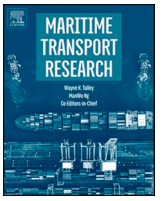
Maritime Transport Research 4 (2023) 100082

Contents lists available at ScienceDirect

Maritime Transport Research

journal homepage: www.sciencedirect.com/journal/maritime-transport-research



Self-labelling of tugboat operation using unsupervised machine learning and intensity indicator

Januwar Hadi [a], Dimitrios Konovessis [b], Zhi Yung Tay [a,*]

[a] Engineering Cluster, Singapore Institute of Technology, 10 Dover Drive, Singapore 534038, Singapore
[b] Department of Naval Architecture, Ocean and Marine Engineering, University of Strathclyde, 100 Montrose St, Glasgow G4 0LZ, United Kingdom

ARTICLE INFO

Keywords:
Machine learning
Self-labelling
Intensity indicators
K-means clustering
Fuel prediction

ABSTRACT

The actual operational data, such as a time sequence of fuel consumption and speed, is usually unlabeled or not associated with a specific activity like tugging or cruising. The operation type is critical for further analysis, as tugging and cruising operations require different fuel and navigation profiles. This paper aims to develop a self-labelling framework for tugboat operation by using unsupervised machine learning and a proposed intensity indicator. The framework considers two sets of data: the positional data and the fuel consumption rate data. The fuel consumption data is obtained from mass flowmeters installed on tugboats, while the positional data are navigational data purchased from marine data aggregators. The developed self-labelling enables ship operators in identifying operations and locations that require heavy fuel consumption andcan be used for further big data analytics and machine learning for fuel consumption prediction when vessel speeds are known.

1. Introduction

The tugboat is a common harbour craft vessel that assists larger vessels and other marine objects to ensure smooth and uninterrupted operations at the port. This type of harbour craft is known to generally have a very high power-to-tonnage ratio and consumes more fuel per job as compared to its counterparts such as ferries or patrol boats. The stark difference is even more pronounced when compared with the sea-going ships. The harbour tug's 4.0–9.5 kW/Gross-Rated-Tonnage is significantly more than the sea-going ship's 0.35–1.2 kW/Gross-Rated-Tonnage (Balakrishnan and Sasi, 2016). This indicates that the tugboats use fuel at a higher rate, despite a shorter time frame while serving the larger vessels around the harbour.

Despite its high power-to-tonnage ratio, the tugboats (harbour tugs) do not always use their full power, especially while taxiing between work sites (also known as cruising). In general, the tugboat operation could be classified into two categories depending on the power intensity requirement with respect to its speed. The first category is known as the tugging operation, and it is related, but not limited to operations such as bollard pulling/barge pulling. Operations such as assisting/stabilizing other vessel operations as well as anchor handling are also included in this first category (Kang et al., 2020). These operations require the tugboats to exert a large amount of power relative to their size. During these operations, the fuel consumption rate is also unsurprisingly high even though the tugboat may not travel at a high speed or even stationary. In short, fuel consumption has no direct correlation with the travelling speed of the boat. During this operation, the fuel usage is solely dictated by the power required to perform the tugging job.

* Corresponding author.
E-mail address: zhiyung.tay@singaporetech.edu.sg (Z.Y. Tay).

https://doi.org/10.1016/j.martra.2023.100082
Received 30 September 2022; Received in revised form 11 January 2023; Accepted 24 January 2023
Available online 31 January 2023

Fig. 1. The subject vessel, POSH Grace tugboat.

The second category is known as the cruising operation where the tugboats must travel between the various sites where their services are required (Lou et al., 2017). During cruising, the tugboat uses relatively less fuel to achieve higher speed. However, the power requirement, consequently the speed, is usually more constant with relatively less fluctuation as compared to the first category operation, e.g., the pulling operation. The correlation between the power or fuel consumption with the travelling speed could thus be sought and an optimized condition could be achieved for various travelling speeds and weather conditions. Therefore, the tugboat is free from load and its fuel consumption is mostly used for the movement of the vessel. In this second category, there is a fuel usage optimization opportunity by finding a theoretically ideal vessel speed. An example of this method is presented by training a fuel consumption rate prediction model based on vessel speed to demonstrate the potential application of the self-labelling tool.

Another challenge that arises for data analytics is the presence of a group of data points that overwhelms the other group of data points. For instance, the speed data at slow speed (or even zero), are abundant as they may also come from records when the vessel is in idling operations, thus filtering out the *invalid* data points is needed prior to further analysis. The presence of *invalid* data points is mainly due to the non-discriminative nature of data collection. The non-discriminative data collection means that the data is always being recorded regardless of whether the vessel is in, or not in operation. Thus, this one particular scenario from the idling condition contributes to a condition known as an imbalanced dataset (Kotsiantis et al., 2005) where the *valid* dataset for data analytics may be overwhelmed by the *invalid* ones. Another cause of an imbalanced dataset may also occur when the vessel is using fuel while not in motion, which is explained in Section 3.1.

One method to mitigate the imbalanced dataset is over-sampling the under-represented data. Another method is to under-sample the over-represented data (Kumar et al., 2021). Either way, some information related to the portion of the data that are under- or over-sampled must first be obtained. One way to discover the imbalanced data is through statistical analysis of the data population (Dixon and Massey, 1951). To present ideas throughout this paper, various data visualization tools are used for knowledge discovery (Grinstein and Wierse, 2002).

With the background and the problem statement mentioned in the earlier paragraphs, this paper aims to present a solution to self-label data points according to their intensity indicators by using unsupervised machine learning. In this paper, an example is presented to show the benefit of obtaining the label information. This is demonstrated by a selection of data points to improve the fuel prediction model using a neural network for a certain operation.

2. Data sources

A lack of good quality data is often the main challenge in data analytics involving harbour craft vessels, especially tugboats. Fuel consumption of tugboats relies on many factors, among them are the crew's experience on how to manoeuvre and navigate along the shore and port, as well as on how to engage the larger vessels to perform a particular tugging task/operation. These operations that require the utmost attention from the crew to complete the task safely and optimally, however, are rarely recorded in real-time in the tugboat's logbook/report. They are typically recorded after the fact, or not recorded at all in the logbook (i.e., ship's noon report) (Bialystocki and Konovessis, 2016). In addition to the manual record keeping, most if not all of the existing tugboats in Singapore do not have high-time resolution digital logs such as per-minute parameters logs throughout their daily tasks. This poses challenges in analyzing the operations in which the data available are only from manually or even verbally recorded sources prone to human error during recording.

Table 1
Tugboat specifications.

Main particulars	Value
Length overall (LOA)	29 m
Displacement	665 tons
Maximum speed	12 knots
Main engines	NIIGATA 6L26HLX
Number of engines	2
Total BHP	4000 BHP
Type of propulsion	Azimuth pod
Number of propulsors	2

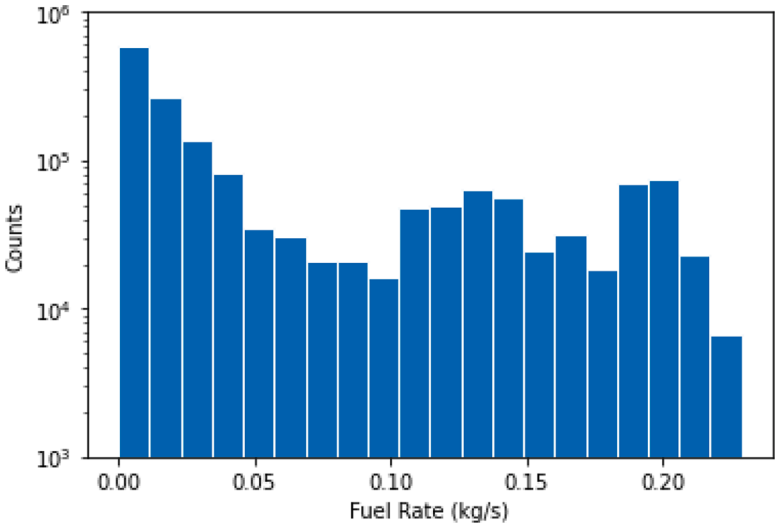


Fig. 2. Histogram of fuel rate.

