Journal of Soft Computing Exploration
Homepage: htttps://shmpublisher.com/index.php/joscex
p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

# Automatic identification system big data-driven maritime traffic density prediction in surabaya port using PCA and k-means clustering

**Afif Zuhri Arfianto[1*], Muhammad Izzul Haj[2], Muhammad Khoirul Hasin[3], Noorman Rinanto[4], Imam Sutrisno[5], Dimas Pristovani Riananda[6], Dwi Sasmita Aji Pambudi[7]**
[1,3,4,5,6,7]Department of Marine Electrical Engineering, Politeknik Perkapalan Negeri Surabaya, Indonesia
[2]Department of Mechanical Engineering, Chung Yuan Christian University, Taiwan

**Article Info**

***Article history:***

Received Mar 16, 2026
Revised Mar 26, 2026
Accepted Apr 07, 2026

***Keywords:***

AIS big data
Maritime traffic density
K-means clustering
Principal component analysis
Surabaya port

**ABSTRACT**

The management of maritime traffic directly determines the level of operational efficiency and safety achievable at major ports, including Tanjung Perak in Surabaya, which serves as a critical logistics node for eastern Indonesia. This study presents a comprehensive analysis of maritime traffic density prediction using Automatic Identification System (AIS) big data combined with Principal Component Analysis (PCA) and K-Means clustering techniques. The dataset comprises 1,173 vessel movements recorded in December 2025, encompassing various vessel types, port operations, and voyage characteristics. Through dimensionality reduction using PCA and unsupervised clustering with K-Means, we identified 10 distinct traffic patterns representing different operational profiles. The analysis revealed significant temporal patterns, with peak traffic occurring at 14:00 (79 vessels) and lowest traffic at 02:00 (18 vessels). The clustering results achieved a silhouette score of 0.3863, effectively segmenting vessels based on voyage distance, capacity, speed, draught, and temporal features. The results of this research offer practical guidance for port authorities seeking to improve resource allocation, traffic management, and operational efficiency based on empirical evidence.



***Corresponding Author:***

Afif Zuhri Arfianto
Department of Marine Electrical Engineering,
Politeknik Perkapalan Negeri Surabaya
Jalan Teknik Kimia, Kampus ITS Sukolilo, Surabaya.
Email: afifzuhri@ieee.org
https://doi.org/10.52465/joscex.v7i1.22

## 1. INTRODUCTION

Maritime transportation plays a vital role in global trade, with seaports acting as critical nodes in the supply chain network [1], [2]. Surabaya Port (Tanjung Perak), located in East Java, Indonesia, ranks among the busiest ports in Southeast Asia, handling diverse cargo types and vessel classes amid steadily rising traffic volumes. As one of Indonesia's highest volume ports, Tanjung Perak consistently handles substantial cargo throughput on an annual basis. Port congestion in such high traffic environments tends to extend vessel waiting

times significantly, generating additional demurrage costs and disrupting broader supply chain operations. Between 2020 and 2024, Indonesia recorded a significant rise in near miss incidents and vessel waiting times at major ports, with economic losses from port congestion estimated in the hundreds of billions of rupiah per year [1]. This surge in traffic volume, combined with increasingly complex logistics chains, necessitates advanced analytical approaches to predict traffic density patterns, optimize port operations, and bolster safety measures before congestion or accidents escalate into irreversible disruptions [3].

Through the periodic transmission of vessel data, including identity, position, speed, course, and voyage particulars, the Automatic Identification System (AIS) facilitates situational awareness for both neighboring vessels and shore based maritime authorities [4]. For instance, a single AIS transponder transmits position updates every 2–10 seconds, generating hundreds of thousands of data points per vessel per day [5]. "The sheer volume of AIS data generated monthly at a major port such as Tanjung Perak far exceeds what can be processed through conventional manual approaches, necessitating the use of automated computational methods to conduct meaningful traffic analysis [6]. These datasets, growing into massive volumes due to continuous transmissions, demand sophisticated analytical techniques to yield actionable insights [7], [8]. AIS data has proven increasingly significant in enhancing maritime domain awareness through comprehensive and continuous vessel monitoring [9]. Prior research has leveraged AIS for diverse applications, including route optimization, anomaly detection, and traffic flow analysis [4], [10]. However, the big data characteristics of AIS, which encompass high volume, velocity, and variety, necessitate advanced computational methods to extract meaningful patterns that support informed operational decisions [11].

Several studies have addressed maritime traffic analysis using AIS data and machine learning techniques, providing the primary reference baseline for this work. Li et al applied spatiotemporal graph neural networks across multiple Chinese ports and demonstrated that incorporating both spatial adjacency and temporal dependencies significantly improves vessel traffic flow prediction accuracy [2]. A key limitation of this approach, however, lies in its reliance on dense inter port connectivity data, a requirement that is difficult to fulfill in port settings where data infrastructure remains underdeveloped . Liu et al. combined trajectory clustering with density based methods in a Chinese port area, demonstrating that hybrid clustering models can effectively distinguish different vessel behavior patterns. However, the study was restricted to a single vessel type and did not extend to multi type operational profiling [6], [12]. Although Zhang et al. successfully constructed a dynamic AIS-based traffic pattern recognition framework by combining K-Means clustering with online data cleaning and partitioning to yield interpretable traffic segments, their method did not incorporate dimensionality reduction, a shortcoming that reduces its effectiveness on high dimensional feature sets [13]. More recently, Khodamoradi et al. proposed a federated learning approach for vessel traffic density prediction that addresses data privacy concerns across distributed port environments [14]. Despite these advances, the existing literature has not addressed the integration of PCA-based dimensionality reduction and K-Means clustering for AIS-driven traffic density prediction in Indonesian port settings, where vessel composition and inter island ferry operations present characteristics that are markedly distinct from East Asian port environments. This study is specifically designed to address that unresolved limitation.

Among analytical methods, Principal Component Analysis (PCA) stands out as a widely adopted dimensionality reduction technique, enabling high dimensional datasets to be analyzed while preserving most of the original variance [15]. In maritime contexts, PCA excels at pinpointing the most influential factors shaping vessel movements and traffic dynamics, transforming correlated variables into a compact set of uncorrelated principal components [15], [16]. This process not only simplifies data interpretation but also enhances visualization of intricate maritime systems, laying the groundwork for subsequent analyses [11], [17]. Beyond traffic analysis, PCA has been applied in various maritime machine learning tasks, including vessel classification and engine model selection, demonstrating its versatility as a feature extraction and dimensionality reduction tool in the maritime domain [18], [19].

Complementing PCA, K-Means clustering has proven highly effective in maritime traffic analysis by identifying vessel behavior patterns, grouping navigation routes, and detecting behavioral anomalies [20]. Such anomalies encompass sudden AIS transponder deactivation in restricted zones, abnormally high vessel speeds in congestion areas, significant deviations from declared voyage plans, and unjustified loitering near restricted anchorages, each of which poses potential safety risks or regulatory compliance concerns [21]. K-Means' computational efficiency and intuitive interpretability render it ideal for processing large scale AIS data [22], [23]. When integrated with PCA, it partitions traffic data into distinct operational categories, revealing deeper insights into port activities, navigational trends, and congestion hotspots [24].

This study pursues several interconnected objectives to address these challenges at Surabaya Port. It first analyzes AIS big data to elucidate traffic patterns and density distributions. Building on this, it employs PCA for dimensionality reduction and feature extraction from the high dimensional dataset then deploys K-Means clustering to delineate unique traffic segments grounded in operational traits. Subsequently, it constructs predictive models for traffic density incorporating both temporal and operational variables, and ultimately delivers practical, data backed recommendations to aid port traffic management decisions.

This study pursues several interconnected objectives to address these challenges at Surabaya Port. It first analyzes AIS big data to elucidate traffic patterns and density distributions. Building on this, it employs PCA for dimensionality reduction and feature extraction from the high dimensional dataset then deploys K-Means clustering to delineate unique traffic segments grounded in operational traits. Subsequently, it constructs predictive models for traffic density incorporating both temporal and operational variables, and ultimately delivers practical, data backed recommendations to aid port traffic management decisions.

