Applied Ocean Research 121 (2022) 103052

Contents lists available at ScienceDirect

Applied Ocean Research

journal homepage: www.elsevier.com/locate/apor

A real-time data-driven framework for the identification of steady states of marine machinery

Christian Velasco-Gallego *, Iraklis Lazakis

Department of Naval Architecture, Ocean and Marine Engineering, University of Strathclyde, 100 Montrose Street, G4 0LZ, Glasgow, United Kingdom

ARTICLE INFO

Keywords:
Smart maintenance
Marine machinery
Maritime transportation
Connected component analysis
Steady states identification
Markov chains

ABSTRACT

While maritime transportation is the primary means of long-haul transportation of goods to and from the EU, it continues to present a significant number of casualties and fatalities owing to damage to ship equipment; damage attributed to machinery failures during daily ship operations. Therefore, the implementation of state-of-the-art inspection and maintenance activities are of paramount importance to adequately ensure the proper functioning of systems. Accordingly, Internet of Ships paradigm has emerged to guarantee the interconnectivity of maritime objects. Such technology is still in its infancy, and thus several challenges need to be addressed. An example of which is data preparation, critical to ensure data quality while avoiding biased results in further analysis to enhance transportation operations. As part of developing a real-time intelligent system to assist with instant decision-making strategies that enhance ship and systems availability, operability, and profitability, a data-driven framework for the identification of steady states of marine machinery based on image generation and connected component analysis is proposed. The identification of such states is of preeminent importance, as non-operational states may adversely alter the results outlined. A case study of three diesel generators of a tanker ship is introduced to validate the developed framework. Results of this study demonstrated the outperformance of the proposed model in relation to the widely implemented clustering models *k*-means and GMMs with EM algorithm. As such, the proposed framework can adequately identify steady states appropriately to guarantee the detection of such states in real-time, whilst ensuring computational efficiency and model effectiveness.

1. Introduction

According to the fourth IMO Greenhouse Gas (GHG) Study 2020, emissions are projected to increase about 90–130% of 2008 emissions by 2050 (IMO, 2021). Although these projections can fluctuate based on the impact of COVID-19, as well as long-term economic and energy scenarios, there is no doubt whatsoever about how critical the application of effective strategies is to contribute towards Net Zero goals. To address such a challenge, various technologies are being investigated, such as energy-saving technologies (Tacar et al., 2020), renewable energies (Dong et al., 2021), alternative fuels (Balcombe et al., 2021), and ship speed (Taskar and Andersen, 2020). If maintenance approaches are considered, these have demonstrated their ability to enhance efficiency, reliability, profitability, and performance of the vessel, while facilitating the emissions reduction all along its operational lifetime (Cheliotis and Lazakis, 2018; Lazakis and Ölçer, 2016). Therefore, the analysis of smart maintenance is of preeminent importance to ensure marine vessels efficiency, and thus reduce emissions and improve ships operational capacity.

Maintenance activities are usually presented in three different typologies: reactive maintenance, time-based maintenance, and Condition-Based Maintenance (CBM) (Lu et al., 2018; Emovon et al., 2018). However, due to an increase in data utilisation and accessibility, the prosperity of CBM has been possible in order to enhance the anticipation of forthcoming failures in marine machinery, thus promoting cost diminution by averting random preventive maintenance and crisis-related reactive maintenance of critical machinery. Accordingly, a large number of sensors are installed alongside the most critical components and around the environment where the assets are operating in order to effectively monitor their condition (Jamshidi et al., 2018; Su et al., 2019; Zhu et al., 2019). As a consequence, the Internet of Ships (IoS) paradigm has emerged due to the need of smart interconnecting maritime objects applicable to monitor the assets' condition (Lazakis et al., 2018; Raptodimos and Lazakis, 2018, 2019). Despite the undeniable enhancements of IoS, there are several challenges that are yet to be addressed. Examples of these are related to satellite communications,

* Corresponding author.
E-mail address: christian.velasco@strath.ac.uk (C. Velasco-Gallego).

https://doi.org/10.1016/j.apor.2022.103052
Received 19 July 2021; Received in revised form 27 October 2021; Accepted 6 January 2022
Available online 9 February 2022

security and privacy, data collection, and data management and analytics (Aslam et al., 2020), which preclude the data standards and quality required to provide reliable data analysis.