Having said that, the authority requires harbour crafts such as tugboats normally operating in Singapore waters to be equipped with AIS transponders, which serves as a saving grace. These transponders enable the retrieval of positional coordinates. The retrieval of these positional coordinates can be done retroactively through marine data aggregators such as Marine Traffic (www.marinetraffic.com). This paper presents a solution to label every data point with an operation intensity by combining the positional data as well as the fuel consumption rate data collected from the mass flowmeter.

There are two sets of data used to self-label the tugboat operations by using unsupervised machine learning and intensity indicator. The two sets of data are: the vessel speed and fuel consumption data throughout a six-month period between April and October 2020. The fuel consumption and position data are described in detail in the following section followed by the description of the parameters and indicators used in achieving the self-labelling of tugboat operations.

2.1. Fuel consumption data

The subject tugboat used in collecting the fuel consumption is the Pacc Offshore Service Holding (POSH) Grace which operates in and around the Singapore port limits. The fuel consumption data are collected from POSH Grace between April to October 2020. Together with externally sourced data such as positional data, this makes up the dataset used for the analysis. POSH Grace tugboat is shown in Fig. 1 and the vessel's specifications are given in Table 1. The main activity of the tugboat consists of anchoring, assisting in large vessel docking, and piloting around the southern sea of Singapore. The purpose of collecting the operational data from the subject vessel was to conduct a research study in predicting the fuel rate to achieve fuel efficiency via data analytics and machine learning (Tay et al., 2021a, 2021b; Hadi et al., 2022a).

POSH Grace was equipped with mass flowmeters and a data logging device to measure the fuel rate consumed by the main engines. To facilitate the collection of the operational data, two Coriolis mass flowmeter sensors were installed to record the fuel consumption on both port and starboard main engines. The remaining two auxiliary engine fuel consumptions are not considered here as the consumption rate is almost constant thereby predictable and significantly smaller compared to the main engines.

The measurement taken from each mass flowmeter is in the unit of kilogram per second. Both fuel consumption rate measurements are summed together, creating a single fuel consumption rate. The histogram of the single fuel consumption rate is given in Fig. 2. Although mass flowmeters were installed to record the fuel consumption, there were challenges in collecting the measurements of the fuel consumption rate especially in its accuracy while the tugboat was in actual operation. The inaccuracy is mainly due to vibration

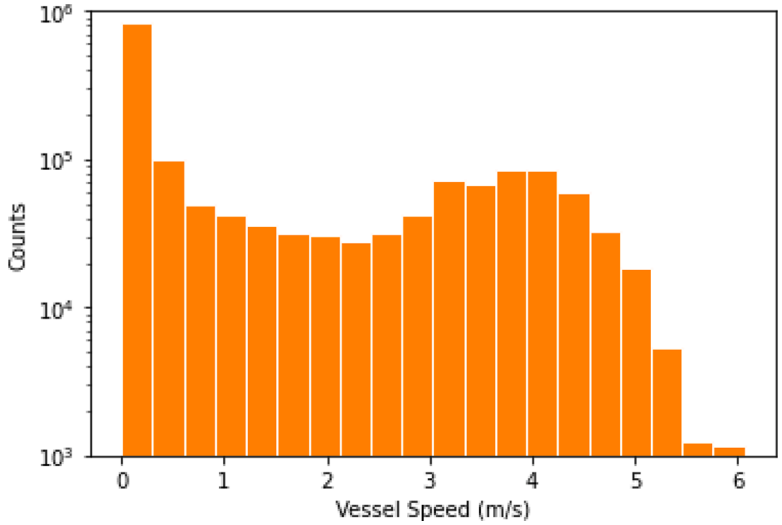


Fig. 3. Histogram of vessel speed.

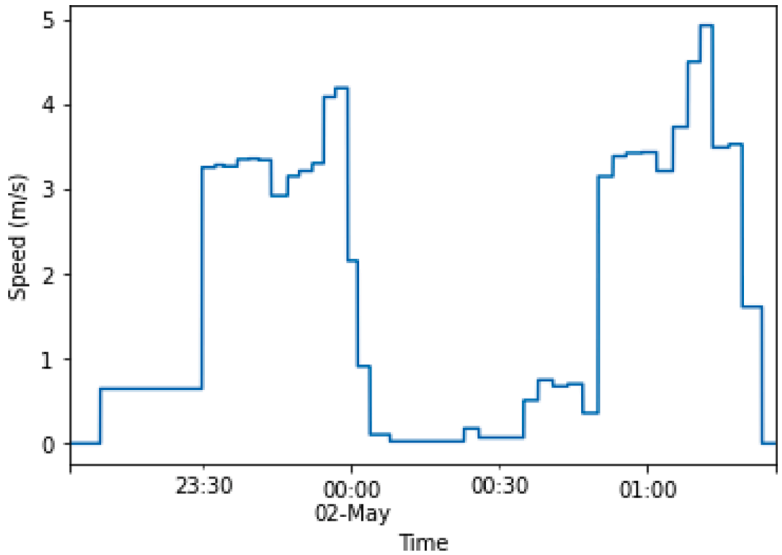


Fig. 4. Sample of sequential vessel speed by parsing the positional data.

noises that may disrupt the Coriolis effect – the principle that the mass flowmeter uses to make measurements. However, the discussion related to solving this challenge is not part of the scope of this paper and has been discussed in another paper by the same authors (Hadi et al., 2022b).

2.2. Vessel speed data

The vessel speeds must be derived from positional data obtained from marine data aggregator because the speedometers in tugboats, as with many other harbour crafts in the Singapore waters, are usually not logged. The positional data of POSH Grace were obtained from the marine data aggregator, Marine Traffic (www.marinetraffic.com) and the data comprise GPS coordinates of latitude and longitude as well as the time stamp of the vessel as it navigated in the Singapore waters. The geodesic distance and time-delta between every pair of sequential position coordinates were then obtained. The segmented vessel speed can then be calculated by the division of the segmented distance over the time-delta. Via this methodology, the series of GPS data were transformed into vessel speeds. Fig. 3 shows the histogram of vessel speeds calculated from the GPS data for POSH Grace.

When parsing vessel speeds from positional data, this however creates an inferior quality of vessel speed data. E.g., the vessel speeds recorded in Fig. 4 for a specific duration in May 2020 show that there are edges between the change in speed giving the impression that the tugboat moves in a discrete manner whereas, in reality, the tugboat changes speed in a more fluidic manner.

