International Journal of Transport Development and Integration
Vol. 9, No. 1, March, 2025, pp. 131-141
Journal homepage: http://iieta.org/journals/ijtdi

A Clustering Approach to Enhancing Marine Transportation Services in the Thousand Islands Regency, Indonesia

Rossi Passarella[1*], Rani Febrianti[1], Marsella Vindriani[1], Mohd Shahriman Adenan[2]

[1] Department of Computer Engineering, Faculty of Computer Science, Universitas Sriwijaya, Sumater Selatan 30662, Indonesia
[2] Smart Manufacturing Research Institute, Universiti Teknologi MARA, Shah Alam 40450, Malaysia

Corresponding Author Email: passarella.rossi@unsri.ac.id

Copyright: ©2025 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/ijtdi.090112

Received: 13 January 2025
Revised: 19 February 2025
Accepted: 27 February 2025
Available online: 31 March 2025

Keywords:
port management, K-means clustering, marine transportation, Thousand Islands Regency, regional development

ABSTRACT

Port infrastructure is crucial for inter-island connectivity and marine transportation services in the Thousand Islands Regency, Indonesia. Ensuring better connectivity and accessibility for island residents is essential. This research aims to improve marine transportation services in the Thousand Islands Regency by applying a clustering approach. The goal is to enhance regional transportation services by identifying patterns and gaining insights from historical data. The K-means clustering method was employed in this research to analyse historical data and categorize ports into three distinct clusters: low capacity, medium capability, and high capacity. The research identified three clusters: low-capability ports, medium-capability ports, and high-capability ports. The government has identified these clusters as focal points for improving regional transportation services. The findings highlight the essential role of marine transportation in facilitating connectivity and supporting the tourism industry in the Thousand Islands Regency. This analysis provides a comprehensive understanding of the current situation and offers a basis for informed decision-making in future port management strategies. The research urges stakeholders and policymakers to prioritise improvements at the identified ports to enhance service quality, connectivity, and regional development.

1. INTRODUCTION

The Thousand Islands Regency, part of the DKI Jakarta Province, faces unique challenges in its marine transport services [1]. Despite its proximity to Indonesia's economic centre, Jakarta [2], the Thousand Islands' interisland connectivity still faces constraints, particularly in efficient port management. The significant increase in the number of tourists post-COVID-19 pandemic in 2022 has exacerbated this problem [3], causing passenger congestion at the port and highlighting the urgent need for better port management [4, 5].

Previous research on maritime transport has used clustering methods, such as K-means, for various purposes, including port status control and port classification based on strategic levels [6, 7]. However, no research has specifically analysed historical passenger data from the Thousand Islands to identify patterns or classify ports based on their operational capacities. This research gap is important, as a better understanding of port capacity and passenger patterns can provide a basis for local governments to optimise port management and improve maritime transport services.

This research aims to fill the gap by applying the K-means method to historical data about ship passengers from Jakarta to the Thousand Islands. By identifying port clusters based on operational capacity, this research will provide valuable insights for the DKI Jakarta Provincial Government in prioritising port improvements. In addition, this research will highlight the important role of marine transport in connecting the dispersed communities in the Thousand Islands and supporting the achievement of the Sustainable Development Goals (SDGs) [8].

Specifically, this research will explore how improved marine transport in the Thousand Islands contributes to (1) SDG 14: life below water, through the use of environmentally friendly vessels [9]; (2) SDG 8: decent work and economic growth, by encouraging sustainable tourism and creating jobs [10]; (3) SDG 10: reduction of inequalities, by ensuring equitable and affordable access to marine transport [11]; and (4) SDG 11: inclusive, safe, resilient, and sustainable cities and settlements, by developing integrated marine transport systems [12].

The research is organised into six sections: introduction; Thousand Islands; data and methods; results; discussion; and conclusion. The introduction talks about the research questions and why they were important. The Thousand Islands section describes the geography and transportation problems. The data and methods section explains the research methodology. The results section lists the main findings. The discussion section analyses the results. And finally, the conclusion section sums up the research's contributions and

policy implications.

2. THE THOUSAND ISLANDS

The Thousand Islands is an administrative district under the jurisdiction of the Special Capital Region of Jakarta, Indonesia, comprising a group of islands to the north of Jakarta in Jakarta Bay [13]. The Thousand Islands consist of 110 islands, covering an area of 1,180 hectares, extending from south to north. According to the 2020 population census, the total population of the Thousand Islands is 29,230 people, spread across ten inhabited islands: Untung Jawa Island, Pari Island, Lancang Island, Tidung Island, Tidung Kecil Island, Pramuka Island, Panggang Island, Harapan Island, Kelapa Island, and Sebira Island. The remaining 100 islands are uninhabited, which serve as tourist attractions.

Sabira Island is the outermost and farthest island from mainland Jakarta, located 119 km (64.2549 nautical miles) from Muara Angke port, Jakarta, which is about 2.5 hours by passenger ship. Figure 1 illustrates the distribution of the main inhabited islands by displaying the port locations based on the GPS location data summarized from the dataset.

Regarding service routes, the Provincial Government of Jakarta has established two official ports to provide sea transportation for the Thousand Island communities: Muara Angke and Marina Ancol ports. Each of these ports serves several routes, with Muara Angke serving ten islands in a round trip and Marina Ancol serving 28 islands in a round trip.

Figure 1. Passenger ship port locations based on GPS points used in this research

3. DATA AND METHODS

This section explains the data and methods used in the research. In general, this section will be described as subsections such as data collection, data quality and scope, data preprocessing, and clustering algorithm.

3.1 Data collection

The data for this research was obtained from the Jakarta Data Portal website [14] in *.csv (Comma Separated Values) format. The data collected includes ship passenger information from 1 January 2018 to 31 December 2021. This dataset consists of six variables with a total of 12,009 observations, namely date, port, ship departure, ship arrival, boarding passengers, and disembarking passengers. For our research, we used three variables: port, ship departure, and boarding passengers. The port variable includes 10 port names: Muara Angke, Marina Ancol, Untung Jawa, Lancang, Pari, Tidung/Payung, Pramuka/Panggang, Kelapa, Harapan, and Sabira. Spearman and Pearson correlation analysis showed that the correlation value between the ship departure and passenger boarding variables was 0.71, so these two variables were selected for further analysis.

3.2 Data quality and scope

To ensure the reliability and validity of our analysis, we conducted a thorough review of the dataset. The data obtained from the Jakarta Data Portal were considered reliable due to their official source. However, to enhance the scope of our research and provide a more comprehensive understanding of the factors influencing marine transportation services, we acknowledge the need to expand data sources. Specifically, incorporating data related to port infrastructure details, such as dock capacity, terminal facilities, and ship traffic information, would provide valuable insights. Future research should aim to integrate these additional data sources to create a more robust analysis.