Therefore, both the analysis and development of innovative data preparation approaches are critical to adequately consider the specificities of the maritime industry. Although several studies have been performed in relation to data preparation (Cheliotis et al., 2019; Velasco-Gallego and Lazakis, 2020; Dalheim and Steen, 2020, Velasco-Gallego and Lazakis, 2021), further analysis is required to promote the implementation of real-time intelligent systems in order to enhance ship availability, and thus increase company profitability. In this study, the steady states identification step is analysed, as raw data usually contain non-operational states that adversely alter the results outlined when performing data-driven tasks to assess the current and future health of marine machinery. Although the marine engines typically run under steady-state conditions, fluctuations may occur due to, for instance, environmental conditions or variations in the operating condition (Theotokatos et al., 2020). Therefore, if such states are not adequately addressed, a decrease in both computational efficiency and model effectiveness can be perceived. Consequently, we propose an innovative approach constituted by image generation of time series sensor data and connected component analysis to adequately determine such states. Additionally, to address the current demands in relation to the transition from historical analysis to real-time analysis, the real-time implementation of the proposed framework is also introduced.

The following paragraphs are structured as follows. Section 2 presents a critical investigation of steady states identification phase. Section 3 describes the proposed methodology. Section 4 reflects on the results obtained after implementing the proposed methodology through a case study and a comparative analysis. Lastly, in Section 5 the conclusions and future work are outlined.

2. Literature review

Several efforts have been made to enhance the current practices in relation to Operations & Maintenance (O&M) activities within the shipping sector. Cao et al. (2020) presented an optimised Support Vector Machine (SVM) driven approach by Improved Artificial Bee Colony (IABC) as an effective state estimation method in ship system. Ellefsen et al. (2019) reviewed four well-established deep learning techniques applied in Prognostics and Health Management (PHM) systems: Deep Belief Network (DBN), Auto-Encoder (AE), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN). Also, some of the benefits and challenges to be faced in relation to PHM based on Deep Learning (DL) were introduced. In relation to the benefits, the authors suggested that the provision of high-speed broadband connections to ships at sea would enable online PHM systems based on DL, which could facilitate autoships without onboard maintenance personnel and achieve zero-downtime performance. Hence, it was thought that when PHM systems based on DL were introduced they could contribute to reduce errors occurring due to personnel, as systems were less dependant on prior knowledge and human influence. Brandsæter et al. (2017) presented a cluster-based anomaly detection methodology. This was based on an original methodology that was divided into two main steps: signal reconstruction, through the implementation of Auto Associative Kernel Regression (AAKR), and residual analysis, by performing Sequential Probability Ratio Test (SPRT). The methodology was then modified to include two new steps: cluster analysis, by the utilisation of the *k*-means algorithm, and the selection of a set of closest points per cluster, which would replace the original dataset as training set to reduce the computational cost. The proposed approach was assessed by analysing sensor signals on a marine diesel engine. Fault data were simulated to be implemented as the test set, as no fault data were available. The technique demonstrated to be successful in detecting anomalies and the computation time was reduced in relation to the original methodology. The proposed methodology is expanded in Brandsæter et al. (2019), which introduced a comprehensive description of the generalisations and modifications performed in the original methodology. As mentioned, cluster analysis was applied to replace the original dataset with rectangular boxes that referred to different clusters. In addition, the distance measure was altered to treat the variables differently based on the credibility of the signal and to distinguish between explanatory and response signals. Credibility estimation was also performed. Cheliotis et al. (2020) combined Expected Behaviour (EB) models with the Exponential Weighted Moving Average (EWMA) for fault detection. Four different regression models were assessed: OLS single linear regression, multiple linear ridge regression, OLS single polynomial regression, and multiple polynomial ridge regression. Multiple polynomial ridge regression was identified as the most accurate to detect faults manifesting in both the main engine cylinder exhaust gas temperature and the main engine scavenging air pressure. As the collected data represented fault-free operating conditions, a total of four different fault cases were simulated in the form of a sensitivity analysis. The estimated residuals were analysed in an EWMA control chart that contained upper and lower control limits to detect faults. It was concluded that the proposed approach could successfully detect imminent faults by analysing the residuals from the recorded and expected occurrences. Data preparation was of paramount importance in this study due to the characteristics of the raw data and the models that were implemented. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was applied effectively to remove outliers and transient states of operation, and thus induce its applicability when dealing with these types of data. Lazakis et al. (2019) proposed a methodology for the monitoring and detection of operating anomalies in ship machinery based on a one-class Support Vector Machine (SVM). The model was trained by using data that corresponded to the normal behaviour of a diesel generator under varying operating conditions. Abnormal data was simulated in the form of a sensitivity analysis. The proposed approach was effective for identifying anomalies. Tan et al. (2020) investigated the performance of the following one-class classifiers: One Class Support Vector Machine (OCSVM), Support Vector Data Description (SVDD), Global *k*-Nearest Neighbors (GKNN), Local Outlier Factor (LOF), Isolation Forest (IF), and Angle-Based Outlier Detection (ABOD). To that end, a real-data validated numerical simulator developed for a Frigate characterised by a combined diesel-electric and gas propulsion plant was utilised for a case study implementation. Based on the outlined results, the authors sorted the performance of the six analysed algorithms as follows: ABOD $>$ OCSVM $\approx$ SVDD $>$ GKNN $>$ IF $\approx$ LOF. Coraddu et al. (2019) applied anomaly detection methods based on SVMs and *k*-NN to predict the hull condition using available parameters measured on-board. Data collected from the Research Vessel The Princess Royal was utilised to perform a case study, the results of which demonstrated the applicability and the effectiveness of the proposed methodology.