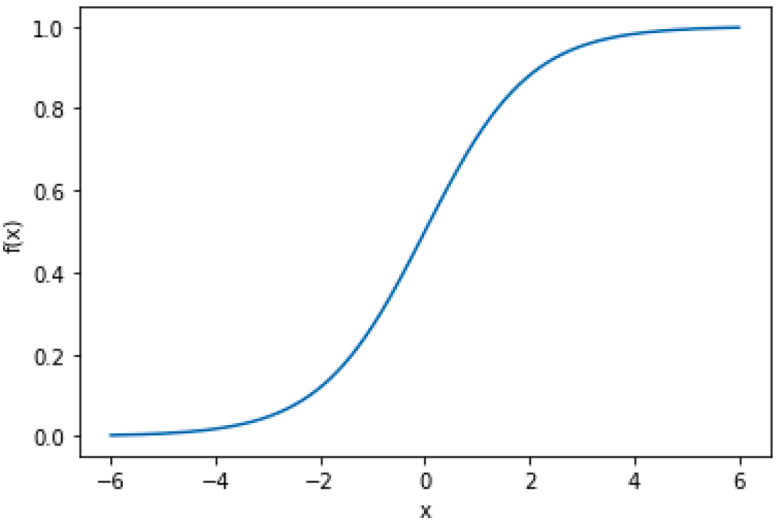


Fig. 5. Sigmoid curve.

Nonetheless, despite the limitation of the quality of the vessel speed data, this paper still aims to present a solution to the self-labelling of the dataset that represents the various tugboat operations by using operational intensity indicators.

3. Methodology

This section discusses a series of methodologies used to produce results for the sample application in the next section. A secondary variable, Fuel per Distance, is derived from the two variables (fuel consumption and vessel speed) to gauge the fuel usage in relation to the tugboat movement. The Fuel per Distance is transformed into operational intensity coefficient to emphasize the Fuel per Distance values that correspond to the cruising operation. The intensity's two most important statistics (mean and standard deviation) are geographically localized and clustered to create distinction of various intensities and to also group them as clusters. Finally, the cluster centers indicate the label of operational intensity (i.e., high and low intensity).

3.1. Fuel per distance

Using the fuel consumption rate data coupled with the vessel speed data, a secondary parameter can be calculated, i.e., the fuel usage per unit distance. It is referred to as the fuel per distance (FPD) throughout this paper. In simple terms, it is the amount of fuel used (in kilograms), in order to travel over a unit distance (in meters). While FPD is an essential measurement for cruising efficiency, the same cannot be said for tugging. During tugging, exerting a large amount of power (consequently, large fuel usage) over a smaller or zero distance is a necessity as explained in the Introduction section (Section 1). In other words, a particular power requirement is a must in an actual tugging operation and this power requirement is usually high during tugging.

3.2. Intensity indicator

The FPD is transformed into the intensity indicator to help with self-discovery or self-labelling of the operation. The intensity indicator is a coefficient introduced to indicate how intense the tugboat is using fuel relative to its movement. Hence, the use of the operational intensity indicator is highly applicable to tugboats considering that various pulling conditions involve varying power requirements.

Coupled with the FPD parameter, rationales for operation intensity can be drawn out as follows:

i When the FPD value is very high or even at infinity, it can be assumed that the tugboat is performing a high-intensity operation. This is a condition where the tugboat is using fuel without producing any movement, i.e. heavy-tugging job.
ii When the FPD value is very low or zero, the tugboat could be idling.

Using the two rationales alone, the FPD produces extreme values. At the low end (zero), the information at this point is not useful, as the tugboat is not using any fuel. At the high end (infinity), the information cannot be processed, as infinity is not a finite number. The useful information related to the relationship between vessel speed and fuel consumption rate lies between these two extremes. This paper uses the simple fact that the fuel and vessel speed data must converge (being proportional to each other) at low-intensity operations such as the cruising operation. In such an operation, the increase in fuel usage must amount to an increase in vessel speed. The opposite is true for high-intensity operations such as pulling operations where the fuel and vessel speed are divergent from each

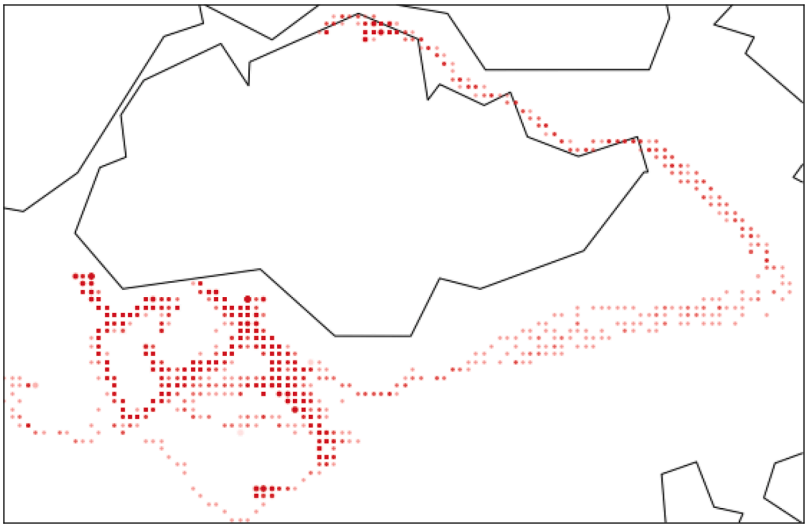


Fig. 6. Intensity indicator after localization.

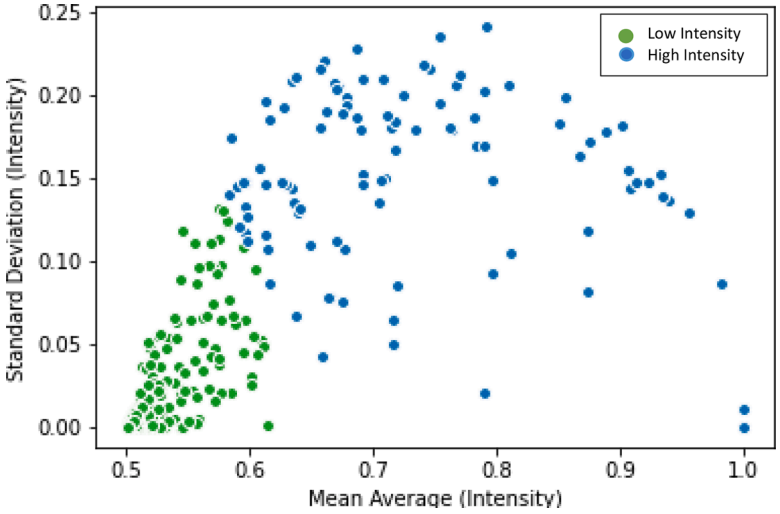


Fig. 7. Scatter plot of intensity indicator showing the two clusters and cluster members.

other. In such cases, the increase or high usage of fuel does not amount to higher vessel speeds.

The FPD values may range from zero to infinity. As it is impossible to work with data with very large values to infinity, FPD must be scaled down by shrinking the large values (or infinity) into finite values. It is also to scale down the large FPD numbers near infinity while to scale up smaller numbers near zero FPD. This can be done by transforming the FPD values using the logistic function (also known as the sigmoid function as shown in Fig. 5) given in Eq. (1)

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

where x is the FPD value in the problem studied here. The sigmoid function will transform the FPD into an intensity indicator represented by values between 0.5 to 1.0. The transformation shrinks the distance between large FPD values, especially large numbers closer to infinity.

