina_scale)
# Group profiling
>data_ukraina %>%
mutate(Cluster = final$cluster) %>%
group_by(Cluster) %>%
summarise_all("mean")
```

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v5i2.215
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023



```
K-means clustering with 3 clusters of sizes 67, 106, 145

Cluster means:
   personnel    aircraft helicopter       tank         APC field artillery        MRL
1 -1.1637283 -1.58291250 -1.4323405 -1.4439043 -1.52097424      -1.2841091 -1.4973940
2  1.2159377  0.92181154  1.0741144  1.1467623  1.09721782       1.2197624  1.1206282
3 -0.3511696  0.05753872 -0.1233746 -0.1711394 -0.09930907      -0.2983414 -0.1273185
       drone naval ship
1 -1.3386451 -1.7907062
2  1.1963192  0.7106422
3 -0.2560043  0.3079258

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [44] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [87] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[130] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[173] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2
[216] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[259] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[302] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1] 136.39153  97.69972 135.63607
 (between_SS / total_SS =  87.0 %)
```

Figure 5. Application of K-Means clustering

In the fourth stage of implementing k-means[18]–[20], based on Figure 5, there are 3 clusters with details 67, 106, and 145. Where for the average value of the variable clusters personnel, aircraft, helicopters, tanks, APCs, Field Artillery, MRL, drones, and the naval ship can be seen in clusters 1, 2, and 3.

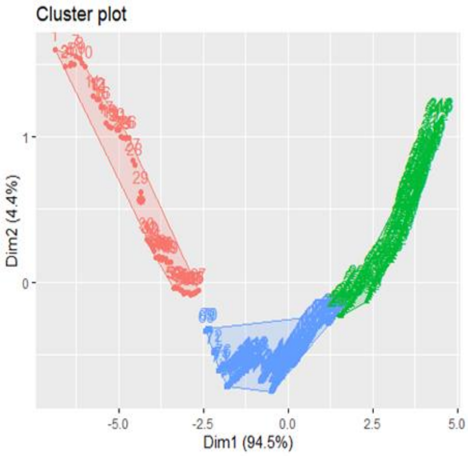
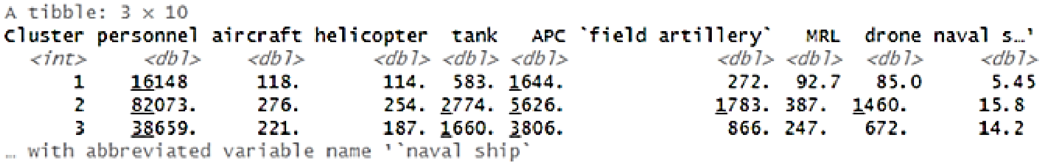


Figure 6. Data Visualization

In addition, in clusters, the number of squares indicates the distance between objects in the cluster. It can be seen that the distance for cluster 1 is 136.39153, for cluster 2 is 97.69972, and for cluster 3 is 135.63607. Therefore, the distance value for each cluster is 87.0%.

In the fifth stage, from the results of the cluster analysis using the k-means algorithm, the results are in the form of three clusters as shown in the visualization of Figure 6, namely the results of the K-Means Clustering visualization plot which consists of three clusters distinguished between three colors, namely red, blue and green. The red color describes the results of cluster 1, the blue color describes the results of cluster 2, and the green color explains the results of cluster 3. It can be seen that each plot color has a different number of members. The following is a display of K-Means Clustering results.