However, if data preparation, specifically steady states identification, is considered, it can be perceived that the establishment of data-driven models for such a matter is not yet widely formalised. Of a total of 7 publications that consider the steady states identifications as a pre-processing phase, only 4 of them implemented data-driven models. Perera and Mo (2016) implemented Gaussian Mixture Models (GMMs) with an Expectation Maximization (EM) algorithm and Principal Component Analysis (PCA) to both classify and analyse frequent operating regions of marine engines. Dalheim and Steen (2020b) developed a new computationally efficient method to identify those parts of the time series data that refer to steady states by assuming that the underlying system behaviour could be modelled by a deterministic linear trend model. Velasco-Gallego and Lazakis (2021) implemented the *k*-means clustering technique to identify substantial steady states. Although the results of such methodologies demonstrated promising results, the case studies implemented to analyse their performance were assessed by only considering different engine loads, although there are states other than engine operating regions that may need to be analysed, such as transient and idle states. Moreover, both *k*-means and GMM with EM require the

selection of its hyperparameters prior to the selection of the different states. This may be unfeasible when deploying this pre-processing step in real time, as new operating regions may arise. Accordingly, the hyperparameters that determine the number of clusters need to be updated, which lead to an increase of the computational time. Therefore, further studies need to consider both the need for labelling a large number of historical data with different states (such as engine loads and transient and idle states), and the current demand for transitioning from historical analysis to real-time analysis. In addition, several studies are still considering non-data-driven models to address this matter, which is more time consuming due to the need for human resources; thus, increasing the probability for human error. For instance, Ellefsen et al. (2020) manually divided the engine loads into five distinct operating conditions to perform multiregime normalization. However, the need for automating such a process was also indicated in this study, as new operating conditions may be encountered in real-life systems. Accordingly, an innovative approach for addressing the steady states identification phase through the implementation of image generation and connected component analysis is presented. Although time series imaging has demonstrated promising results when applying forecasting (Li et al., 2020), such a matter has not been applied to the best of the authors' knowledge. As such, the analysis of such techniques is explored within this enquiry in other to assess its possible potential in the identification of steady states. Additionally, challenges such as the lack of availability in operations monitoring, the quality of raw data, and the provision of the information in real time are also considered for the methodology development. To assess its effectiveness, a case study is also presented where, not only different engine loads are presented, but transient and idle states are also perceived. Furthermore, a comparative study, in which both *k*-means and GMM with EM are also analysed for validation purposes.