The logistic function in Eq. (1) can exponentially scale down any number between minus infinity and positive infinity to values between 0.5 to 1. Once scaled, the larger FPD values towards infinity are associated with values closer to 1, indicating a high-intensity operation. At the other end of the spectrum, the low-intensity operator is indicated by values closer to 0.5. As there are no negative values, the lowest possible intensity value is 0.5. Transforming FPD to intensity indicator value using the sigmoid function compresses the larger FPD values while stretches the smaller FPD values. This gives more distinction to the lower FPD values (which is normally associated with the cruising operation) at the expense of the larger FPD values.

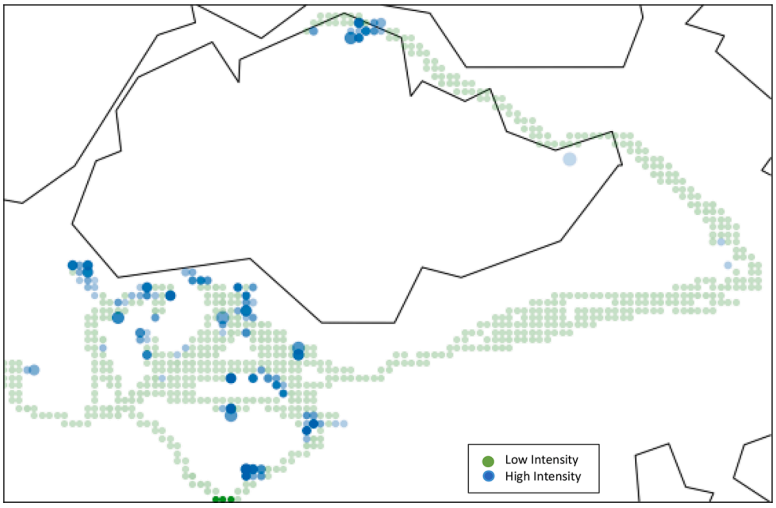


Fig. 8. Localized intensity indicator with cluster information.

3.3. Localization of intensity indicator

Each intensity indicator coefficient originated from a particular geographical location or latitude-longitude coordinate. By binning each of the intensity indicator values to a pre-determined geographical grid, the spectrum of intensity can be visualized, as presented in Fig. 6. The bin interval of the geographical grid in Fig. 6 is set at 0.005° for both latitude and longitude. It is noted that for simplicity, the grid shape is assumed to be square-like. The error in assuming the grid to be square-like is negligible as the data for vessel position used in this paper is close to the equator.

Each point in Fig. 6 is a mean (average) representation of the intensity indicator values in the bin. Visually, it can be observed that the locations where high-intensity (bolder red dots) operations occurred took place near the berthing points and the anchorages (the bolder red dots in the lower left region of Fig. 6) (Bialystocki and Konovessis, 2016). Geographical localization provides a useful insight into how the location could help in grouping the high- and low-intensity areas. In addition to the mean value, variance or standard deviation can also be calculated to help indicate if a particular coordinate is consistently experiencing a particular intensity.

3.4. Data point self-labelling

Each point in Fig. 6 is a representation of a group of intensity values at a particular coordinate bin. From each group, the mean and standard deviation are calculated. Both mean and standard deviation values are then populated in a scatter plot as shown in Fig. 7. By using the unsupervised machine learning technique – *K*-Means clustering (Celebi and Aydin, 2016; Hartigan and Wong, 1979), two clusters that represent two different operational profiles can be identified as shown in Fig. 7. Note that other clustering methods, such as hierarchical clustering and DBSCAN could be adopted but the main objective here is to develop a self-labelling framework for tugboat operations and the *K*-means clustering is used for demonstration purposes. These two clusters are grouped in accordance with their mean and standard deviation obtained from Fig. 6 as follows:

(i) First, the blue cluster with higher mean values relates to the high-intensity operations. It is shown by the mean values of the intensity indicators that are higher than the green ones. The blue cluster also has a higher standard deviation. It indicates that there are more variations of intensity in its data points. Hence, the members of the blue cluster should have higher adversity of operations. Both cruising and tugging jobs may take place in these data points.
(ii) The green cluster indicates that there are fewer variances as well as lower intensity. It may indicate that there are lower adversities of operation associated with the members of the green cluster.

After adding the cluster information from Fig. 7 to localize the intensity indicators from Fig. 6, the locations in which the high-intensity operations took place have become more obvious as presented in Fig. 8. The high-intensity operations took place in the southwestern and northern parts of Singapore between April and October 2020. In the eastern and southeastern parts of Singapore, the tugboat did not perform high-intensity operations. It is validated by the account from the tugboat crew/operator that the region in the lower right of Fig. 8 is where the tugboat would normally cruise to get to the northern region.

Relevant personnel could benefit from making informed decisions such as planning, bunkering, etc. based on the information on high-intensity operations as shown in Fig. 8. The description of the methodology up to this point uses two clusters to determine high- and low-intensity operation. To demonstrate the application/usage of this self-labelling methodology, a greater number of clusters are used to enhance a prediction model in the following Result and Discussion section.

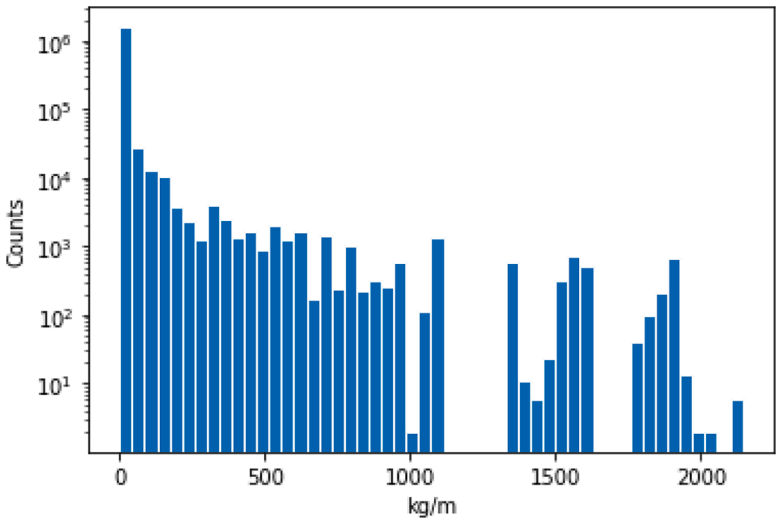


Fig. 9. Histogram of FPD (fuel per distance).

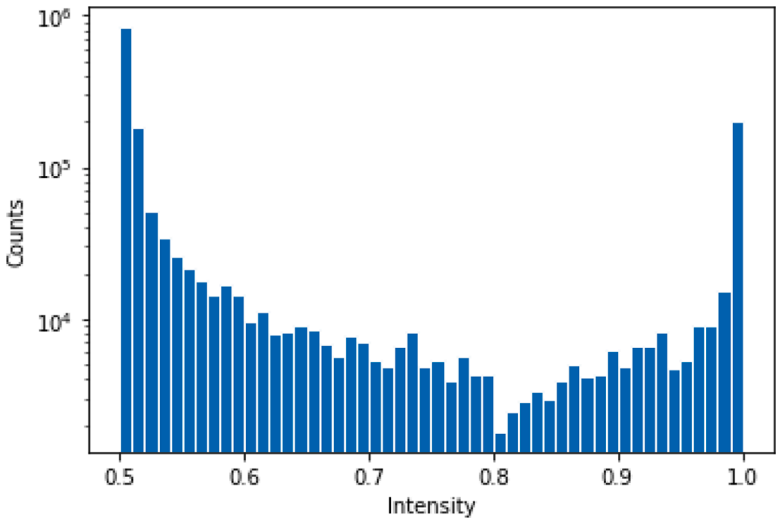


Fig. 10. Histogram of intensity indicator.