The study's significance stems from its targeted illumination of traffic density patterns at Surabaya Port, fostering multifaceted benefits across operational, safety, environmental, and economic dimensions. Operationally, the findings enable optimized berth allocation and resource scheduling for greater efficiency. From a safety standpoint, precise traffic pattern recognition helps mitigate congestion and reduce collision risks. Environmentally, refined traffic management curtails vessel idling and associated emissions. Economically, enhanced vessel flows promise higher throughput and reduced operating costs, positioning the port for sustainable long term growth. Recent studies on Tanjung Perak's Ro-Ro terminal have revealed operational inefficiencies linked to queuing and congestion, further underscoring the urgency of data driven traffic management approaches.

## 2. METHOD

This study presents a systematic methodological framework comprising seven sequential stages: (1) Data Collection and Description, which outlines the sources and characteristics of the AIS dataset employed; (2) Research Framework, which describes the overall design and workflow of the proposed approach; (3) Data Preprocessing, which addresses data cleaning, filtering, and normalization procedures to ensure data quality; (4) Feature Preparation for PCA, which involves the selection and transformation of relevant variables prior to dimensionality reduction; (5) PCA Implementation, which applies Principal Component Analysis to extract the most significant features from the dataset; (6) K-Means Clustering, Optimal Cluster Selection, which determines the appropriate number of clusters through established optimization criteria; and (7) Traffic Density Classification, which categorizes vessel traffic patterns based on the resulting cluster assignments.

### A) Data Collection and Description

The dataset consists of 1,173 vessel movement records from Surabaya Port collected in December 2025 as shown at Table 1. It contains 34 attributes that cover vessel information, temporal and operational variables, voyage characteristics, and port infrastructure details. Vessel information includes the vessel's name, MMSI, IMO, vessel type, flag, capacity (DWT), and draught. Temporal data capture actual time of arrival and departure (ATA/ATD), time spent in port, and voyage time. Operational attributes describe the port call type, load condition, and type of port operation performed. Voyage characteristics include origin and destination ports, distance traveled, average and maximum speed, and idle time during the voyage. Infrastructure related attributes specify the berth name and terminal name used during each port call.

Table 1. Dataset statistics

| Metric | Value |
|---|---|
| Total Records | 1,173 |
| Arrival Records | 656 (55.9%) |
| Departure Records | 517 (44.1%) |
| Unique Vessels | 847 |
| Date Range | December 1-31, 2025 |
| Vessel Types | 5 Categories |

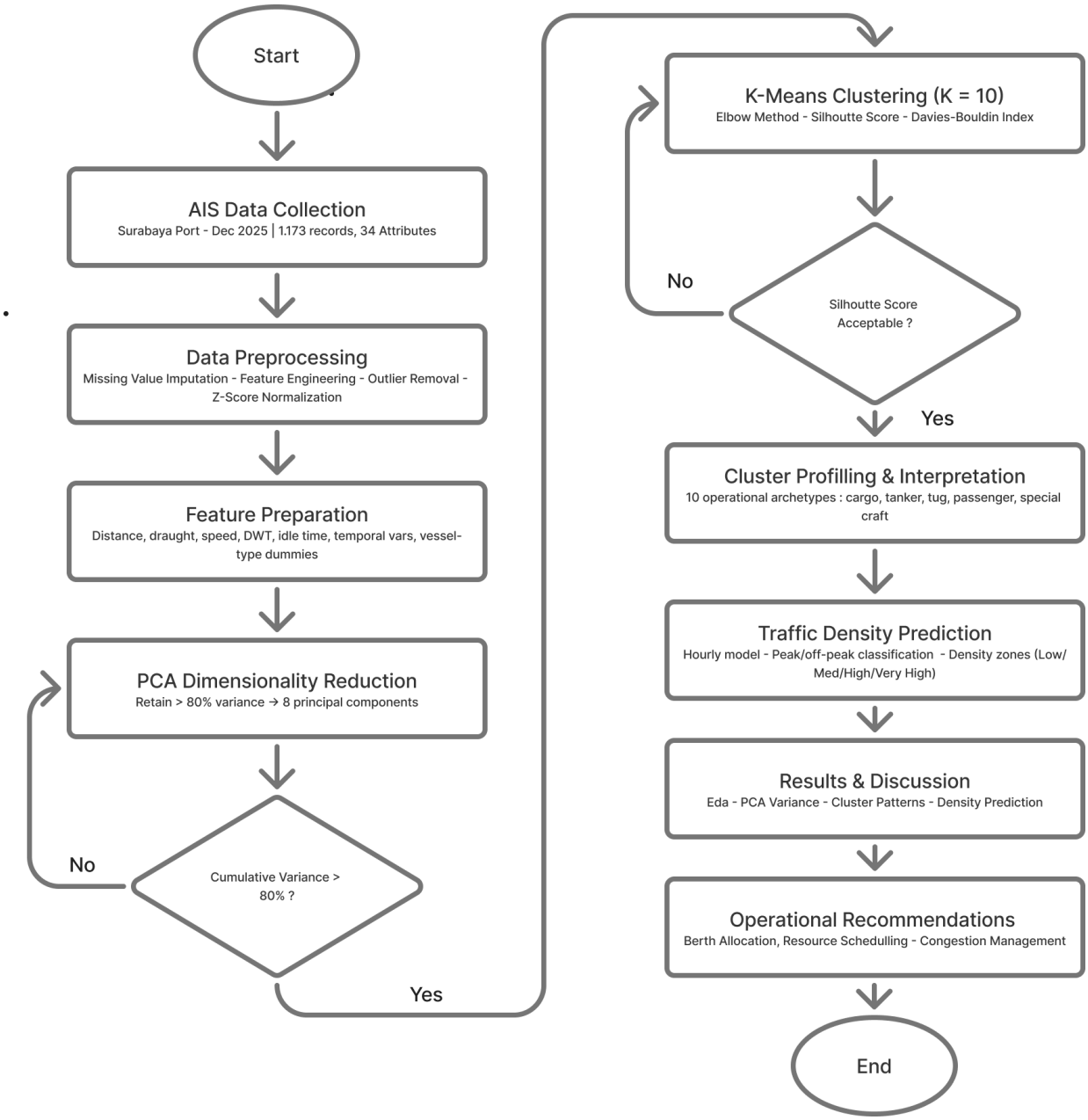


Figure 1. Research flowchart

B) Research Framework

Figure 1 presents the research flowchart illustrating the complete analytical pipeline adopted in this study. The process begins with raw AIS data collection from Surabaya Port, followed by a systematic data preprocessing stage that includes missing value imputation, feature engineering, outlier treatment, and z-score normalization. The preprocessed data then undergoes PCA-based dimensionality reduction to extract the most informative principal components that capture the dominant structure of vessel movement patterns. The reduced feature space is subsequently fed into K-Means clustering to identify distinct traffic segments with coherent operational profiles. Finally, the cluster profiles are used to construct traffic density prediction models and derive data driven operational recommendations for port management. This end-to-end pipeline ensures that each analytical step is grounded in the output of the preceding stage, maintaining methodological coherence throughout the study.

C) Data Preprocessing

The preprocessing pipeline for this study consisted of several sequential steps designed to prepare the vessel movement data for Principal Component Analysis (PCA). Missing values in key operational features such as capacity, speed, and idle time were first handled through median imputation to preserve the central tendency of the data while reducing the influence of extreme values. Next, feature engineering was applied by

extracting temporal features (hour, day of week, and day of month), generating a binary indicator for port call type (arrival or departure), and applying one-hot encoding to the five most frequent vessel types to better capture categorical variation in the dataset. Outlier detection procedures were then used to identify and treat anomalous values in distance and speed, ensuring that extreme observations did not distort subsequent analysis. Finally, all continuous variables used as input to PCA were standardized using z-score normalization so that each feature had zero mean and unit variance, allowing all variables to contribute comparably to the resulting principal components.

D) Feature Preparation for PCA

The feature preparation step determined which variables would be input to PCA based on their relevance to capturing both operational intensity and temporal patterns of vessel movements. It is important to clarify that PCA is a dimensionality reduction and feature extraction technique — not a feature selection method. Unlike feature selection, which ranks and discards original variables, PCA mathematically transforms the entire input space into a new set of uncorrelated axes (principal components) that capture the maximum variance in the data. No original feature is eliminated; rather, information from all features is redistributed across the components according to its contribution to overall variance. Fifteen features were retained as inputs, including voyage distance travelled, draught, hour of day, day of week, and day of month, alongside vessel capacity, average voyage speed, voyage idle time, and categorical encodings of vessel type and port call type. These variables collectively represent the multidimensional operational profile of each vessel visit, providing a rich basis for the subsequent PCA transformation

E) PCA Implementation

The PCA transformation was applied to the standardized 15-dimensional feature space so that each variable contributed on a comparable scale to the resulting components. The analysis was designed to retain at least 80% of the total variance in the dataset. In practice, this resulted in a reduced representation of 8 principal components that together explained 80.9% of the variance, with the first two components accounting for 33.3% and the first three components for 44.5% of the cumulative variance, while a total of 12 components would be required to capture 95% of the variance. Each principal component is a linear combination of all original standardized features, ordered by the amount of variance explained. The first few components capture the dominant sources of variation in vessel movement patterns, while subsequent components account for progressively finer nuances. This lower-dimensional representation substantially reduces computational complexity while preserving the core structure of the data for downstream clustering.