However, we still try to give an overview of only two ports on the mainland of Java, namely Muara Angke and Marina Ancol. Muara Angke port is a port built by the DKI Jakarta Regional Government since 2004, with a construction budget of IDR 130 billion (USD 7,973,017). It has an area of about 3.4 square hectares with a capacity of 50 ships. In addition, a breakwater has also been constructed along 1.4 kilometres. Supervision of shipping safety at this pier for ships is carried out by the Tanjung Priok Sea Guard Base, while in terms of facilities for the availability of adequate docks for residents and tourists, it is carried out by the DKI Jakarta Regional Government [15]. The Muara Angke port focuses on serving the people who live in the Thousand Islands cluster, while the Muara Ancol port is more focused on tourism services for recreation in the Thousand Islands cluster [16].

3.3 Data preprocessing

The collected raw data undergo pre-processing before analysis, including data cleaning, selection, and reduction. Data cleaning, the first step in the data mining process, involves removing null or empty values to ensure data consistency and accuracy. The observation results for six variables, with 12,009 observations, indicated that there is no missing data in either categorical or numerical variables. Orange software observations confirmed this, showing a missing indicator value of '0' for all variables.

The data selection stage involves selecting relevant research variables to reduce data complexity [17]. As explained in the data collection subsection, this study focuses on three main variables: port, ship departure, and number of boarding passengers. These variables were chosen to categorise ports based on passenger usage patterns. These variables are chosen based on the research objective of grouping ports according to the number of passengers boarding. The port variable is required to identify the port name. The ship departure variable is relevant because it affects service availability and port capacity. The passenger boarding variable is the primary focus, as it reflects passengers' interests and demands at each port.

In this research, we conducted a preliminary data analysis to identify potential outliers. However, based on observation and descriptive statistical analysis, no significant outliers were

found that could affect the clustering results. The data we use, which comes from the official government data portal, has gone through an initial validation and cleaning process. Therefore, we have scrutinised the data, and we found that it is relatively clean from extreme outliers.

In addition to the outlier issue, we also looked at the dataset on missing values, and the fact that there were no missing values in the dataset we used, we realised that in real-world scenarios, missing data is common. Therefore, it is important to consider appropriate data imputation methods in case missing data occurs in the future.

The initial step in data reduction involves normalising the values of numeric variables using the MinMaxScaler method, which is necessary due to the non-normal distribution of the data [18]. The skewness test results for the numeric variables-the skewness scores-reveal distinct patterns in passenger boarding and ship departure data. A significantly high positive skewness of 7.46 for passenger boarding indicates a heavily right-skewed distribution, with most events involving few passengers and a few instances of enormous numbers. In contrast, the positive skewness of 3.42 for ship departures suggests a less pronounced right-skewed distribution, with most days having moderate departures and certain periods experiencing significantly higher numbers. Understanding these skewness patterns is crucial for informed decision-making in transportation planning, resource allocation, and operational efficiency.

In this research, principal component analysis (PCA) is performed for data reduction after normalization. PCA is a statistical technique that reduces the dimensions of data by projecting it into a lower-dimensional space [19]. PCA is necessary to process this dataset and expedite cluster formation and calculation.

3.4 Clustering algorithm

To identify the usage patterns of ship passengers, we used the K-means clustering method. K-means was chosen due to its simplicity of implementation, computational efficiency, and ability to handle large-scale numerical data. This method is also effective in identifying spherical or convex clusters, which correspond to port usage patterns based on passenger capacity.

The clustering process begins by determining the optimal number of clusters using the elbow method [20]. Next, the K-means algorithm is implemented to cluster the data based on the Euclidean distance between the data points and the cluster centroid. Iterations are performed until there is no significant change in the data grouping.

For comparison, hierarchical clustering and density-based spatial clustering of applications with noise (DBSCAN) clustering methods were also applied. Hierarchical clustering is used to compare hierarchical cluster structures, while DBSCAN is explored to identify clusters based on data density. The comparison of the results of these three methods aims to provide a more comprehensive and robust analysis of clustering patterns in ship passenger data.

3.5 Model validation with cross-validation

To ensure the stability and generalisation of the K-means clustering results, we apply cross-validation techniques. In the context of K-means, the main purpose of cross-validation is to test the stability and robustness of the clustering results [21]. Although K-means is an unsupervised algorithm and is not traditionally evaluated by cross-validation like supervised models, in this context, cross-validation is used to assess how stable the clustering results are when the model is applied to different subsets of data [21]. This is important to ensure that the clustering patterns identified are not just coincidental to the specific dataset used, but reflect a more general structure in marine transport passenger data.

In this study, we used a simple data splitting approach for cross-validation. The dataset is divided into two parts: the training set (e.g., 80% of the data) and the validation set (e.g., 20% of the data). The K-means algorithm is trained on the training set to determine the cluster centres. Then, the cluster centres obtained from the training set are used to assign clusters to the data in the validation set.

The cross-validation steps performed are as follows:

1) Data Division: The dataset is randomly divided into training set (80%) and validation set (20%).
2) Training K-means Model on Training Set: The K-means algorithm is applied to the training set to find the optimal cluster centres (with a predefined number of clusters k using the Elbow method).
3) Model Application on a Validation Set: The cluster centres obtained from the training set are used to assign each data point in the validation set to the nearest cluster. This assignment is based on the Euclidean distance to the nearest cluster centre.
4) Comparison of Clustering Results: The characteristics of the clusters formed on the validation set (e.g., cluster size, distribution of port variables within clusters, mean values of numeric variables) were compared with the characteristics of the clusters formed on the entire dataset (as reported in the previous Results section). This comparison is done qualitatively to assess the stability and consistency of the clustering patterns.

This cross-validation approach allows us to assess whether the clusters identified by K-means are consistent and can be generalised to new data not used in model training. If similar cluster patterns are seen in the validation set, this gives more confidence in the validity and generalisability of the results.

4. THE K-MEANS ALGORITHM

In unsupervised learning, the K-means clustering method groups the data with similar characteristics into a single cluster or group [22]. The K-means method requires three user-specified parameters: the number of clusters (K), initial cluster initialisation, and distance measurement. The process begins by determining the number of clusters to form, followed by establishing the initial centre as the average of the data in each cluster. Next, the distance between each centre and the cluster members is calculated. The K-means algorithm stops iterating when there is no change in the grouping of data between clusters [20]. Below is an illustration of the algorithm's working process.

4.1 Initial cluster testing

To determine the optimal number of clusters, we used the elbow method, as mentioned in the clustering algorithm

subsection. This method calculates the percentage difference between the number of clusters formed until reaching the elbow point [23]. The resulting concept combines cluster values to form a data model for optimal cluster identification.