3.5. Development toolkits

To present the ideas and findings in the paper, a visualization tool is necessary. There are various commercialized visualization toolkits available, such as Tableau, which is one of the industry-leading software (Murray, 2013). Nonetheless, the visualization toolkit used in this paper is a highly customizable open-source software package, the `Matplotlib` library (Hunter, 2007). The visualization toolkit helps to explain the data structure and to make informed decisions such as discovering which part of the dataset needs to be removed for further analysis. In addition to the visualization toolkit, a data structure toolkit is also essential to import, contain, and manipulate the data. This paper uses the `Pandas` software package as the data structure toolkit (McKinney, 2010).

While processing the dataset, this paper utilizes the array vectorization toolkit – `Numpy` (Harris et al., 2020; Walt et al., 2011). Array is a popular term in computer science, which means a collection of data with the same data type. Array vectorization is the parallelization of computer calculation for arrays that do not have execution dependencies among one another (Weinhardt and Luk, 2001). The benefit of vectorization is a significant reduction in calculation time (Lemire and Boytsov, 2015). The statistical and scientific computing algorithm toolkit used in this paper is `Scipy` (Virtanen et al., 2019). Finally, the machine learning toolkits are `Scikit-learn` and `Tensorflow` (Abadi et al., 2016; Pedregosa et al., 2011). `Scikit-learn` is an open-source software suite for data analytics as well as machine learning (clustering). `Tensorflow` machine learning toolkit is used to construct, train, and predict a neural network (Jain et al., 1996; Fam et al., 2022, 2021), as a sample application for the solution presented in this paper.

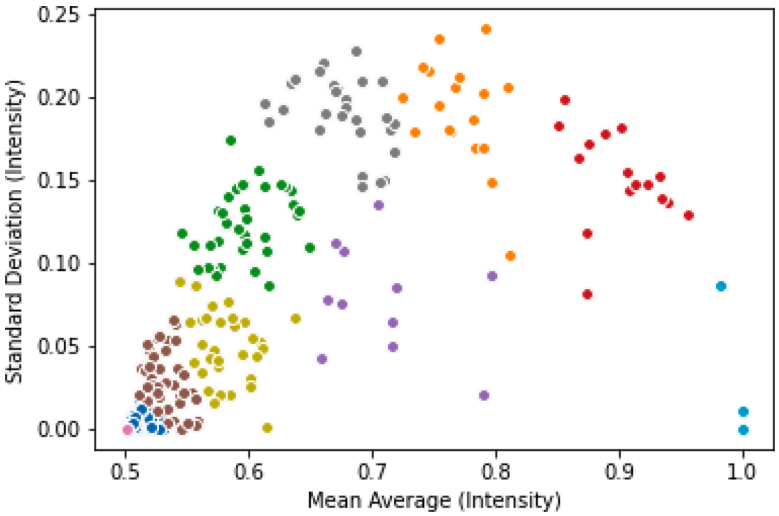


Fig. 11. Scatter plot of with intensity indicator with 10 clusters.

4. Results and discussion

This section discusses the application of the methods explained in Section 2. In the first two parts, this section addresses the imbalanced data issues (as explained in Section 1) and ways to overcome them. Next, the benefits of solving the imbalanced data with a better prediction model are presented.

As stated in Section 3.4, this paper uses a multi-cluster approach (more than two clusters) to discuss a sample application using the solution presented in this paper. Prior to discussing the application using the multi-cluster approach in Section 4.3, Section 4.1 discusses the data anomaly at the extreme ends of the FPD values (zero and infinity). It is followed by Section 4.2 which describes the solution to the issue discussed in Section 4.1.

4.1. Issues with zero & infinity FPD data counts

The FPD (the fuel consumption rate per unit distance) value is a gauge to measure the relationship between fuel consumption and vessel speed. It is obtained by dividing the fuel consumption rate by the vessel speed of the same rate so that this secondary parameter/variable produces values ranging from zero to positive infinity. Figs. 9 and 10 show the histogram of FPD and intensity indicator, respectively. It is to note that both figures use a logarithmic scale for the vertical *y*-axis. Also, Fig. 9 only shows the finite range of FPD values where the number of counts for infinity is not shown. In the Introduction section (Section 1), the condition for *invalid* data points is described to occur during the idling condition. Nonetheless, the *invalid* data points may also occur outside the idling condition, such as when the tugboat is operating at maximum intensity value. It is the condition when the tugboat is stationary while consuming fuel at any rate.

From the histograms of fuel rate (FR) and vessel speed (VS) shown in Figs. 2 and 3, respectively, both FR and VS values have a large number of counts at zero, which is associated with situations where the tugboat is not in operation or when it is idling. This large number of counts of zero value in VS and FR produces a high number of counts of zero FPD values and zero intensity indicators as shown both in Figs. 9 and 10. The high number of counts of the intensity indicator equal to 1.0 in Fig. 11 is due to the high counts of infinity FPD values caused by the extremely high-intensity operation when dividing the fuel rate by the distance value close to zero. In such conditions, the fuel being consumed is high, possibly while tugging but with minimal to no vessel movement. Therefore, the large counts of zeros for both FR and VS consequently result in large counts of zero and infinity FPD values.

The fuel consumption rate during extremely high-intensity operations (such as tugging) is highly unpredictable, for the reason explained earlier. On the other end, the number of data points associated with fuel consumption rates of extremely low intensity is overwhelming. This creates an imbalanced dataset as shown by the two extreme ends of Fig. 10. An imbalanced dataset means that a particular range of value datasets is over- or under-represented in the dataset. Therefore, it is good to exclude these two extremes in Fig. 10 as they are irrelevant. However, the exclusion of the two extremes is not necessary if the data are only used for visualization as shown in Fig. 9.

4.2. Solutions to zero & infinity FPD and selection of data points

To handle the *invalid* data point, a simple rule-based approach is used (Grosan and Abraham, 2011) by dropping or removing the data point with an extreme intensity value of either 0.5 or 1.0. After removing the extremes, the intensity now appears as a spectrum of values between 0.5 and 1.0. Application of clustering (*K*-Means clustering) is then utilized in grouping the intensity indicator in the

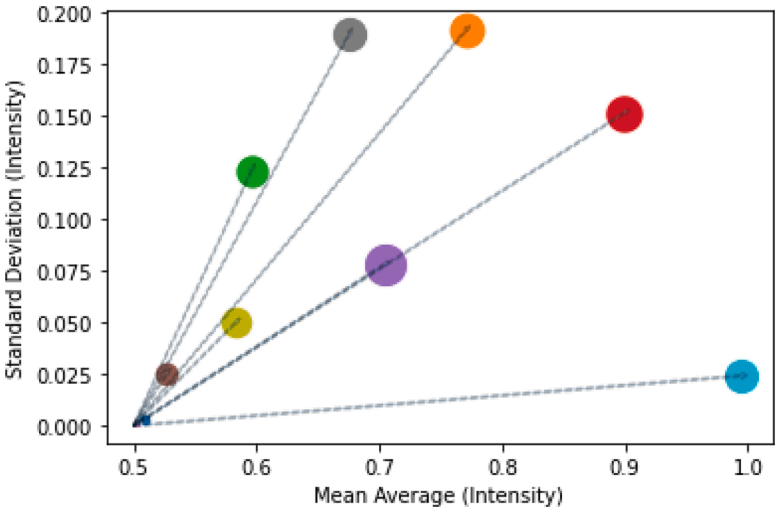


Fig. 12. Scatter plot of 10 cluster centers and its distance to point of origin.