F) K-Means Clustering, Optimal Cluster Selection

Optimal cluster selection in this study relied on three complementary internal validation techniques to balance model simplicity with clustering quality. First, the Elbow Method was used by plotting the within-cluster sum of squares (inertia) against different values of k to identify the point where additional clusters yield diminishing returns in variance reduction. While the Elbow Method remains commonly used, recent studies have highlighted its limitations and advocated for more rigorous alternatives such as the silhouette score and the Davies-Bouldin index [27]. Second, Silhouette Analysis measured how well-separated and cohesive the clusters were, providing insight into whether data points were appropriately assigned to their nearest cluster. Finally, the Davies-Bouldin Index was computed to quantify cluster compactness and separation, where lower index values indicate tighter, more distinctly separated clusters. Based on the combined analysis, K=10 was selected as the optimal number of clusters, achieving the highest silhouette score of 0.3863, as shown at Table 2.

Table 2. K-Means clustering parameter

| K | Inertia | Silhouette Score |
|---|---|---|
| 2 | 11,690.06 | 0.2088 |
| 3 | 10,095.25 | 0.2287 |
| 4 | 8,922.25 | 0.2597 |
| 5 | 7,830.70 | 0.2951 |
| 6 | 6,560.67 | 0.3535 |
| 7 | 5,465.03 | 0.3583 |
| 8 | 4,807.32 | 0.3738 |
| 9 | 4,307.68 | 0.3731 |
| 10 | 3,957.60 | 0.3863 |

G) Traffic Density Classification

Traffic density levels in this study were categorized using hourly vessel counts to distinguish different intensities of port activity. Low density corresponds to fewer than 10 vessels per hour, medium density to 10-30 vessels per hour, high density to 30-50 vessels per hour, and very high density to more than 50 vessels per hour. These thresholds were defined based on the operational capacity benchmarks of Surabaya Port's traffic management system and are used consistently throughout the analysis to characterize temporal traffic patterns.

## 3. RESULTS AND DISCUSSIONS

A) Exploratory Data Analysis — Traffic Distribution Patterns

The analysis of the vessel movement data showed clear patterns in port operations across port call types, vessel categories, and time. Arrivals slightly outnumbered departures, with 656 arriving vessels (55.9%) compared to 517 departing vessels (44.1%), indicating a marginally higher inbound traffic intensity over the study period. Cargo vessels dominated the traffic mix with 478 calls (40.8%), followed by passenger vessels with 279 calls (23.8%), tug boats with 196 calls (16.7%), tankers with 103 calls (8.8%), and special craft with 61 calls (5.2%), highlighting the predominance of cargo operations but also the significant roles of passenger services and port-support vessels in overall port activity. Figure 2 visualizes port call types and vessel type distributions. The bar chart on the left panel shows the arrival-departure balance, while the right panel presents the breakdown by vessel category, illustrating the dominance of cargo vessels and the substantial supporting role of tug boats and passenger ferries.

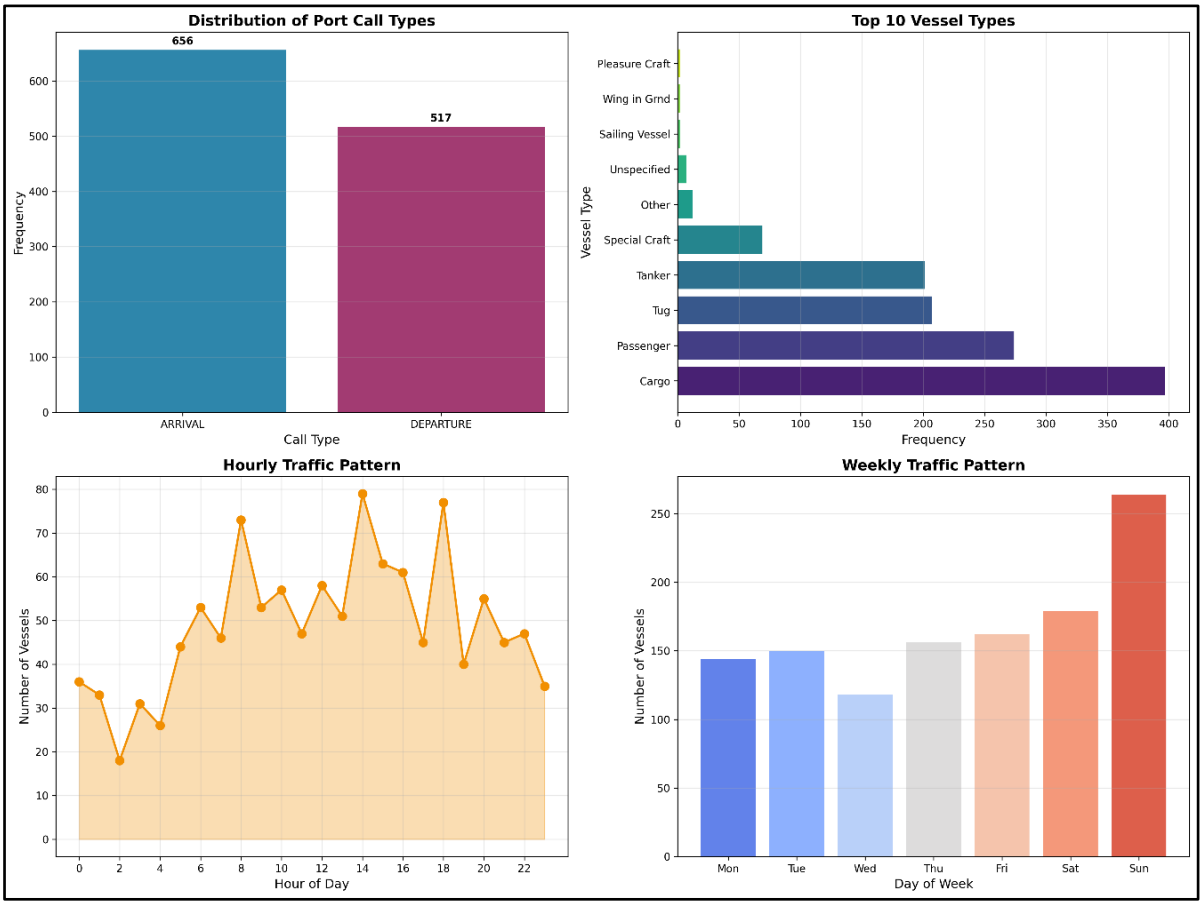


Figure 2. Traffic distribution by call type and vessel type

B) Temporal Traffic Patterns

The temporal analysis of vessel movements at Surabaya Port reveals pronounced peaks and consistently high utilization throughout the day. Peak traffic occurs at 14:00 (79 vessels), 18:00 (77 vessels), and 08:00 (73 vessels), while the lowest activity is observed at 02:00 (18 vessels), 04:00 (26 vessels), and 03:00 (31 vessels), with an average hourly traffic of 48.9 vessels over the observation period.

Figure 2 presents the hourly traffic distribution as a line chart overlaid with density zone bands. The chart clearly shows the bimodal daily pattern with a primary afternoon peak and a secondary morning peak. On a weekly scale, traffic remains relatively stable across weekdays with only a slight reduction during weekends,

indicating that port operations are driven by continuous commercial demand rather than strongly weekday-bound cycles. When mapped to predefined density categories, 11 hours (45.8%) fall into the high density band and another 11 hours (45.8%) into the very high density band, while only 2 hours (8.3%) are classified as medium density and none as low density, confirming that Surabaya Port operates near consistently high traffic density levels throughout the study window.