3. Methodology

The proposed methodology is graphically represented in Fig. 1. The first step refers to the pre-processing of the input time series data, in which the overall time series is sectioned into sequences by applying the sliding window algorithm. Subsequently, each sequence is transformed into an image by estimating the transition matrix obtained from the implementation of the first-order Markov Chain. To adequately determine the different regions identified in each of the images, connected component analysis is conducted and, consequently, post-processing is performed on the outcoming images to transform them into sequences. As the states are labelled per sequence, results from the preceding phase need to be also pooled to achieve the input time series with the resulting labels that specify the different steady states identified.

3.1. Data pre-processing

To adequately apply pre-processing, a previous data understanding phase needs to be applied to establish the steps required based on the characteristics of the data set. A frequent challenge to be addressed when dealing with data of marine systems is data imputation, as missing values are usually encountered. In addition, data denoising is also applied due to the susceptibility of the time series to contain high noise. Accordingly, Exponentially Weighted Moving Average (EWMA) is applied. The last step applied in the data pre-processing step is the division of the time series into sequences by the application of the sliding window algorithm.

3.2. Image generation

To adequately identify the different steady states all along the analysed data set, the input time series is transformed into an image by the implementation of the first-order Markov chain. By applying such a process, it is determined that the occurrence at time *t* just hinges on the previous value and not on all values at time before *t*. Thus, if the time series values are clustered in a finite number of states, the first-order Markov chain transition matrix can be estimated, which will be considered as an image, and thus each of the pixels will be considered as an element of the matrix.

To estimate such a matrix, the definition of the discrete time stochastic process is considered. A discrete time stochastic process, $(X_n)_{n\in\mathbb{N}}$, which takes values in a finite set S, is considered to have the Markov property if the probability distribution of X_{n+1} at time $n+1$ only hinges on the previous state X_n at time n, and not on all the past values of X_k for $k \leq n-1$. Thus,

$$\mathbb{P}(X_{n+1}=j|X_n=i_n,\ X_{n-1}=i_{n-1},\ \ldots,\ Z_0=i_0) = \mathbb{P}(Z_{n+1}=j|Z_n=i_n) = p(i,j) \tag{1}$$

where $i_0,\ i_1,\ \ldots,\ i_n.j\ \in S$.The probability $p(i,j)$ indicates the probability that the previous state i is followed by the current state j. All the possible transition probabilities of a process can be collected in a *rxr* matrix, where each (i,j) entry P_{ij} is $p(i,j)$,

$$P = \left(P_{ij}\right)_{1\leq i,\ j\leq r} \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,r} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,r} \\ \vdots & \vdots & \ddots & \vdots \\ p_{r,1} & p_{r,2} & \cdots & p_{r,r} \end{pmatrix} \tag{2}$$

and that satisfies

$$0 \leq P_{ij} \leq 1,\ 1 \leq i,j \leq r, \tag{3}$$

$$\sum_{j=1}^{r} P_{ij} = 1,\ 1 \leq i \leq r. \tag{4}$$

3.3. Connected component analysis

By considering the transition matrix estimated in the preceding step as a collection of discrete cells, a.k.a., pixels, the transformation from time series to image is achieved. Thus, each pixel is associated with a pixel value, which lies between 0 and 1 (inclusive) and refers to the probabilities formerly estimated.

In turn, to facilitate the implementation of connected component analysis, the image outlined is converted to a binary one with only two possible intensity values. Such a conversion is performed according to (Eq. (5)), in which the binary image is obtained by classifying the different pixel values into either 0, if the probability associated with the pixel is equal to 0, or 1, otherwise. Thus, those pixels that present information about a transition between states can be efficiently identified.

$$P_{ij} = \begin{cases} 0, & if\ P_{ij} = 0 \\ 1, & otherwise \end{cases} \tag{5}$$

By applying this conversion, the distinct transition clusters presented within the image can be labelled. Accordingly, pixel connectivity is analysed, which characterises the relationship between pixels. To consider that two neighbouring cells are connected, they must present the same pixel value. For this enquiry such a connectivity is formulated by applying the 4-neighbours adjacency criterion (see Fig. 2). Thus, the notation of neighbourhood for such a case is expressed hereunder.