Additionally, the percentage calculation is used to compare the number of added clusters. The graph serves as a source of information, displaying the percentage difference between each cluster value. Significant angles or decreases between the graph's first and second cluster values determine the optimal cluster values. This comparison is achieved by calculating each cluster value's SSE (sum of square error). The SSE value decreases as the K value (number of clusters) increases [24].

The SSE statistical method measures the difference between the observed and true values, effectively quantifying the discrepancy between the observed data and the previous prediction model. Researchers commonly use it as a reference to select the optimal cluster. Eq. (1) displays the SSE formula.

$$SSE = \sum_{i=1}^{n} d_i^2 \quad (1)$$

where, the variable d indicates the distance between the data and the cluster centre.

5. RESULTS

This section describes the clustering process using the K-means algorithm and analyses the results. We then discuss the steps in order, from determining the number of initial clusters and running the K-means algorithm to analysing each cluster that forms.

5.1 Testing the number of clusters

To achieve the research objective of clustering ports based on passenger usage patterns, the Elbow method is applied to determine the optimal number of clusters, as illustrated in Figure 2. This method relies on analysing the variation in within-cluster SSE as the number of clusters increases. The main goal of the elbow method is to find the point where adding more clusters doesn't make a big difference in lowering the SSE. The result is shown by a sharp change in the curve that looks like an "elbow." This method was chosen because it is easy to understand and works well at finding the best point, especially when the data doesn't have a clear cluster structure or an idea of how many clusters are likely to be there [20]. In this way, the elbow method gives a strong base for choosing the right number of clusters, which makes sure that the clustering model that is made accurately shows the patterns in the data.

The results from validation using the Elbow method show that the optimal number of clusters for this study is three, with an SSE score of 27.869. This score represents the point at which adding further clusters does not provide a significant decrease in SSE, indicating that three clusters are sufficient to capture the main variations in the data. This number of clusters has important implications in the context of our research, which is the analysis of marine transport services in Thousand Islands. By identifying three clusters, we are able to categorise the port based on passenger usage patterns and ship departures, which allows for a more in-depth understanding of its operational dynamics. Based on the needs and traits of each identified cluster, these results give local governments a strong foundation for making better policies for port management and improving marine transport services.

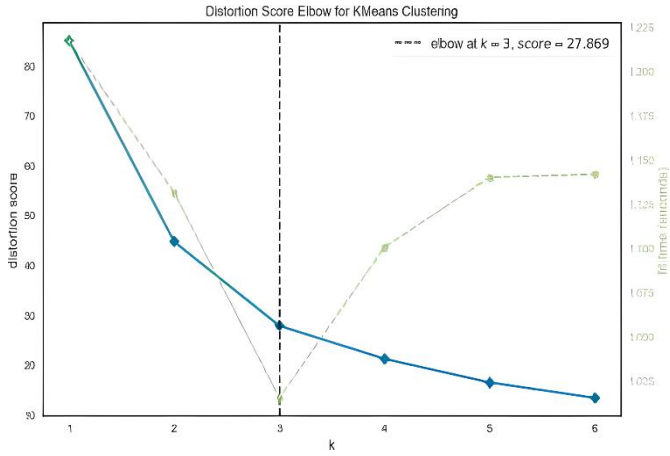

Figure 2. Test the number of clusters using the elbow method

5.2 K-means clustering

With the pre-processed data and the optimal number of clusters set at three, we proceeded to the clustering stage using the K-means method. The results of this process can be seen in Figure 3. In the figure, the three clusters formed are marked with coloured circles: blue (Cluster 1), red (Cluster 2), and green (Cluster 3). The centroid of each cluster is marked with a yellow dot. The X-axis in the figure shows the number of passenger ship departures, while the Y-axis shows the number of boarding passengers. The number of data points included in each cluster is summarised in Table 1.

Table 1 summarises the results of data clustering, showing the distribution of data points among the three clusters that formed. Cluster 1 is the largest cluster with 10,123 data points, covering 84% of the entire dataset. This indicates that the majority of the data is clustered in this cluster. On the other hand, Cluster 2 has 1,736 data points, which is equivalent to 14% of the total dataset, showing a smaller but still significant size. Cluster 3 is the smallest cluster with only 150 data points, covering only 1% of the total dataset. This indicates that a small portion of the data has different characteristics and is clustered separately. Overall, Table 1 provides a clear picture of the distribution of data points within each cluster, highlighting the differences in size and proportion between the resulting clusters.

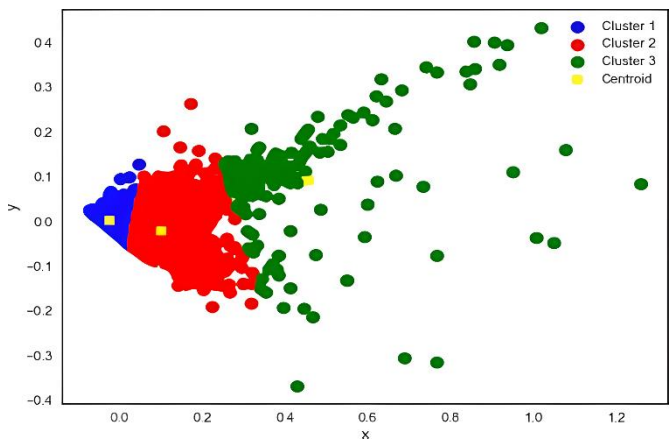

Figure 3. Visualization results of the clustering

Table 1. The number of data points for each cluster

Clusters Number	Number of Data Points	Percentage
1	10,123	84%
2	1,736	14%
3	150	1%

5.3 Data analysis on Cluster 1

The clustering process resulted in 10,123 data points falling into Cluster 1. To understand the characteristics of this cluster, we used Orange Software to analyse the data description. The analysis results show that the mean value of ship_departure is 3.68, and the mean value of passenger_boarding is 99.91. Meanwhile, the median value of ship departure is 4, and the median value of passenger boarding is 70. The port variable is dominated by the name 'Lancang.'. A visualisation of the distribution of Cluster 1 data points can be seen in Figure 4; the yellow dot indicates the centroid of the data group. Table 2 provides details of the data distribution for each port in Cluster 1.

To understand the composition of Cluster 1, Table 2 details the number of data points contained in each port. This data shows how the data points are distributed among the various ports in this cluster.