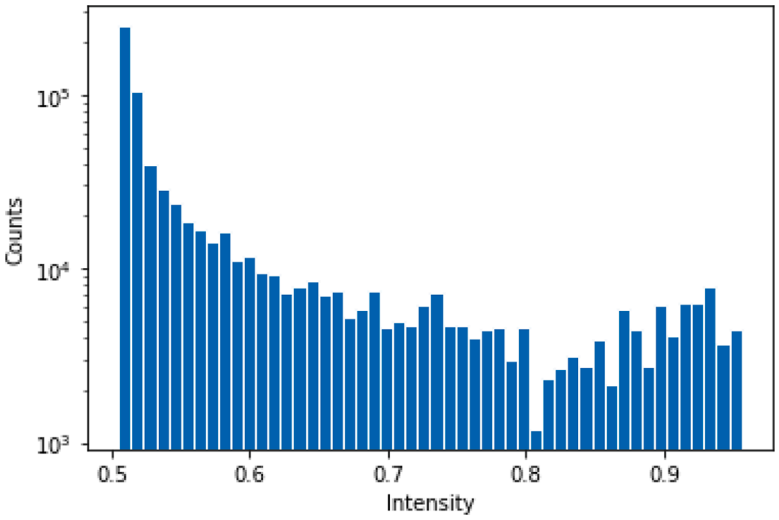


Fig. 13. Histogram of intensity indicator without members of two clusters.

mean and standard deviation space as shown in Fig. 11. It has the same members as Fig. 7, except that the members in Fig. 11 are associated with ten clusters instead of only two clusters. The means of data points from each geographical bin are arbitrarily coloured to visually distinguish the members from one another. The clusters of interest are the two clusters at the extremes. The pink cluster at the lower left of the scatter plot represents the lowest level of intensity whereas the cyan cluster at the rightmost of the scatter plot represents the highest level of intensity. The members of the pink and cyan clusters and data points are therefore associated with the two extremes in Fig. 10. Next, each cluster in Fig. 11 is represented by its cluster centre in Fig. 12, where the colour legend that represents different clusters is the same as in Fig. 11. By calculating the Euclidian distance between the origin (0.5, 0) to each cluster centre, the distance of each cluster to the point of origin can be calculated. Here, the point of origin is defined as the condition when the tugboat is idling. For the instance shown in Figs. 11 and 12, the two extremes are shown by the pink and cyan clusters.

The inclusion of data points from these two clusters for the prediction model (i.e., the pink and cyan clusters) tends to produce worse results. It is due to that the information from the pink and cyan clusters has a strong relationship with the idling and heavy tugging job as explained in Section 2.4. Therefore, the omission of data points from the pink and cyan clusters is necessary. The histogram of the intensity indicator showing the counts of the population without the two clusters and the redistribution of the counts with respect to the intensity indicator are presented in Fig. 13. It is to note that the values near the extreme ends (0.5 and 1) have been dropped/removed. Fig. 14 shows the localized intensity indicator after discarding the members of the two clusters. The self-labelling technique allows the mitigation of imbalanced data, by removing irrelevant data points. Thus, selectively under-sampling the dataset is favourable for a better prediction of operation with a certain range of intensity indicator values (i.e., intensity range while cruising). This completes the self-labelling technique using the operational intensity indicator.

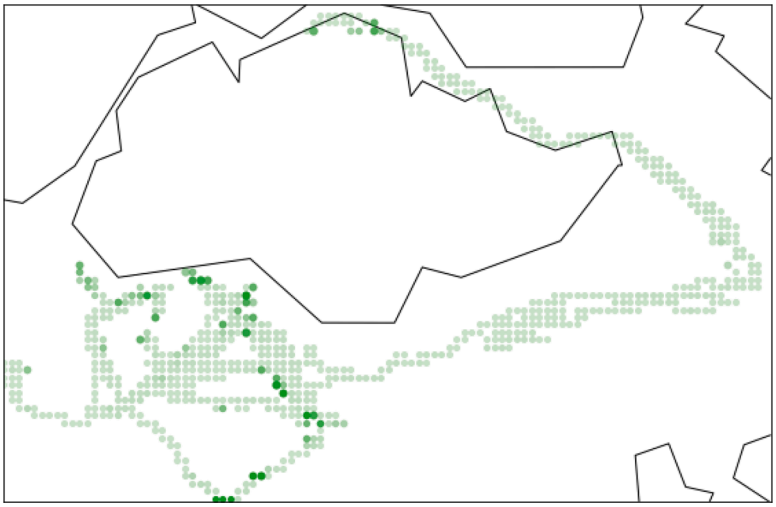


Fig. 14. Localized intensity indicator without members of two clusters.

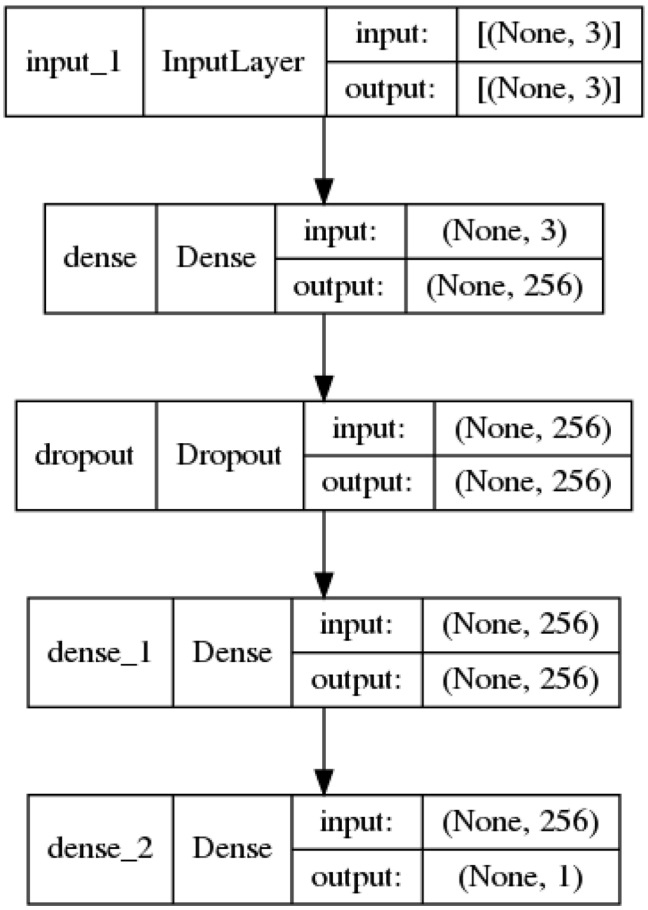


Fig. 15. Fully connected neural network.