C) Geographical and Operational Analysis

The origin-destination analysis highlights concentrated inter-port linkages centered on Surabaya. Kamal is the most frequent origin with 145 vessels, followed by Banjarmasin (58), Gresik (43), Patimban (35), and Balikpapan (32), reflecting strong short- and medium-range feeder connections into Surabaya. Surabaya River Anchor is the dominant destination with 187 vessels, ahead of Kamal (81), Gresik (38), Banjarmasin (29), and Lembar (25), indicating that anchorage and nearby regional ports form the primary sinks for outbound traffic.

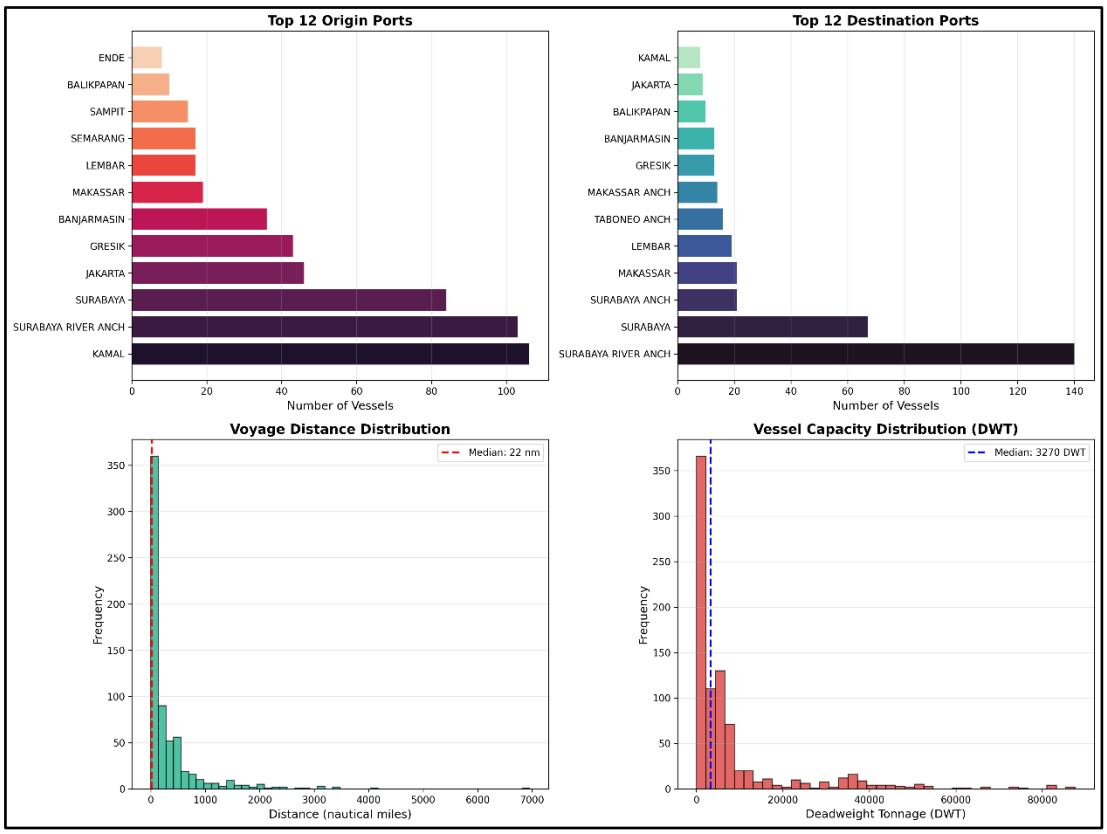


Figure 3. Geographical and operational analysis

Figure 3 maps the geographical distribution of origin-destination flows, presenting the main corridors feeding and served by Surabaya Port as a chord diagram or flow map. The concentrated connectivity pattern underscores the port's role as a regional hub within the Java-Kalimantan-Nusa Tenggara shipping network. The operational characteristics confirm a strong concentration of short-range movements with a smaller number of long-distance voyages. Voyage distance has a median of 1 nautical mile and a mean of 174.2 nautical miles (range: 0-6,955 nm), revealing a highly skewed distribution dominated by short-distance operations. Vessel capacity shows a median of 3,270 DWT and a mean of 7,849 DWT (range: 150-39,829 DWT), suggesting a fleet composed mainly of small to medium-sized ships with a few larger vessels. Average voyage speed is 6.9 knots with a median of 5.7 knots, consistent with port approaches and coastal sailing. Draught values have a mean of 3.3 meters and a range from 0 to 25.5 meters, indicating predominantly shallow-draft vessels. Figure 4 presents the distribution of speed, draught, and capacity metrics through box plots and histograms, highlighting the contrast between typical short-range traffic and the less frequent deep-sea or high-capacity operations.

D) PCA Results — Variance Explanation

The PCA transformation achieved effective dimensionality reduction while retaining most of the variation present in the original feature set. The first principal component (PC1) explains 18.7% of the total variance, followed by PC2 with 14.6% and PC3 with 11.2%, indicating that these three axes already capture a substantial share of the dominant structure in the data. In total, the first 8 principal components account for 80.9% of the cumulative variance, showing that the original 15-dimensional feature space can be compactly represented in a much lower dimension with limited information loss.