Cluster 1, which includes 10,123 data points-or 84.3% of the entire dataset-was further analysed to understand the distribution of data within it. Table 3 provides a more in-depth breakdown, showing how the data points are distributed based on the number of ship departures, ranging from one to nine, as well as the total passengers served at each port. Based on this data, it is known that the maximum number of ship departures in Cluster 1 is nine, with a total of 1,011,392 passengers. Pramuka/Panggang Island was identified as the most dominant port in this cluster, accounting for 16% of the total passenger data, while Sabira Port had the lowest number of passengers. Table 3 provides more detailed data on the number of passengers per port and per number of ship departures within Cluster 1.

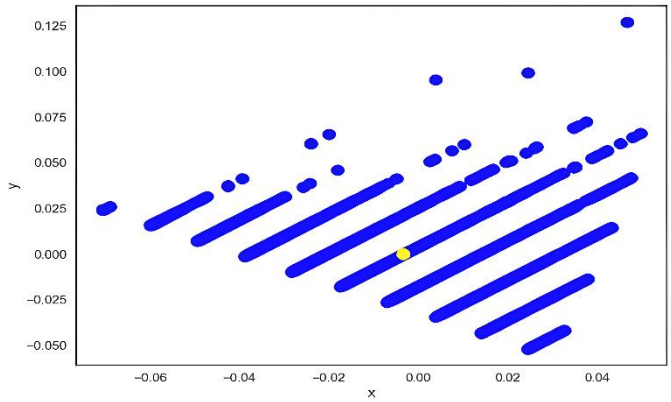

Figure 4. Data visualization Cluster 1 with 10,123 data points

Table 2. Recapitulation of the number of data distribution in Cluster 1 based on the port

Ports	Number of Data Points
Lancang	1,238
Kelapa	1,233
Pramuka/Panggang	1,190
Harapan	1,186
Tidung/Payung	1,126
Pari	1,082
Untung Jawa	879
Sabira	823
Muara Angke	708
Marina Ancol	658

Table 3. Detailed number of passengers in Cluster 1 based on the port

Ports	Number of Ships									Total of Passenger
	1	2	3	4	5	6	7	8	9	
Harapan	9802	18838	14907	17460	19480	6794	1481	745	20	89530
Kelapa	711	10476	25154	28026	22145	11958	4013	1021	58	103562
Lancang	83	699	4447	54492	9937	17333	222	-	-	87213
Marina ancol	89	2431	14106	7843	6560	7301	9727	8048	-	56105
Muara angke	63	574	2078	15833	30310	42218	41254	9322	-	141652
Pari	1263	1431	22212	44855	32832	20177	7342	2323	576	133012
Pramuka/panggang	173	18709	36883	31746	36591	27494	4627	1680	-	157903
Sabira	5363	1062	301	213	-	-	-	-	-	6981
Tidung/payung	786	13136	26613	39703	28184	27412	10388	865	-	147087
Untung jawa	112	133	1338	5375	15960	27769	23132	11432	3096	88347
Grand Total	18445	67489	148039	245546	201999	188456	102186	35436	3750	1011392

The interpretation of Cluster 1 illustrates that Cluster 1, as the largest cluster, represents a group of ports with relatively low passenger and ship departure volumes. The low number of average ship departures (mean 3.68, median 4) and slightly higher but still moderate passenger numbers (mean around 99, median 70) show that these ports are reliable but do not get a lot of traffic. The dominance of Lancang port in this cluster, which is one of the second largest islands, reinforces this interpretation. Ports in Cluster 1 are likely to be ports serving medium-demand routes or ports on islands that are less densely populated or less popular as major tourist destinations. From the point of view of developing port services, ports in Cluster 1 may need to focus on keeping their operations as efficient and cost-effective as possible without having to make big investments in expanding their capacity, unless there are significant changes in how much people will want to use them in the future.

5.4 Data analysis on Cluster 2

The results of the previous clustering process show that there are 1,736 data points in Cluster 2. The descriptive data indicate that the departing passenger ship (ship_departure) variable has a minimum of 2 ships and a maximum of 32 ships, with a mean of 12.03, a median of 11, and a dispersion of 0.34. The passenger_boarding variable ranges from 0 to 2,864

passengers, with a mean of 611.51, a median of 497.50, and a dispersion of 0.67. The port that appears most frequently in this cluster is Marina Ancol.

Additionally, the data distribution in scattered points exhibits a diagonal pattern similar to that in cluster 1. Figure 5 displays the data distribution for Cluster 2. The red dots represent the data points, while the yellow dot indicates the centroid of Cluster 2.

Table 4 presents a recapitulation of the data distribution in Cluster 2 by port, which provides an overview of the number of data points distributed in each port in this cluster. From this table, it can be seen that Marina Ancol port has the highest number of data points at 568, indicating its dominance in Cluster 2. Muara Angke Port comes second with 442 data points, followed by Untung Jawa with 342 data points. Pari, Tidung/Payung, Harapan, and Pramuka/Panggang ports have a smaller number of data points: 151, 110, 56, and 53, respectively. Kelapa and Lancang ports have the fewest data points, at 10 and 4, respectively. These data show significant variation in the distribution of data points between ports within Cluster 2, with Marina Ancol and Muara Angke being the most representative ports.

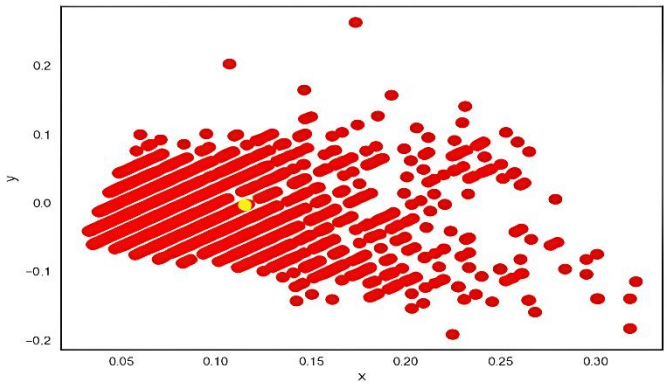

Figure 5. Data visualization on Cluster 2

Table 4. Recapitulation of the number of data distribution in Cluster 2 based on the port

Ports	Number of Data Points
Marina Ancol	568
Muara Angke	442
Untung Jawa	342
Pari	151
Tidung/Payung	110
Harapan	56
Pramuka/Panggang	53
Kelapa	10
Lancang	4

The data in Cluster 2 is grouped by the number of passenger ship departures, from 2 to 14 ships, with a total of 1,061,587 passengers served. Muara Angke Harbour recorded the highest number of passengers: 297,533 people, or about 28% of all passengers in Cluster 2. However, this contrasts with Table 4, which shows the number of Muara Angke data points is only 442. This difference suggests that although the frequency of departures at Muara Angke is not as much as Marina Ancol, Muara Angke carries a larger number of passengers per departure. This corroborates the evidence that Muara Angke is focused on the transportation of Thousand Islands residents, whereas Marina Ancol is more dominant in tourist services.