$$N_4(p) = \{(x+1,\ y).\ (x-1,\ y),\ (x,\ y+1),\ (x,\ y-1)\} \tag{6}$$

All possible neighbouring pixel connectivity is evaluated to determine the distinct sets of connected pixels, a.k.a. connected components. Therefore, the last step of this phase, named connected components labelling, is achieved, in which the different connected components are clustered to identify the different states, and in turn determine those that only refers to steady states. A graphical representation of such a phase is expressed in Fig. 3.

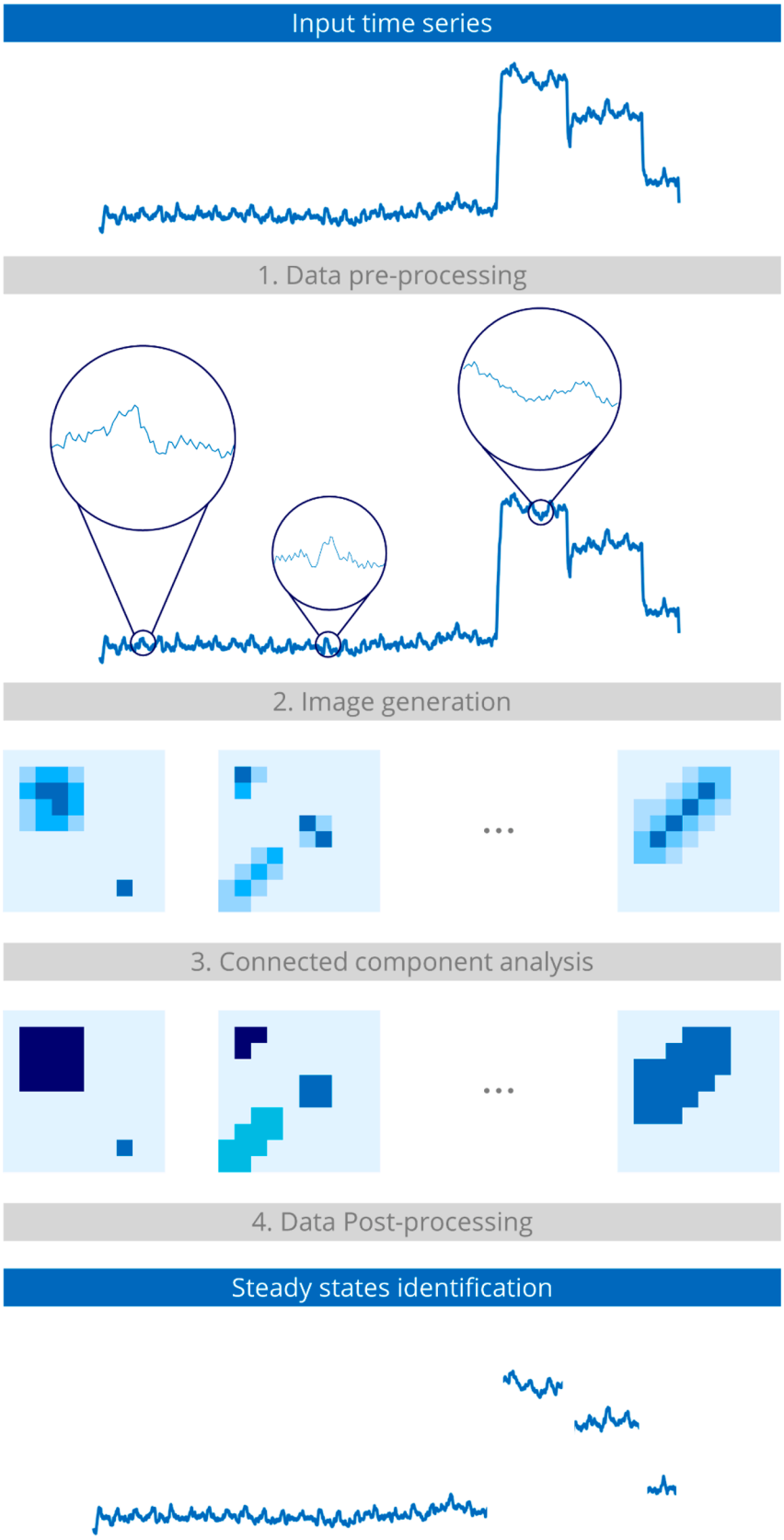

Fig. 1. Graphical representation of the proposed methodology.