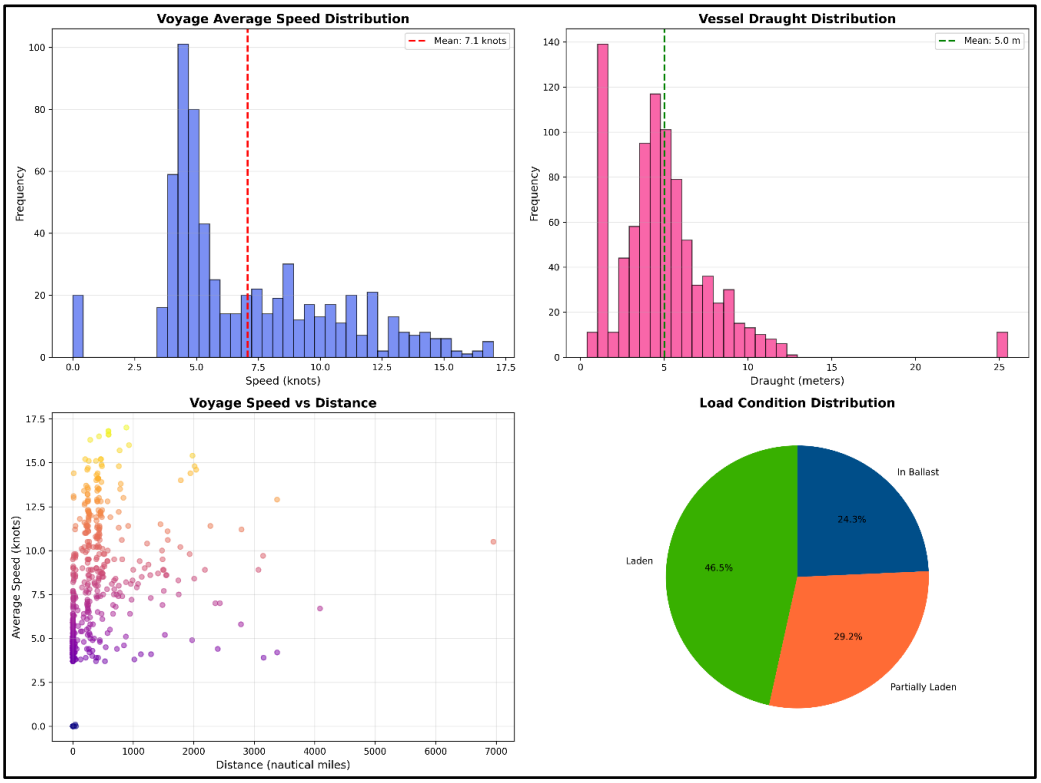


Figure 4. Speed and performance metrics

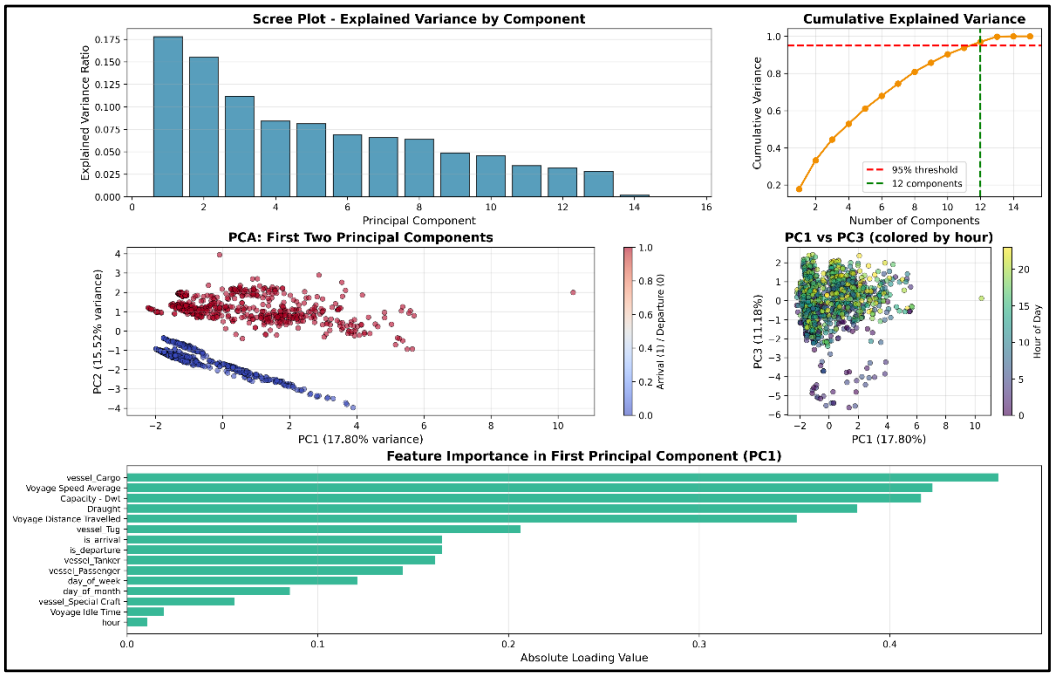


Figure 5. PCA analysis results

Figure 5 presents the scree plot and cumulative variance curve from the PCA analysis. The scree plot shows the explained variance ratio of each component as a bar chart, while the cumulative variance curve overlaid as a line clearly identifies the elbow point at which additional components contribute diminishing returns. The visualization confirms that 8 components represent a well-justified truncation point balancing dimensionality reduction with information preservation.

E) PCA Feature Loadings Heatmap

Figure 6 presents the PCA loading matrix as a heatmap. Each cell in the heatmap represents the loading value — that is, the correlation coefficient between an original feature (displayed on the rows) and a principal component (displayed on the columns). Warm colors (red and orange) indicate strong positive loadings, meaning that the corresponding feature contributes substantially to that principal component in a positive direction, while cool colors (blue) represent negative loadings or weak contributions toward the corresponding component. The numerical values printed within each cell provide the exact loading coefficient, enabling precise interpretation of each feature's directional influence across components.

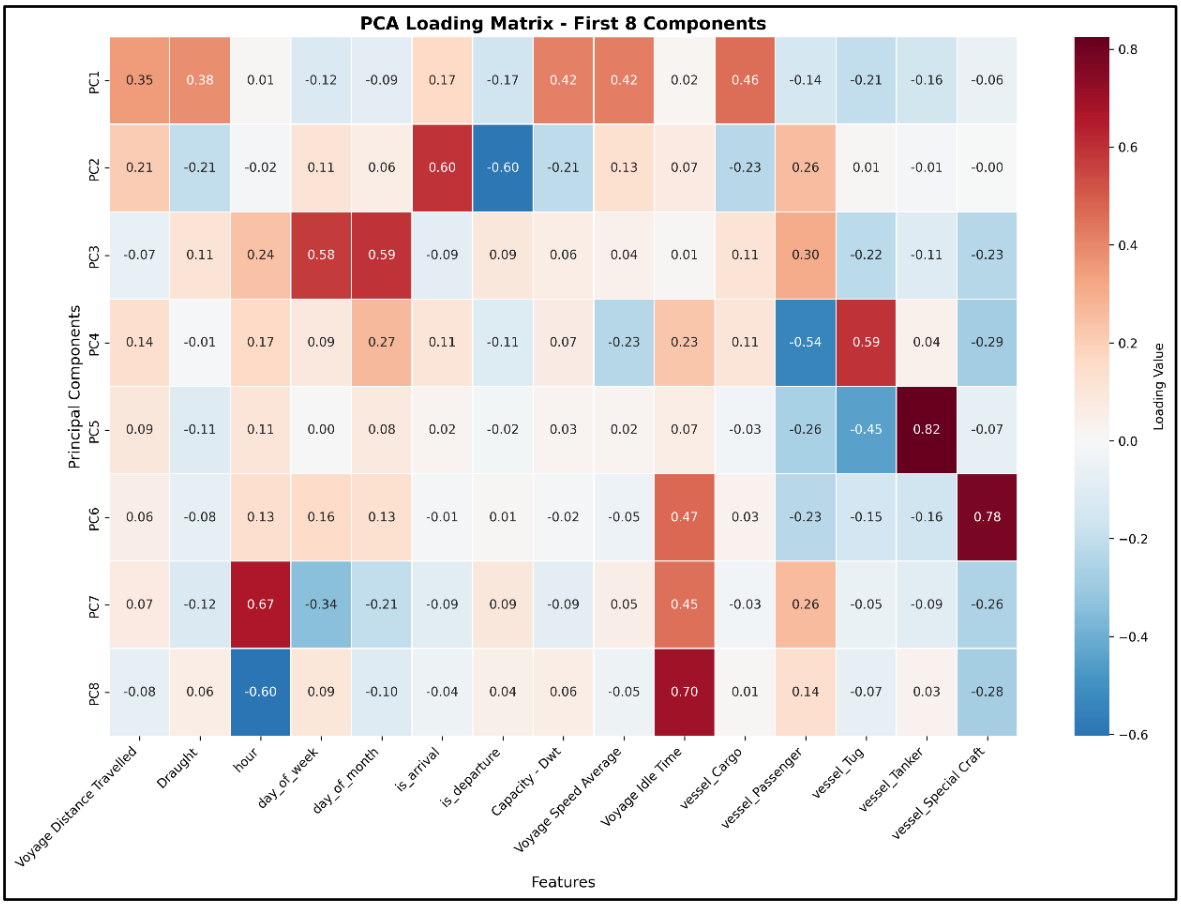


Figure 6. PCA loadings heatmap

PC1, which explains 18.7% of the variance, is characterized by high positive loadings for voyage distance travelled and moderate positive loadings for vessel capacity and average voyage speed, indicating that PC1 captures the scale and reach of vessel operations (long-range, large-capacity voyages). PC2, accounting for 14.6% of the variance, is mainly driven by temporal features (hour and day variables) and arrival/departure indicators, capturing systematic temporal patterns and operational characteristics. The heatmap was generated using seaborn's diverging color palette centered at zero, ensuring symmetric and perceptually balanced visual representation of positive and negative contributions so that the zero-loading boundary is immediately apparent to the reader.

F) Clustering Results

The K-Means algorithm applied to the PCA-reduced feature space identified 10 distinct clusters representing different operational profiles at Surabaya Port. The clustering achieved a silhouette score of 0.3863 and a Davies-Bouldin Index of 0.8674, indicating moderate but meaningful cluster separation given the inherent heterogeneity of maritime traffic data.

Table 3. K-Means cluster profiles

| Cluster | Size | % | Avg Distance (nm) | Avg Capacity (DWT) | Avg Speed (knots) | Avg Draught (m) | Dominant Type |
|---|---|---|---|---|---|---|---|
| 0 | 181 | 15.4% | 115.1 | 1,237 | 7.2 | 2.4 | Passenger |
| 1 | 197 | 16.8% | 0.0 | 14,505 | 5.7 | 6.6 | Cargo |
| 2 | 95 | 8.1% | 0.0 | 3,987 | 5.7 | 2.6 | Tanker |
| 3 | 38 | 3.2% | 253.3 | 6,995 | 6.7 | 3.3 | Cargo |
| 4 | 196 | 16.7% | 49.8 | 2,221 | 4.7 | 2.2 | Tug |
| 5 | 103 | 8.8% | 135.8 | 3,178 | 5.4 | 2.2 | Tanker |
| 6 | 61 | 5.2% | 69.5 | 3,063 | 6.2 | 2.7 | Special Craft |
| 7 | 203 | 17.3% | 717.0 | 14,717 | 9.5 | 5.9 | Cargo |
| 8 | 1 | 0.1% | 325.0 | 3,270 | 4.3 | 0.0 | Sailing Vessel |
| 9 | 98 | 8.4% | 0.0 | 2,467 | 5.7 | 3.5 | Passenger |

Figure 7 presents the cluster visualization projected onto the first two principal components. The scatter plot reveals spatial separation between the larger clusters while acknowledging some overlap in transitional traffic regimes. Cluster 7 (long-distance cargo) and Cluster 1 (port-based large cargo departures) are well separated from the tug and short-distance passenger clusters, reflecting their distinct operational profiles. Cluster 8 (a single sailing vessel) appears as an isolated outlier distant from all other clusters, confirming its anomalous nature relative to the commercial traffic patterns at Surabaya Port.