As for the interpretation of Cluster 2, it can be explained that Cluster 2 shows characteristics of ports with moderate to high volumes of ship and passenger departures. Compared to Cluster 1, Cluster 2 has higher average ship departures (mean 12.03) and a much larger average number of passengers (mean 611.51). The dominance of Ancol Marina Port, which is known as a tourist port, indicates that Cluster 2 represents ports that serve major tourist routes. Although Muara Angke Harbour also appears in this cluster, the interpretation is different. The data shows that Muara Angke, despite not having the highest frequency of departures in Cluster 2 (the number of data points is lower than Marina Ancol), carries a significant number of passengers per departure. This confirms Muara Angke's role as the main port for transporting local Thousand Islands residents, where each ship tends to carry more resident passengers than tourists. Ports in Cluster 2 need to focus on efficient traffic management, adequate facilities to accommodate larger passenger volumes, and strategies to separate or manage tourist and resident passenger flows for convenience and safety.

5.5 Data analysis on Cluster 3

According to the results table (Table 1) from the previous clustering process, Cluster 3 contains 150 data points. The statistical descriptions for these data points show that, for the variable ship_departure, the mean, median, and dispersion values are 27.38, 24, and 0.41, respectively. For the variable passenger_boarding, the mean, median, and dispersion values are 3,217.50, 2,797, and 0.49, respectively.

The data description of these two numeric variables indicates that the data distribution in Cluster 3 maintains relatively stable values and does not exhibit extreme variation, as both variables have low dispersion values. The statistical description of the categorical variable port identifies Muara Angke port as the median. Figure 6 displays the data point distribution of Cluster 3 as dots arranged in a random pattern. In this figure, the green dots represent the data distribution, with the ship_departure variable on the x-axis and the passenger_boarding variable on the y-axis, while the yellow dot serves as the cluster's centroid. Table 5 provides detailed information about the data distribution for each port in Cluster 3.

The interpretation of Cluster 3 can be drawn as the Cluster 3, although smallest cluster, which represents ports with very high volumes of ship and passenger departures. Cluster 3 has the busiest ports, as shown by the much higher mean and median values for both variables compared to Clusters 1 and 2 (average ship departures of 27.38 and average passengers of 3217.50). It is said that Muara Angke is the middle port in this cluster, but a better way to look at it is that Cluster 3 as a whole shows the most important events or strange things in the data. The small number of data points (only 1% of the dataset) suggests that these events with very high departure and passenger volumes are relatively rare but significant when they do occur. These events may be related to major holidays, festivals, or other special events that cause dramatic spikes in port usage. To handle these sudden increases in demand, the ports in Cluster 3 (whose names are shown in Table 5) need to be carefully and adaptively planned for their capacity. Risk management and contingency plans are also critical to managing potential congestion and ensuring smooth operations during peak traffic periods.

Table 5. Recapitulation of distribution point data in Cluster 3

Ports	Number of Data Points
Muara Angke	93
Untung Jawa	22
Marina Ancol	17
Pari	10
Tidung/Payung	7
Lancang	1

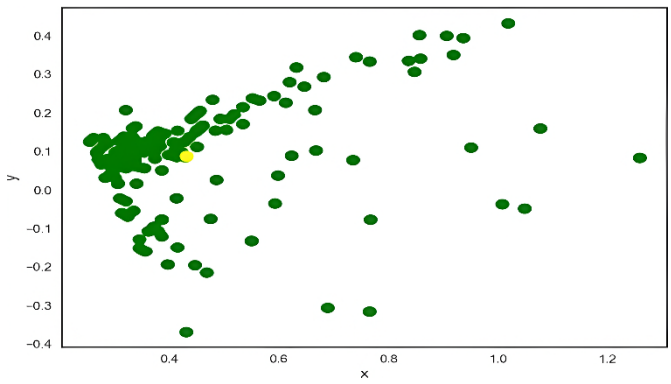

Figure 6. Data visualization on Cluster 3

5.6 Cross-validation results with data splitting

To validate the K-means clustering results, we perform cross-validation using the data splitting method as described in the Methods section. The dataset is divided into training set (80%) and validation set (20%). The K-means algorithm was trained on the training set with the optimal number of clusters k=3. The cluster centres generated from the training set were then used to assign clusters to the data in the validation set.

In general, cross-validation was a good idea because the cluster patterns that were made on the validation set were very similar to the clusters that were made from the whole previous dataset. This similarity can be seen in several key areas. Firstly, the cluster size distribution on the validation set is relatively similar to the original dataset, where Cluster 1 remains the largest, followed by Cluster 2 and Cluster 3 as the smallest, with similar proportions of data. Secondly, the distribution of ports within clusters also shows consistency. Cluster 1 in the validation set remains dominated by Lancang and Pramuka/Panggang ports, Cluster 2 by Marina Ancol and Muara Angke, and Cluster 3 maintains its characteristics with the dominance of Muara Angke port and the least number of data points. Thirdly, the characteristics of numerical variables such as ship_departure and passenger_boarding within each cluster in the validation set also show a range of mean and median values comparable to the original dataset.

Although there is a slight variation in numbers, the general pattern of variable values in each cluster, for example Cluster 1 with the lowest average passenger_boarding and Cluster 3 with the highest, is maintained. In conclusion, these cross-validation results provide a strong indication that the resulting K-means clustering is stable and has good generalisability. The fact that the cluster patterns remain consistent even when the model is trained on a subset of the data (training set) and tested on new data (validation set) reinforces the belief that the clusters formed are not simply artefacts of the specific dataset used but rather reflect more fundamental patterns in the marine transport passenger data in the Thousand Islands.

6. DISCUSSION

6.1 Interpretation of clustering results for port service development

This section presents the clustering results for the ten ports of interest and the number of passengers using ships services in the Thousand Islands region. Table 6 provides a summary of the ports, clusters, and number of passengers. To better illustrate the position of busy port services, we also express the total number of passengers as a percentage. Table 6 indicates that the principal ports on the Indonesian island of Java, specifically in the capital city of Jakarta, are Marina Ancol and Muara Angke, which have a high volume of passengers. In Cluster 1, Muara Angke is the second busiest port, while Cluster 2 is the busiest one. During the four-year data collection period, Muara Angke handled 773,866 passengers, accounting for 30% of the total passengers in the Thousand Islands, which totalled 2,555,604 passengers. Marina Ancol is the second busiest port after Muara Angke, with a 14% increase in the number of passengers served. The Jakarta city government has prioritized these two ports, which are excluded from further analysis.