```
A tibble: 3 × 10
Cluster personnel aircraft helicopter  tank   APC `field artillery`   MRL  drone naval s…'
  <int>     <dbl>    <dbl>      <dbl> <dbl> <dbl>             <dbl> <dbl>  <dbl>      <dbl>
      1     16148     118.       114.  583. 1644.              272.  92.7   85.0       5.45
      2     82073.    276.       254. 2774. 5626.             1783. 387.  1460.       15.8
      3     38659.    221.       187. 1660. 3806.              866. 247.   672.       14.2
… with abbreviated variable name '`naval ship`
```

Figure 7. Group profiling

Based on the results above, the last stage can be profiling (figure 7) for each group formed (Chantaramanee et al., 2022; Lee & Chung, 2016). Cluster 1 has the lowest category of loss or loss of personnel and war equipment compared to other groups. Cluster 2 has a very high category of loss and loss of personnel and war equipment. Meanwhile, Cluster 3 experienced moderate category losses and losses of personnel and war equipment in Ukraine.

CONCLUSIONS AND SUGGESTIONS

Conclusion

Based on the results of a clustering analysis of the impact of Russia's war on Ukraine, there were losses in personnel and war equipment. Three cluster optimization methods are used using k-means, where two methods can be recommended for analysis: elbow, which produces K=3, and Silhouette, which also produces K=3. The profiling results show that losses or losses in Ukrainian personnel and war equipment are categorized into three clusters. Cluster one is in the lowest category, cluster two is in the very high category, and cluster three is in the medium category.

Suggestion

In order to broaden the scope of further research and make it more objective, the dataset must also use datasets originating from Russia. In addition, other techniques can be combined with this research to find a more optimal k value or compare it with other clustering methods.

REFERENCES

Chantaramanee, A., Nakagawa, K., Yoshimi, K., Nakane, A., Yamaguchi, K., & Tohara, H. (2022). Comparison of Tongue Characteristics Classified According to Ultrasonographic Features Using a K-Means Clustering Algorithm. *Diagnostics*, *12*(2). https://doi.org/10.3390/diagnostics12020264

Cohn, R., & Holm, E. (2021). Unsupervised Machine Learning Via Transfer Learning and k-Means Clustering to Classify Materials Image Data. *Integrating Materials and Manufacturing Innovation*, *10*(2), 231–244. https://doi.org/10.1007/s40192-021-00205-8

Darmayadi, A., & Megits, N. (2023). the Impact of the Russia-Ukraine War on the European Union Economy. *Journal of Eastern European and Central Asian Research*, *10*(1), 46–55. https://doi.org/10.15549/jeecar.v10i1.1079

Dmitry, N., & Yerkebulan, B. (2022). Clustering of Dark Patterns in the User Interfaces of Websites and Online Trading Portals (E-Commerce). *Mathematics*, *10*(18). https://doi.org/10.3390/math10183219

Govender, P., & Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). In *Atmospheric Pollution Research* (Vol. 11). Turkish National Committee for Air Pollution Research and Control. https://doi.org/10.1016/j.apr.2019.09.009

Haque, U., Naeem, A., Wang, S., Espinoza, J., Holovanova, I., Gutor, T., ... Nguyen, U. S. D. T. (2022). The human toll and humanitarian crisis of the Russia-Ukraine war: the first 162 days. *BMJ Global Health*, *7*(9), 1–11. https://doi.org/10.1136/bmjgh-2022-009550

Kaparang, D. R., Moningkey, M. J. M., & Sumual, H. (2021). The Distribution Pattern Of New Students Admissions Using The K-Means Clustering Algorithm. *International Journal of Information Technology and Business*, *3*(2), 52–60. Retrieved from https://ejournal.uksw.edu/ijiteb/article/view/4632

Lee, C., & Chung, M. (2016). Digital Forensic for Location Information using Hierarchical Clustering and k-means Algorithm. *Journal of Korea Multimedia Society*, *19*(1), 30–40. https://doi.org/10.9717/kmms.2016.19.1.030

Mailund, T. (2017). Beginning Data Science in R. In *Beginning Data Science in R*. https://doi.org/10.1007/978-1-4842-2671-1

Nerlinger, M., & Utz, S. (2022). The impact of the Russia-Ukraine conflict on energy firms: A capital market perspective. *Finance Research Letters*, *50*(May), 103243. https://doi.org/10.1016/j.frl.2022.103243

Osokina, O., Silwal, S., Bohdanova, T., Hodes, M., Sourander, A., & Skokauskas, N. (2022). Impact of the Russian Invasion on Mental Health of Adolescents in Ukraine. *Journal of the American Academy of Child and Adolescent Psychiatry*, 1–9. https://doi.org/10.1016/j.jaac.2022.07.845

Paul, A. (2015). The EU in the South Caucasus and the Impact of the Russia-Ukraine War. *International Spectator*, *50*(3), 30–42. https://doi.org/10.1080/03932729.2015.1054223

Peng, R. D. (2015). R Programming for Data Science. *The R Project; R Foundation*, 132. https://doi.org/10.1073/pnas.0703993104

Shelly, Z., Burch, R. F. V., Tian, W., Strawderman, L., Piroli, A., & Bichey, C. (2020). Using K-means clustering to create training groups for elite american football student-athletes based on game demands. *International Journal of Kinesiology and Sports Science*, *8*(2), 47–63. https://doi.org/10.7575//aiac.ijkss.v.8n.2p.47

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*, *2*(2), 226–235. https://doi.org/10.3390/j2020016

Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, *17*(January), 100179. https://doi.org/10.1016/j.imu.2019.100179