4.3. A sample application of the self-labelling dataset: improving a prediction model

This subsection discusses the application of the intensity indicator to improve fuel rate prediction. As explained, at high-intensity operations as well as low-intensity operations, the fuel usage of the tugboat can be unpredictable. Hence, it is important to remove

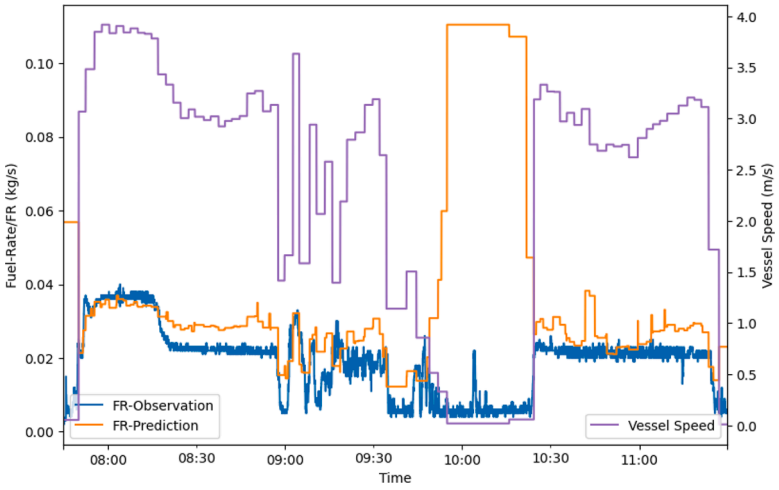


Fig. 16. Result with the balanced data.

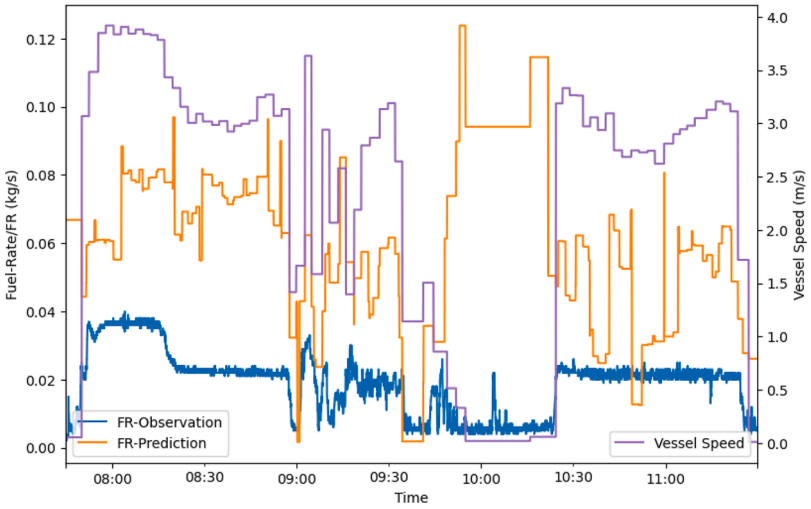


Fig. 17. Result with the imbalanced data.

them from the dataset. The removal process has been explained in the previous section using unsupervised machine learning, i.e., by applying the *K*-Means clustering on the statistics of intensity indicators.

A two-hidden layer neural network with 256 neurons each and one dropout layer (dropout rate is 0.125) is designed to showcase the application of the solution presented in this paper. The fully connected neural network is shown in Fig. 15. The statistics, i.e., the mean and standard deviation of the localized intensity indicator together with the vessel speed are used as input features to the neural network. The neural network is expected to predict fuel consumption from vessel speed. The fully connected neural network as shown in Fig. 15 was compiled with the `Adam` optimizer and the loss function of mean squared error (Bae et al., 2019; Chai and Draxler, 2014). Two neural networks were built and trained for 200 epochs. The first build-and-train of the neural network was done with the balanced data, producing a training loss value of 0.015. The second build-and-train was carried out with the imbalanced data, producing a training loss value of 0.051. `Tensorflow`, the machine learning framework mentioned in Section 3.5, is used to build and train the prediction model.

Figs. 16 and 17 show the comparison of the results of the neural network in fuel prediction. The first neural network in Fig. 16 was trained with the balanced data and made a very decent fuel prediction except for the duration at the beginning and ending of the operation, as well as the duration around 10:00 AM. The balanced data, that is used to produce Fig. 16, is all the data points in which the intensity indicator values fall between the Lower and Upper Limits (the Lower and Upper Limits of Intensity are shown in Fig. 18). The data points that are outside the Lower and Upper Limits are still included in the balance data. The number of data points that is outside the Lower and Upper Limits is 10% proportional to the number of data points that is inside the Lower and Upper Limits. In contrast, Fig. 17 shows the training with imbalanced data where the prediction results are far from the ideal.

Fig. 18 shows the intensity indicator for the same period as Figs. 16 and 17. Fig. 18 shows that at the durations when the prediction

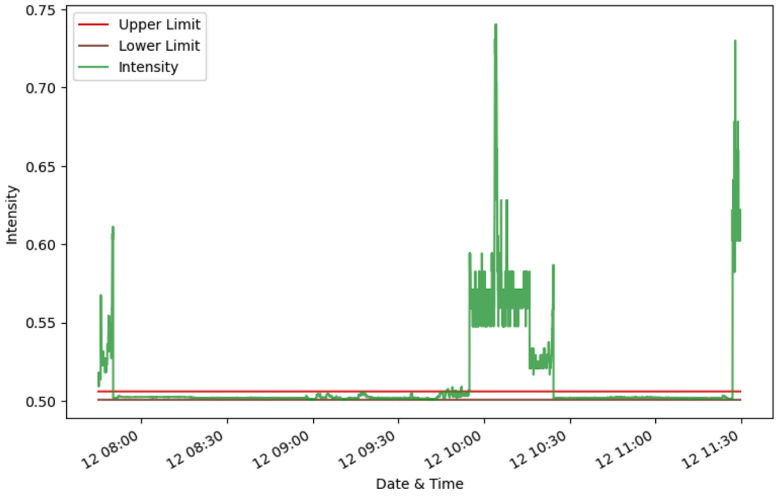


Fig. 18. Intensity indicator for the test period.

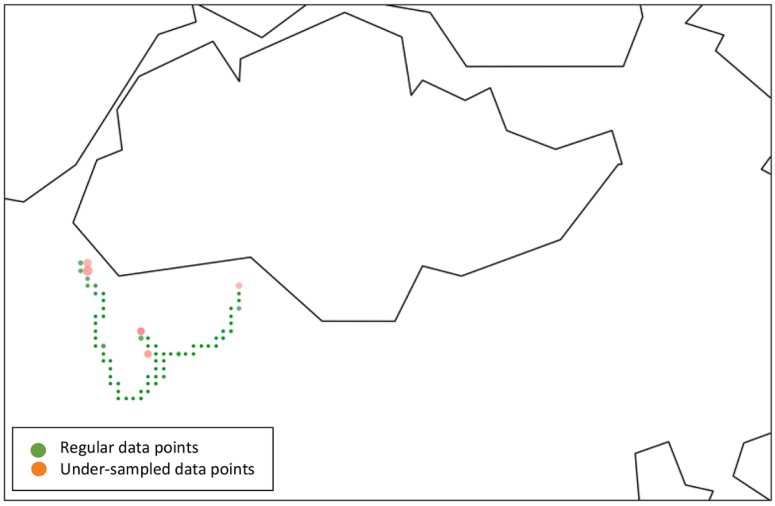


Fig. 19. Intensity indicator for the test period.

results in Fig. 16 do not produce good predictions (i.e., during the beginning and ending, as well as the portion around 10:00 AM), the intensity indicator is high. As the model was trained by data points between the Lower and Upper Limits, the prediction model, therefore, performs unfavourably when required to make predictions outside the limits. Having said that, the use of prediction model at these durations is practically insignificant as the vessel speeds are near zero (see Fig. 16). In Fig. 19, the periods in which the bad prediction are made are shown as red dots where these are the areas where the vessel does not usually cruise (i.e., berthing points). The limits may be derived from the Lowest and Highest Intensity Values from one cluster or a group of adjacent clusters. The data points associated with these clusters may have a strong relationship with the cruising activity, and finally may be labelled as cruising data points.