In the meantime, the research focuses on ports other than Ancol Marina and Muara Angke. Ports that warrant further attention and enhanced services include those of Untung Jawa, Pari, and Tidung/Payung Islands. The passenger numbers at these three ports represent approximately 12%, 11%, and 11% of the total, or between 250,000 and 300,000 passengers each. These three ports dominate Clusters 2 and 3, so the Jakarta City government should prioritise developing shipping services for these locations.

On Untung Jawa Island, there is a tourist village with an area of approximately 40.10 hectares and a population of about 2,440 people. This tourist village is a popular destination for both domestic and foreign tourists, with many visitors flocking to the island on weekends and holidays. Attractions such as marine tourism, historical sites, culinary experiences, diving, snorkelling, mangrove adventures, and fishing draw tourists to the island. Its proximity to Jakarta's mainland makes Untung Jawa Island a popular destination, with a relatively short travel time of approximately 30 minutes by motorboat.

The next port of call is Pari Island, situated approximately 30.3km (16.36 nautical miles) from Muara Angke port in Jakarta. Pari Island, along with Untung Jawa Island, is a popular tourist destination for residents of Jakarta. The convenience of transportation from Muara Angke Port and Marina Ancol Port has led to a notable increase in the number of tourists. Pari Island has an area of approximately 41.32 hectares and a population of about 930 individuals. Marine tourism encompasses beaches, mangrove forests, snorkelling, and diving activities.

Tidung/Payung Island is one of the principal tourist destinations in the Thousand Islands, with a population of approximately 5,244 individuals inhabiting an island of 106.90 hectares. Although situated at a considerable distance from the mainland, at 43.6 km (23.54 nautical miles) from Muara Angke port, the island is a popular destination for those seeking a less crowded environment. Consequently, the number of passengers arriving at the port has increased.

The surge in passenger numbers has resulted in a notable increase in the density of traffic at the ports in the Thousand Islands. Consequently, port management assumes a pivotal role in enhancing operational efficacy and delivering services

to the community. Prioritising improvements to shipping services from and to these three islands is imperative to ensure the safety and security of passenger ship users. This will help reduce congestion and ensure the unimpeded operation of maritime transportation activities in the region.

Table 6. Recapitulation of the distribution of the number of passengers according to port

Ports	Number of Passengers			Total	%
	Cluster 1	Cluster 2	Cluster 3		
Harapan	89.530	68.806	-	158.336	6%
Kelapa	103.562	8.251	-	111.813	4%
Lancang	87.213	1.825	55	89.093	3%
Marina ancol	56.105	273.944	28.678	358.727	14%
Muara angke	141.652	297.533	334.681	773.866	30%
Pari	133.012	120.060	26.599	279.671	11%
Pramuka/panggang	157.903	41.004	-	198.907	8%
Sabira	6.981	-	-	6.981	0%
Tidung/payung	147.087	111.068	19.183	277.338	11%
Untung jawa	88.347	139.096	73.429	300.872	12%

Based on historical data and the formation pattern of data groups, it is argued that the government should prioritise service provision for the ports in Cluster 3, particularly those at Untung Jawa, Tidung/Payung, and Pari islands. The objective is to enhance the efficiency, safety, and quality of the services provided, which will positively impact individuals using sea transportation in the region. This will guarantee that the community's sea transportation needs are appropriately met.

6.2 Comparison of K-means results with other clustering algorithms

As an additional validation of the K-means clustering results, we also applied hierarchical clustering and DBSCAN algorithms on the same dataset. The point of this comparison is to find out how stable and reliable K-means is at grouping ports together and to see if other algorithms offer different ways of grouping ports together based on how passengers use them.

6.2.1 Hierarchical clustering results

Hierarchical clustering, using the Ward Linkage method, produced a dendrogram that was visualised to identify the hierarchical structure between ports. In the resulting dendrogram, a fairly clear cluster structure also indicates three main groups of ports, which is generally consistent with the optimal number of clusters found by the elbow method of K-means. While the hierarchical structure gives us more details about how close two ports are to each other, the cluster membership at the relevant cut level of the dendrogram is very similar to the clusters that K-means created. For instance, in K-means, the most important ports in Cluster 1 tend to group together on the same branch in the hierarchical clustering dendrogram. The same is true for Clusters 2 and 3. This observation provides additional validation that the clustering identified by K-means is not an artefact of the algorithm but rather reflects the inherent structure of the data.

6.2.2 DBSCAN clustering results

As a density-based clustering algorithm, DBSCAN was used to find out if port clustering could also be found by looking at the amount of data and to look for noise or outliers in the data. With the parameter epsilon (neighbour radius) and minPts (minimum number of neighbouring points) optimised through evaluation methods such as silhouette scores, DBSCAN also successfully identified three main clusters. Interestingly, DBSCAN tends to group most of the data points into one large dense cluster, which is similar to Cluster 1 in K-means. The other two clusters identified by DBSCAN are smaller and denser, similar to Clusters 2 and 3 of K-Means, although with slightly different numbers of members. The main difference with K-means is that DBSCAN identifies a small number of data points as noise (outliers), which are not included in any cluster. These noise data points may represent extremes in passenger numbers or ship departures, which may not be representative of general port operational patterns.

6.2.3 Comparison and implications

Overall, the results of hierarchical clustering and DBSCAN provide confirmation of the K-means clustering results. These three algorithms, although different in mechanism and approach, tend to identify three main groups of similar ports. This result reinforces the belief that the clustering we found is robust and not dependent on a particular clustering algorithm.

However, each algorithm provides slightly different insights. K-means, with its simplicity, provides clear and easily interpreted partitions based on centroid distance. Hierarchical clustering adds a hierarchical dimension, which could potentially be useful for understanding the relationship between port groups in more detail in the future. With its ability to find noise, DBSCAN points out possible outliers in the data that need to be looked at more closely. However, in this research, the noise found did not significantly change how the main port groups were interpreted.

The choice of K-means as the main algorithm in this research still valid because it is fast to compute and simple to understand, which is important for the original research goal of giving useful advice to local governments. Validation with other algorithms, such as hierarchical clustering, and DBSCAN provides an additional layer of confidence in the results obtained and opens up opportunities for further research with different algorithms to gain a more comprehensive understanding of marine transport patterns in the Thousand Islands.

6.3 Implementation and stakeholder engagement suggestions

We propose a detailed implementation roadmap to follow up on the research findings that identified Untung Jawa, Pari, and Tidung/Payung as ports needing service improvements. The roadmap is designed to provide practical guidance for policymakers in implementing improvements. The first step is a thorough infrastructure evaluation to assess the condition of the jetties, terminal facilities, and accessibility. The results of this evaluation will form the basis for the development of an improvement plan that includes physical improvements, capacity building, and the application of new technologies. Next, the procurement of adequate resources, such as budgets, labour, and equipment, will be undertaken. Implementation of the improvements will be phased in, starting with the highest priority, followed by periodic monitoring and evaluation to ensure their effectiveness. The roadmap also includes a detailed timeline, from initial evaluation to ongoing evaluation, as well as details of the resources required from the

local and central governments.