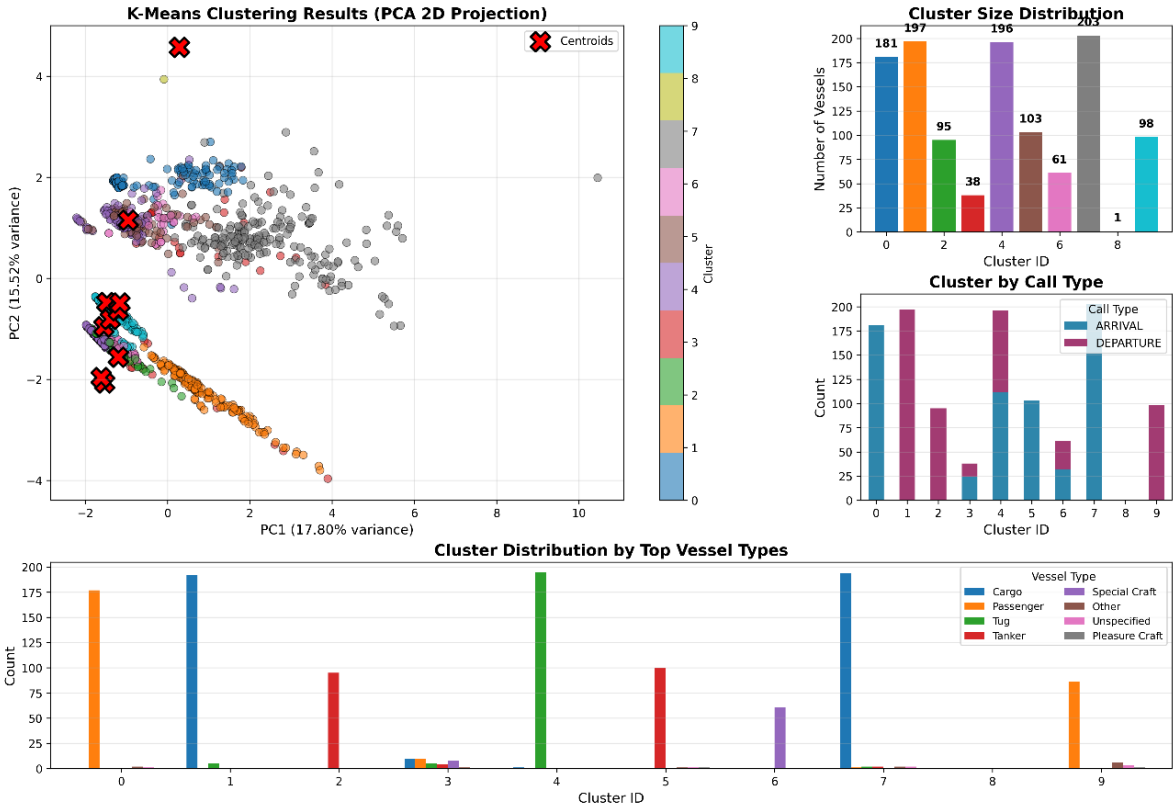


Figure 7. Cluster visualization

G) Cluster Interpretation

Figure 8 presents the cluster characteristics, Cluster 0 groups medium-distance passenger vessels (around 115 nautical miles on average) with consistent arrival patterns, characteristic of scheduled inter-island ferry services operating on the Java-Madura and Java-Lombok corridors. Cluster 9 contains medium-sized passenger

vessels engaged in port-based, all-departure operations, likely representing outbound ferry or passenger services starting from Surabaya.

Cluster 3 represents medium-distance cargo traffic with balanced arrivals and departures, medium routes of about 253 nautical miles, and moderate-capacity vessels, indicating regional freight connections to ports such as Banjarmasin and Balikpapan rather than purely local or deep-sea services. Cluster 1 captures large cargo vessels (about 14,505 DWT on average) that are predominantly departures with deep draughts (around 6.6 meters) and zero recorded voyage distance, suggesting fully loaded ships leaving after port-based loading operations. Cluster 7 corresponds to long-distance cargo services, with the highest average distance (about 717 nautical miles), the highest speeds (around 9.5 knots), large capacities (about 14,717 DWT), predominantly arrivals, and the second-largest cluster size (203 vessels), making it representative of international or long-haul cargo operations calling at Surabaya.

Cluster 2 groups medium-sized tankers that are predominantly departures with moderate draughts, indicating partially loaded outbound tanker movements. Cluster 5 represents arriving tankers on medium-distance routes (about 136 nautical miles) with consistent operational characteristics, highlighting inbound liquid bulk flows likely originating from Pertamina terminals along the Kalimantan coast. Cluster 4 consists of tug operations characterized by short distances (around 50 nautical miles), mixed arrivals and departures, lower speeds, and the largest membership (196 vessels), reflecting intensive support activities such as berthing assistance and local maneuvering. Cluster 6 aggregates special craft performing short- to medium-distance voyages with balanced arrivals and departures and diverse operational patterns, consistent with auxiliary or service vessels such as pilot boats, survey vessels, and dredging support craft. Cluster 8 is an outlier cluster containing a single sailing vessel with a unique operational profile that differs substantially from all other traffic patterns identified in the analysis.

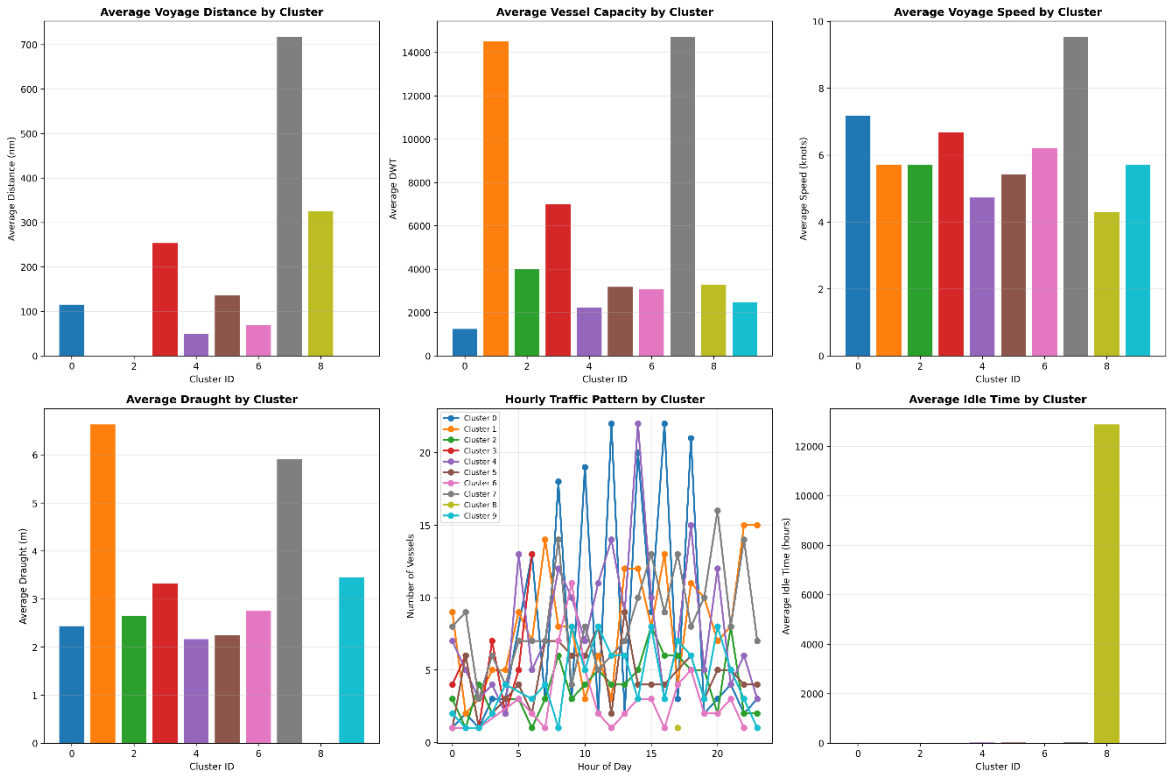


Figure 8. Cluster characteristics

H) Traffic Density Prediction

The traffic density prediction model, built from historical patterns and cluster-based analysis, identifies several peak periods with high congestion risk and a set of lower-traffic windows suitable for maintenance activities. Very high density is forecast at 14:00 (79 vessels), 18:00 (77 vessels), 08:00 (73 vessels), 15:00 (63 vessels), and 16:00 (61 vessels), indicating that these hours are the most critical for managing berthing, channel

access, and tug allocation. In contrast, the lowest predicted traffic occurs around 02:00 (18 vessels, medium density) and 04:00 (26 vessels, medium density), offering comparatively optimal windows for scheduling planned maintenance and other capacity-reducing activities without severely impacting operations.