5. Conclusion

This paper presents a solution to self-label data points according to their intensity indicator by using unsupervised machine learning, i.e., the *K*-means clustering. It is achieved by using only two sets of data, the positional data and fuel consumption rate data. The intensity indicators and the labels are particularly useful as the manually recorded report (daily report) does not usually contain this information. Even if it is recorded by the crew, it is not tagged to each data point. The intensity indicator may be utilized as a visualization of operational intensity. Any person with interest could make a more informed decision making by utilizing the localized intensity indicator on every coordinate in a map. However, the use of the proposed intensity indication is not only limited to visualization. For instance, it can also be used when working with imbalanced data for improving the fuel prediction model as demonstrated in Section 3.3.

The aim of this paper was achieved using a mix of statistical analytics tools and machine learning methods. Both branches of machine learning, supervised and unsupervised, were discussed to make use of the intensity indicator as label. This paper also presents future opportunities to extend the dimension of the intensity indicator (from a scalar to a vector). Finally, a more complex and robust neural network can be designed to accommodate multiple steps of intensity ranges for a more generalized prediction model.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors wished to acknowledge the resources supported by SMI (R-SMI-A403- 706 0001) and MOE (R-MOE-A403-C002/MOE2018-TIF-1-G-008).

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., et al., Tensorflow: large-scale machine learning on heterogeneous distributed systems, ArXiv Prepr. arXiv: 1603 (2016). http://arxiv.org/abs/1603.04467.

Bae, K., Ryu, H., Shin, H., Does Adam optimizer keep close to the optimal point?, ArXiv Prepr. ArXiv1911.00289, (2019).

Balakrishnan, P., Sasi, S., 2016. Technological and economic advancement of tug boats. IOSR J. Mech. Civ. Eng. 87–96.

Bialystocki, N., Konovessis, D., 2016. On the estimation of ship's fuel consumption and speed curve: a statistical approach. J. Ocean Eng. Sci. 1, 157–166. https://doi.org/10.1016/j.joes.2016.02.001.

Celebi, M.E., Aydin, K., 2016. Unsupervised Learning Algorithms. Springer.

Chai, T., Draxler, R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? Geosci. Model Dev. 7 https://doi.org/10.5194/gmdd-7-1525-2014.

Dixon, W.J., Massey Jr, F.J., 1951. Introduction to Statistical Analysis. McGraw-Hill.

Fam, M.L., Tay, Z.Y., Konovessis, D., 2021. An artificial neural network based decision support system for cargo vessel operations. In: Proceedings of the 31st European Safety and Reliability Conference. Research Publishing Services, pp. 3391–3398. https://doi.org/10.3850/978-981-18-2016-8_758-CD.

Fam, M.L., Tay, Z.Y., Konovessis, D., 2022. An artificial neural network for fuel efficiency analysis for cargo vessel operation. Ocean Eng. 264, 112437 https://doi.org/10.1016/J.OCEANENG.2022.112437.

Grinstein, G.G., Wierse, A., 2002. Information Visualization in Data Mining and Knowledge Discovery. U.M.F. Morgan Kaufmann.

Grosan, C., Abraham, A., 2011. Rule-based expert systems. Intell. Syst. A Mod. Approach 149–185. https://doi.org/10.1007/978-3-642-21004-4_7.

Hadi, J., Tay, Z., Konovessis, D., 2022a. Ship navigation and fuel profiling based on noon report using neural network generative modeling. J. Phys. Conf. Ser. 2311, 12005. https://doi.org/10.1088/1742-6596/2311/1/012005.

Hadi, J., Konovessis, D., Tay, Z.Y., 2022b. Filtering harbor craft vessels' fuel data using statistical, decomposition, and predictive methodologies. Marit. Transp. Res. 3, 100063 https://doi.org/10.1016/j.martra.2022.100063.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., et al., 2020. Array programming with NumPy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: a k-means clustering algorithm. J. R. Stat. Soc. Ser. C (Applied Stat. 28, 100–108.

Hunter, J.D., 2007. Matplotlib: a 2D graphics environment. Comput. Sci. Eng. 9, 90–95. https://doi.org/10.1109/MCSE.2007.55.

Jain, A.K., Mao, J., Mohiuddin, K.M., 1996. Artificial neural networks: a tutorial. Computer 29, 31–44. https://doi.org/10.1109/2.485891.

Kang, L., Gao, S., Meng, Q., 2020. Capacity analysis of ship-tugging operations in a large container port. Asian Transp. Stud. 6, 100011 https://doi.org/10.1016/j.eastsj.2020.100011.

Kotsiantis, S., Kanellopoulos, D., Pintelas, P., 2005. Handling imbalanced datasets: a review. GESTS Int. Trans. Comput. Sci. Eng. 30, 25–36.

Kumar, P., Bhatnagar, R., Gaur, K., Bhatnagar, A., 2021. Classification of imbalanced data:review of methods and applications. IOP Conf. Ser. Mater. Sci. Eng. 1099, 12077. https://doi.org/10.1088/1757-899x/1099/1/012077.

Lemire, D., Boytsov, L., 2015. Decoding billions of integers per second through vectorization. Softw. Pract. Exp. 45, 1–29.

Lou, D.M., Bao, S.J., Hu, Z.Y., Tan, P., 2017. Cruise speed optimization of tugboat based on real fuel consumption and emission. Jiaotong Yunshu Gongcheng Xuebao J. Traffic Transp. Eng. 17, 93–100.

McKinney, W., 2010. Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference, 445, pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

Murray, D.G., 2013. Tableau Your Data!: Fast and Easy Visual Analysis with Tableau Software. John Wiley & Sons.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.

Tay, Z.Y., Hadi, J., Chow, F., Loh, D.J., Konovessis, D., 2021a. Big data analytics and machine learning of harbour craft vessels to achieve fuel efficiency: a review. J. Mar. Sci. Eng. 9, 1351. https://doi.org/10.3390/JMSE9121351.

Tay, Z.Y., Hadi, J., Konovessis, D., Loh, D.J., Tan, D.K.H., et al., 2021b. Efficient harbor craft monitoring system: time-series data analytics and machine learning tools to achieve fuel efficiency by operational scoring system. In: Proceedings of the ASME 2021 40th International Conference on Ocean, Offshore, and Arctic Engineering. American Society of Mechanical Engineers Digital Collection. https://doi.org/10.1115/OMAE2021-62658.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., et al., 2019. SciPy 1.0-fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272. http://arxiv.org/abs/1907.10121.

Walt, S.Van Der, Colbert, S.C., Varoquaux, G., 2011. The NumPy array: a structure for efficient numerical computation. Comput. Sci. Eng. 13, 22–30.

Weinhardt, M., Luk, W., 2001. Pipeline vectorization. IEEE Trans. Comput. Des. Integr. Circuits Syst. 20, 234–248.