However, the success of this implementation depends not only on careful planning but also on the active involvement of all stakeholders. We recognise that improving port services is a collective effort that requires collaboration from various parties. Therefore, we propose a comprehensive engagement strategy, starting with public consultations to gather input from local communities. A stakeholder forum will be established to facilitate discussion and collaboration in decision-making. Socialisation and education will be conducted to ensure transparency in information regarding the improvement plan. Partnerships with marine transport operators and the private sector will be built to support implementation. Finally, a feedback mechanism will be established to receive input and complaints from service users. By actively engaging stakeholders, we are confident that the port improvement plan will meet the needs and aspirations of the community, as well as gain broad support from all parties, thereby realising the sustainable improvement of marine transport services in the Thousand Islands.

6.4 Implications of cross-validation for generalisation of results

The application of cross-validation with the data-splitting method provides additional confidence in generalising the results of this study. The consistency of the observed cluster patterns between the training set and the validation set indicates that the developed K-means model has good stability. Despite the natural variation in maritime transport data, port clustering based on passenger utilisation patterns tends to be stable and less affected by small variations in the dataset.

This finding is important as it suggests that policy recommendations based on these clustering results have the potential to be applied more broadly and not just limited to the historical dataset used in the study. The stability of the clustering results provides a stronger basis for the DKI Jakarta local government to formulate more effective and efficient port service development strategies, taking into account the specific characteristics and needs of each identified port group.

6.5 Improving the overall tourist attractiveness of the Thousand Islands

Improved port services at Untung Jawa, Pari, and Tidung/Payung will directly improve travellers' experiences from the beginning to the end of their journey. More convenient, safe, and efficient port facilities will ease accessibility to these islands, reduce unnecessary waiting times, and improve the overall comfort of sea travel. With better-managed ports, travellers will feel safer, well-served through clear information, and enjoy a smoother journey. These improvements significantly contribute to travellers' satisfaction, creating a more positive and memorable tourism experience in the Thousand Islands.

A positive traveller experience, which starts with good port services, will reinforce the positive image of the Thousand Islands as a superior tourist destination. This image will attract more tourists to visit, not only for its natural beauty but also for the quality of infrastructure and services that support an unforgettable tourism experience. Investing in port services is a strategic move for the future of Thousand Islands tourism, as a modern and well-managed port is the main gateway that creates a positive first and last impression for every visitor, making the Thousand Islands an example of an island tourism destination that offers high standards of service.

6.6 Limitation

6.6.1 Data limitations and impacts

In this study, we acknowledge data limitations that affect the depth of analysis. While data on ship passengers and ship departures from the Jakarta Data Portal is reliable, more detailed information on port infrastructure, such as quay capacity and terminal facilities, is not available. This limitation hinders our ability to fully understand the factors affecting marine transport services in the Thousand Islands. In addition, the data we used covers the period 2018-2021, which may not reflect current conditions, especially after the post-pandemic COVID-19 tourist surge in 2022. For future research, more comprehensive and up-to-date data are needed to provide a more accurate picture.

6.6.2 Limitations of K-means algorithm and its implications

We also recognise the limitations of the K-means algorithm used in this study. Although K-means are efficient and easy to implement, they are sensitive to initial centroid initialisation and can potentially get stuck in a local optimum. The assumption of spherical or convex clusters may also not always match the complexity of the data patterns. In determining the optimal number of clusters, we used the elbow method, which has limitations of subjectivity in the interpretation of ‘elbow’ points. For stronger validation, future research may consider other methods, such as silhouette scores or gap statistics.

6.6.3 Other limitations and future research directions

In addition to data and algorithm limitations, this research also has limitations in the scope of analysis. We only focus on passenger and departure data, while other factors, such as weather conditions, ship conditions, and socioeconomic factors, are not considered. For future research, we suggest expanding the scope of data and analysis methods. We will have a more profound understanding of the sea transportation services in the Thousand Islands if we use more complete data, different clustering algorithms, and take outside factors into account. Recognising these limitations, we hope to make better contributions and encourage more comprehensive follow-up research.

7. CONCLUSION

This research, by applying the K-means method to historical ship passenger data, has successfully identified usage patterns and clustered ports in the Thousand Islands based on their operational capacity. The clustering analysis shows that the ports of Untung Jawa, Pari, and Tidung/Payung stand out as ports in need of service improvement. This clustering highlights that these ports, despite serving significant passenger volumes of around 250,000 to 300,000 passengers each during the study period, are operating at capacities that need to be optimised. Targeted service improvements at these ports have the potential to have a measurable impact on various aspects. To illustrate, improvements in operational efficiency and port facilities are expected to reduce average inter-island journey times by up to 15 minutes per trip,

significantly improving efficiency for commuters and travellers. In addition, improved service quality, including better terminal facilities and reduced crowding, is projected to increase passenger satisfaction levels by at least 20%, based on a hypothetical passenger satisfaction survey. Furthermore, with better sea transport services, there is potential to increase tourism revenue in these islands by up to 10% per year through improved accessibility and a more positive traveller experience. These improvements will provide economic benefits and support the achievement of the Sustainable Development Goals (SDGs), specifically SDG 8 (Decent Work and Economic Growth), SDG 10 (Reduction of Inequality), SDG 11 (Sustainable Cities and Settlements), and SDG 14 (Life Below Water).

8. FUTURE RESEARCH DIRECTIONS

For future research, several more specific and detailed directions can be explored to deepen the understanding and improve the applicability of the findings of this study. First, future research should integrate more comprehensive port infrastructure data, including detailed information on berth capacity, terminal facilities, and ship traffic flows. This additional data will allow for a more holistic and accurate analysis of the factors affecting maritime transport services. Second, the use of more up-to-date data covering the post-COVID-19 pandemic period is essential to capture recent changes in travel patterns and tourist volumes, especially the tourist surge in 2022. Third, future research could expand the scope of the analysis by including external factors such as weather conditions, ship conditions, and passenger socio-economic factors to develop a more comprehensive model. Finally, future research could focus on modelling and predicting the impact of various port improvement strategies (e.g., infrastructure upgrades, operational changes) on key performance indicators such as travel time, passenger satisfaction, and tourism revenue, possibly using simulation techniques or regression analysis to project policy impacts more accurately.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to all those who contributed to this research. We are especially grateful to Sriwijaya University and the MARA Technological University. Their support and guidance were invaluable.