Figure 9 visualizes hourly traffic levels as a stacked area chart showing how different traffic clusters jointly shape the daily vessel load profile. During the morning peak around 08:00, overall density is largely driven by Cluster 7 (international cargo arrivals). In the afternoon peak between 14:00 and 18:00, the chart shows a more mixed cluster composition with a strong presence of departure-oriented clusters, reflecting intense outbound cargo, tanker, and passenger movements. The density-level analysis confirms that approximately 92% of observed hours fall into High or Very High density zones (more than 30 vessels per hour), while only 8% fall into the Medium density band, and no hours record Low density. This distribution implies that the port is operating close to its effective capacity throughout most of the day.

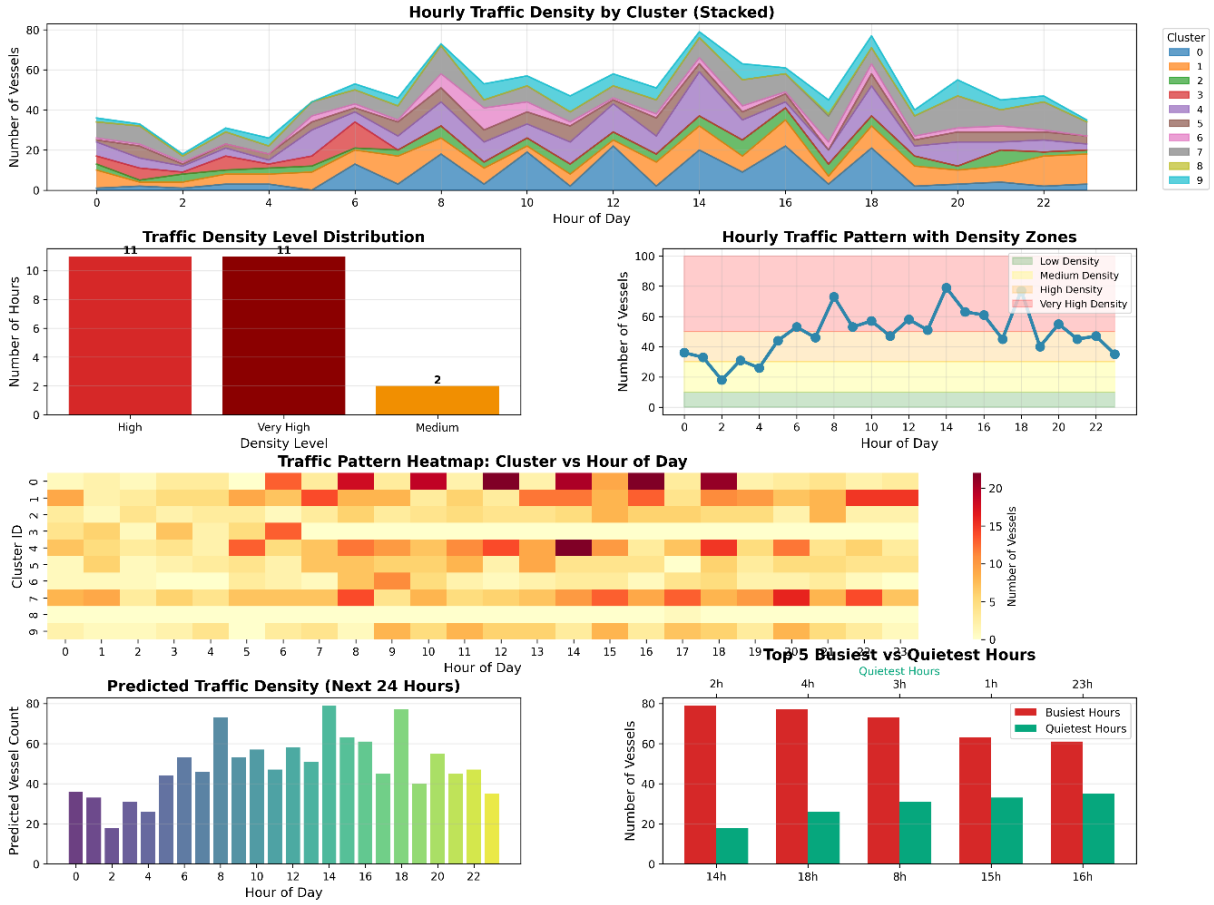


Figure 9. Traffic density prediction

## I) Discussion

*Interpretation of Traffic Patterns*

The 10 clusters identified in this study correspond to meaningfully distinct operational profiles rather than arbitrary statistical groupings. The separation between short-haul coastal vessels and deep-sea ocean-going vessels is particularly pronounced, reflecting the dual nature of Tanjung Perak as both a regional hub for inter-island traffic and an international gateway for larger ocean-going vessels. The strong temporal loading on the second principal component aligns with findings in prior port traffic studies, which have consistently identified time-of-day as a primary driver of traffic density variation [24], [25].

The peak traffic hour at 14:00 corresponds to the mid-afternoon window commonly observed at Indonesian ports, when vessels departing in the morning arrive at their first waypoints and coordinate inbound approaches with port traffic control. The low-traffic trough at 02:00 reflects the natural diurnal rhythm of commercial shipping, though the non-negligible traffic volume even at this hour underscores the continuous operational demand placed on Surabaya Port's resources and the need for round-the-clock traffic management capacity.

*PCA as Dimensionality Reduction, Not Feature Selection*

It is important to distinguish PCA's role in this study from conventional feature selection methods. Feature selection techniques — such as filter methods, wrapper methods, or embedded methods — rank and discard original variables to retain only those deemed most informative [28]. PCA, in contrast, is a dimensionality reduction and feature extraction method: it does not remove original features but mathematically transforms the entire input space into a new orthogonal basis defined by directions of maximum variance. Each principal component is a weighted linear combination of all original features, so no information is entirely discarded; rather, it is redistributed across components in order of importance. This distinction is critical for interpreting the PCA loadings heatmap as shown at Figure 6, since every original feature retains some representation across all components, even if its contribution to any single component is small. The practical implication is that PCA-based dimensionality reduction is more information-preserving than variable elimination, which justifies its selection as the preprocessing step for K-Means clustering in this study.

*Comparison with Existing Studies*

The silhouette score of 0.3863 achieved in this study is consistent with values reported in comparable maritime clustering research. Studies applying K-Means to large-scale AIS datasets at other ports typically report silhouette scores in the range of 0.30-0.45, reflecting the inherent complexity and overlap in maritime traffic patterns rather than any inadequacy in the clustering method [20], [21]. In the Indonesian context, a recent study on the Bali Strait using K-Means clustering on AIS data similarly identified distinct vessel clusters and temporal peak patterns with a comparable silhouette coefficient of 0.3040, supporting the generalizability of the methodology adopted in this study [3].

Compared to studies relying solely on speed-course-position triplets from raw AIS streams, the enriched feature set used here — incorporating voyage distance, vessel capacity, draught, and operational type — yields clusters with stronger operational interpretability, making the results more actionable for port management purposes [4], [10]. The integration of temporal features into the PCA-K-Means pipeline also differentiates this work from purely spatial clustering approaches, enabling time-aware traffic density prediction rather than static pattern description [26], [27]. Studies at major hubs such as Singapore and Rotterdam have likewise identified clear clusters by vessel type and voyage behavior using AIS data, supporting the validity of distinguishing operational profiles in the way conducted here.