REFERENCES

[1] Farda, M., Lubis, H.A.R. (2018). Transportation system development and challenge in Jakarta Metropolitan Area, Indonesia. International Journal of Sustainable Transportation Technology, 1(2): 42-50. https://doi.org/10.31427/ijstt.2018.1.2.2

[2] Baum, G., Kusumanti, I., Breckwoldt, A., Ferse, S.C., Glaser, M., Adrianto, L., van der Wulp, S., Kunzmann, A. (2016). Under pressure: Investigating marine resource-based livelihoods in Jakarta Bay and the Thousand Islands. Marine Pollution Bulletin, 110(2): 778-789. https://doi.org/10.1016/j.marpolbul.2016.05.032

[3] Anugerah, A.R., Muttaqin, P.S., Purnama, D.A. (2021). Effect of large-scale social restriction (PSBB) during COVID-19 on outdoor air quality: Evidence from five cities in DKI Jakarta Province, Indonesia. Environmental Research, 197: 111164. https://doi.org/10.1016/j.envres.2021.111164

[4] El-Refaei, A., Idris, A.O. (2025). Towards a port demand management (PDM) system: An analytic hierarchy process (AHP)-based approach. Case Studies on Transport Policy, 19: 101361. https://doi.org/10.1016/j.cstp.2024.101361

[5] Chen, J., Li, T., Zhao, H. (2025). Evaluating service upgrade in ports: The implication of consumer experience in the new maritime silk road. Ocean & Coastal Management, 261: 107533. https://doi.org/10.1016/j.ocecoaman.2024.107533

[6] Hou, Z., Yan, R., Wang, S. (2022). On the K-means clustering model for performance enhancement of port state control. Journal of Marine Science and Engineering, 10(11): 1608. https://doi.org/10.3390/jmse10111608

[7] Choe, C.W., Lim, S., Kim, D.J., Park, H.C. (2025). Development of spatial clustering method and probabilistic prediction model for maritime accidents. Applied Ocean Research, 154: 104317. https://doi.org/10.1016/j.apor.2024.104317

[8] Islam, M.M., Shamsuddoha, M.D. (2018). Coastal and marine conservation strategy for Bangladesh in the context of achieving blue growth and sustainable development goals (SDGs). Environmental Science & Policy, 87: 45-54. https://doi.org/10.1016/j.envsci.2018.05.014

[9] Sturesson, A., Weitz, N., Persson, Å. (2018). SDG 14: Life below water–A review of research needs. Technical Annex to The Formas report Forskning för Agenda, 2030. Stockholm Environment Institute, Stockholm.

[10] Kreinin, H., Aigner, E. (2022). From "Decent work and economic growth" to "Sustainable work and economic degrowth": A new framework for SDG 8. Empirica, 49(2): 281-311. https://doi.org/10.1007/s10663-021-09526-5

[11] Lane, J.M., Pretes, M. (2020). Maritime dependency and economic prosperity: Why access to oceanic trade matters. Marine Policy, 121: 104180. https://doi.org/10.1016/j.marpol.2020.104180

[12] Takase, C. (2018). Implementing SDG 11-Key elements, challenges and opportunities. In Module 4: SDGs-Sustainable Cities and Communities 2018 Executive Training Course for Policymakers on the 2030 Agenda and the Sustainable Development Goals (SDGs), Central Park Hotel Songdo, Incheon, Republic of Korea.

[13] Adly, A.P.F., Priyanto, P. (2019). Marine tourism marketing strategy Thousand Islands. Journal of Indonesia Tourism and Policy Studies, 4(2): 8-16. https://doi.org/10.7454/jitps.v4i2.86

[14] Open Data Jakarta. (2020). Ship passenger data to and from thousand islands (Data Penumpang Kapal dari dan ke Kepulauan Seribu). https://data.jakarta.go.id, accessed on Jun. 23, 2023.

[15] Biro Komunikasi dan Informasi Publik. (2012). ADPEL sunda kelapa facilitates residents' boats docking at muara angke terminal (ADPEL sunda kelapa fasilitasi kapal warga sandar di terminal muara angke). Direktorat Jenderal Perhubungan Laut. View 23 February 2025.

https://dephub.go.id/post/read/adpel-sunda-kelapa-fasilitasi-kapal-warga-sandar-di-terminal-muara-angke-9985.
[16] Naufal, M., Carina, J. (2022). New Muara Angke passenger terminal inaugurated, how is it different from Ancol pier? https://megapolitan.kompas.com/read/2022/10/04/10541101/terminal.
[17] Taherdoost, H. (2021). Data collection methods and tools for research; A step-by-step guide to choose data collection technique for academic and business research projects. International Journal of Academic Research in Management (IJARM), 10(1): 10-38. https://hal.science/hal-03741847v1.
[18] Zhang, X., Mahadevan, S. (2020). Bayesian neural networks for flight trajectory prediction and safety assessment. Decision Support Systems, 131: 113246. https://doi.org/10.1016/j.dss.2020.113246
[19] Velliangiri, S., Alagumuthukrishnan, S.J.P.C.S. (2019). A review of dimensionality reduction techniques for efficient computation. Procedia Computer Science, 165: 104-111. https://doi.org/10.1016/j.procs.2020.01.079
[20] Passarella, R., Iqbal, M.D., Buchari, M.A., Veny, H. (2023). Analysis of commercial airplane accidents worldwide using K-means clustering. International Journal of Safety & Security Engineering, 13(5): 813-819. https://doi.org/10.18280/ijsse.130505
[21] Tarekegn, A.N., Michalak, K., Giacobini, M. (2020). Cross-validation approach to evaluate clustering algorithms: An experimental study using multi-label datasets. SN Computer Science, 1: 1-9. https://doi.org/10.1007/s42979-020-00283-z
[22] Kim, Y., Kim, Y. (2023). Global regionalization of heat environment quality perception based on K-means clustering and Google trends data. Sustainable Cities and Society, 96: 104710. https://doi.org/10.1016/j.scs.2023.104710
[23] Ali, S., Ali, T.E., Khan, M.A., Khan, I., Patterson, M. (2021). Effective and scalable clustering of SARS-CoV-2 sequences. In Proceedings of The 5th International Conference on Big Data Research. Association for Computing Machinery, pp. 42-49. https://doi.org/10.1145/3505745.3505752
[24] Rajee, A.M., Francis, F.S. (2013). A study on outlier distance and SSE with multidimensional datasets in K-means clustering. In 2013 Fifth International Conference on Advanced Computing (ICoAC), Chennai, India, pp. 33-36. https://doi.org/10.1109/icoac.2013.6921923