*Limitations*

Several limitations should be acknowledged when interpreting the results of this study. First, the dataset covers a single month (December 2025), which may not fully capture seasonal variation in maritime traffic at Surabaya Port; extending the analysis to a full annual cycle would provide a more robust characterization, particularly for identifying seasonal peaks linked to agricultural export cycles or national holiday effects. Second, the AIS data used here reflects declared vessel movements and may not capture vessels that deliberately disable or spoof their AIS transponders — a known limitation of AIS-based surveillance that could introduce selection bias in anomaly detection applications. Third, the moderate silhouette score of 0.3863 suggests that some operational regimes are not fully separable using the current feature set, and incorporating additional contextual variables such as weather conditions, tidal data, or cargo manifest information could improve cluster purity in future work. Fourth, the traffic density prediction component of this study is descriptive and pattern-based rather than formally predictive; incorporating machine learning regression or time-series forecasting methods could yield more precise future-state predictions for operational planning.

## 4. CONCLUSION

This study demonstrated that the combination of PCA-based dimensionality reduction and K-Means clustering is an effective approach for characterizing and predicting maritime traffic density at Surabaya Port using AIS big data. Applied to 1,173 vessel movement records from December 2025, the analytical pipeline identified 10 distinct operational clusters with a silhouette score of 0.3863, revealing clear temporal patterns — peak traffic at 14:00 and minimum traffic at 02:00 — alongside vessel-type-specific behavioral profiles that reflect the port's dual role as a regional and international maritime hub. Approximately 92% of observed hours fell into high or very high density zones, confirming that Surabaya Port operates near its effective capacity throughout most of the day and is vulnerable to cascading delays under demand surges or unplanned disruptions.

The practical implications of these findings are directly applicable to port management operations. The identified traffic patterns provide a data-driven basis for optimizing berth allocation schedules, pre-positioning

pilotage and tug resources ahead of predicted peak periods, and designing targeted interventions to reduce vessel idling and associated emissions. Port authorities can further use the temporal prediction framework developed here to anticipate congestion windows and proactively communicate traffic flow advisories to incoming vessels.

Future research should extend the temporal scope of the dataset to cover full annual cycles, incorporate external contextual variables such as weather and tidal conditions, and explore deep learning architectures — such as LSTM networks or graph neural networks — that may better capture the sequential and spatial dependencies inherent in vessel movement data. Integrating the cluster-based prediction model into a real-time decision support system represents a particularly promising avenue for translating these analytical findings into operational impact.

## CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

**Afif Zuhri Arfianto**: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft,. **Muhammad Izzul Haj**: Software, Data curation, Visualization, Investigation. **Muhammad Khoirul Hasin**: Validation, Resources, Data curation. **Noorman Rinanto**: Formal analysis, Validation, Writing – review and editing. **Imam Sutrisno**: Supervision, Conceptualization, Methodology. **Dimas Pristovani Riananda**: Writing – review and editing, Resources, Project administration. **Dwi Sasmita Aji Pambudi**: Validation, Investigation, Writing – review and editing.

## DECLARATION OF COMPETING INTERESTS

The authors declare that there are no conflicts of interest regarding the publication of this manuscript. The research was conducted independently, and the results presented are based on the objective analysis of the AIS data.

## DATA AVAILABILITY

The data used in this research, consisting of vessel trajectory, speed, draught, and temporal features recorded in Surabaya Port, are not publicly available due to operational privacy and maritime security considerations. However, the processed data used for maritime traffic density prediction and clustering analysis are available from the authors upon reques

## REFERENCES


[1] T. Zhang, J. Yin, X. Wang, and J. Min, "Prediction of container port congestion status and its impact on ship's time in port based on AIS data," *Marit. Policy Manag.*, vol. 51, no. 5, pp. 669–697, 2024.

[2] Y. Li, "Research on Multi-Port Ship Traffic Prediction Method Based on Spatiotemporal Graph Neural Networks," *J. Mar. Sci. Eng.*, vol. 11, no. 7, 2023.

[3] Z. Liu, W. Chen, C. Liu, R. Yan, and M. Zhang, "A data mining-then-predict method for proactive maritime traffic management by machine learning," *Eng. Appl. Artif. Intell.*, vol. 135, p. 108696, 2024.

[4] D. Zissis, K. Chatzikokolakis, G. Spiliopoulos, and M. Vodas, "A Distributed Spatial Method for Modeling Maritime Routes," *IEEE Access*, vol. 8, pp. 47556–47568, 2020.

[5] I.-L. Huang, M.-C. Lee, L. Chang, and J.-C. Huang, "AIS-Based Ship Trajectory Extraction Framework," *J. Mar. Sci. Eng.*, vol. 12, no. 9, p. 1672, 2024.

[6] L. Liu, Y. Zhang, Y. Hu, Y. Wang, J. Sun, and X. Dong, "A Hybrid-Clustering Model of Ship Trajectories," *J. Mar. Sci. Eng.*, vol. 10, no. 3, 2022.

[7] I. AbuAlhaol, R. Falcon, R. Abielmona, and E. Petriu, "Mining Port Congestion Indicators from Big AIS Data," in *IJCNN*, 2018.

[8] R. Mussabayev, N. Mladenovic, B. Jarboui, and R. Mussabayev, "How to Use K-means for Big Data Clustering," *Pattern Recognit.*, vol. 137, p. 109269, 2023.

[9] Z. Xiao, "Big Data Driven Vessel Trajectory Prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 3696–3709, 2022.

[10] D. Filipiak, M. Strozyna, K. Wecel, and W. Abramowicz, "Big Data for Anomaly Detection in Maritime Surveillance," *Zesz. Nauk. Akad. Mar. Wojennej*, vol. 215, no. 4, pp. 5–28, 2018.

[11] H. Li and T. Kim, "A Dimensionality Reduction-Based Multi-Step Clustering Method for Robust Vessel Trajectory Analysis," 2017.

[12] D. Liu, H. Rong, and C. Guedes Soares, "Shipping route modelling of AIS data," *Ocean Eng.*, vol. 289, p. 115868, 2023.

[13] Y.-Q. Zhang and W. Li, "Dynamic maritime traffic pattern recognition," *Sensors*, vol. 22, no. 16, p. 6307, 2022.

[14] A. Khodamoradi, "Vessel Traffic Density Prediction." 2025.

[15] S. et al. Kawashima, "Ship traffic flow by PCA of AIS," *J. Nav. Arch. Ocean Eng.*, vol. 23, pp. 55–63, 2016.

[16] P. Sheng, "Extracting shipping route patterns," *Sustainability*, vol. 10, no. 7, 2018.

[17] X. Han, C. Armenakis, and M. Jadidi, "Modeling vessel behaviours using DBSCAN," *Sustainability*, vol. 13, no. 15, 2021.

[18] K. et al. Skarlatos, "Ship engine model selection," *J. Mar. Sci. Eng.*, vol. 12, no. 1, p. 97, 2024.

[19] L. Wang, "Ship AIS trajectory clustering," *J. Mar. Sci. Eng.*, vol. 9, no. 6, 2021.

[20] Z. et al. Hanyang, "Vessel Sailing Patterns Analysis," in *ICBDA*, 2019.

[21] W. Xing, J. Wang, K. Zhou, H. Li, Y. Li, and Z. Yang, "A hierarchical methodology for vessel traffic flow prediction using Bayesian tensor decomposition and similarity grouping," *Ocean Eng.*, vol. 286, no. P2, p. 115687, 2023.

[22] G. et al. Xu, "AIS data analytics for vessel traffic service," *Ind. Manag. Data Syst.*, vol. 120, no. 4, pp. 749–767, 2020.

[23] W. Peng, X. Bai, D. Yang, K. F. Yuen, and J. Wu, "A deep learning approach for port congestion estimation and prediction," *Marit. Policy Manag*, vol. 50, no. 7, pp. 835–860.

[24] Y. Li, "Deep learning-powered vessel traffic flow prediction," *Eng. Appl. Artif. Intell.*, vol. 126, 2023.

[25] Z. et al. Dong, "Short-Term Vessel Traffic Flow Prediction," *Sustainability*, vol. 16, no. 13, p. 5499, 2024.

[26] E. D. et al. Ozkan, "Capacity analysis of Ro-Ro terminals," *Asian J. Shipp. Logist.*, vol. 32, no. 3, pp. 139–147, 2016.
[27] Y.-Q. Zhang and W. Li, "Dynamic Maritime Traffic Pattern Recognition," *Sensors*, vol. 22, no. 16, 2022.

.