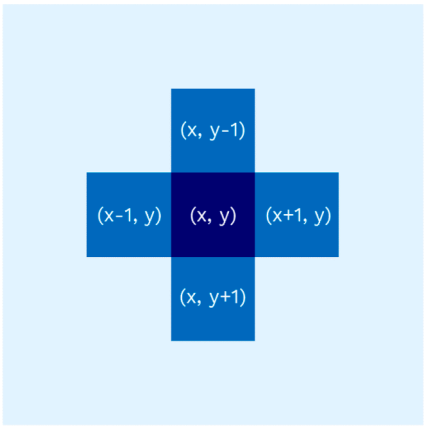

Fig. 2. Graphical representation of the 4-connected neighbourhood.

3.4. Data post-processing

As the results outlined in the preceding section are structured in the form of images, the pixel values of which pertaining to the labels obtained from the connected component analysis, inverse transformation needs to be applied in order to convert such images into distinct time series sequences. Thus, each sequence instance is associated with a temporal label, as different timestamps can be contained in more than one resulting sequence. Thus, such sequences need to be pooled to obtain a unique label per instance of the input time series. To that end, the following approach is applied: if all the temporal labels for a particular instance present the same value, that instance is part of a steady state. Otherwise, if the temporal labels associated with a specific instance differ in regards to their respective values, it is assumed that the instance could not be related to a particular state, and thus such an instance can not be considered for further analysis. A graphical representation of such a process is described in Fig. 4.

4. Case study and results

Having explored the methodology being analysed to identify the different steady states widely observed when dealing with critical marine machinery, a case study is presented to assess its performance. As such, the power parameter collected from a total of three diesel generators of a tanker ship is considered.

The analysed parameter has been collected in a 1-minute frequency and includes more than 65,000 instances in all scenarios. Figs. 5–7 represent graphically the time series of such a parameter for the Diesel Generator 1 (DG1), Diesel Generator 2 (DG2), and Diesel Generator 3 (DG3), respectively. The descriptive statistics are also presented in Table 1.

As observed, distinct typologies of states can be perceived in the time series being analysed. For instance, idle states, transient states, and operational states of machinery are presented. Additionally, various adjustments between operational states can also be distinguished due to either contractual agreements between the charterer and the shipowner in relation to the vessel speed and the fuel oil consumption per day or environmental conditions. Accordingly, all the proposed scenarios can be effectively applied in order to evaluate the performance of the proposed methodology.

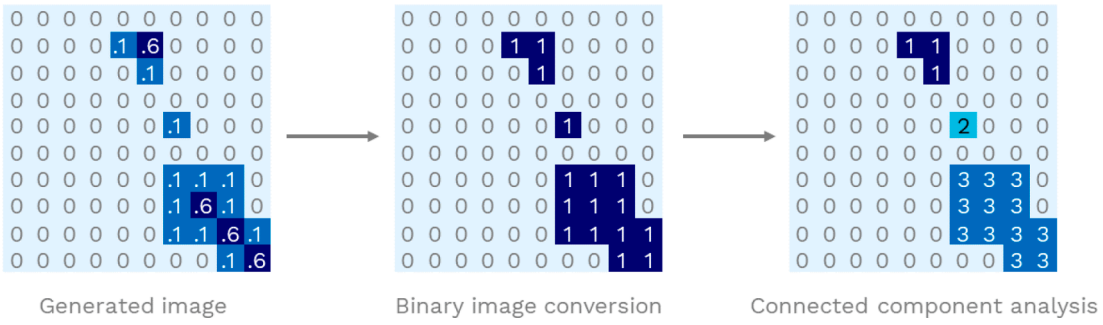

Fig. 3. Connected component analysis phase representation.

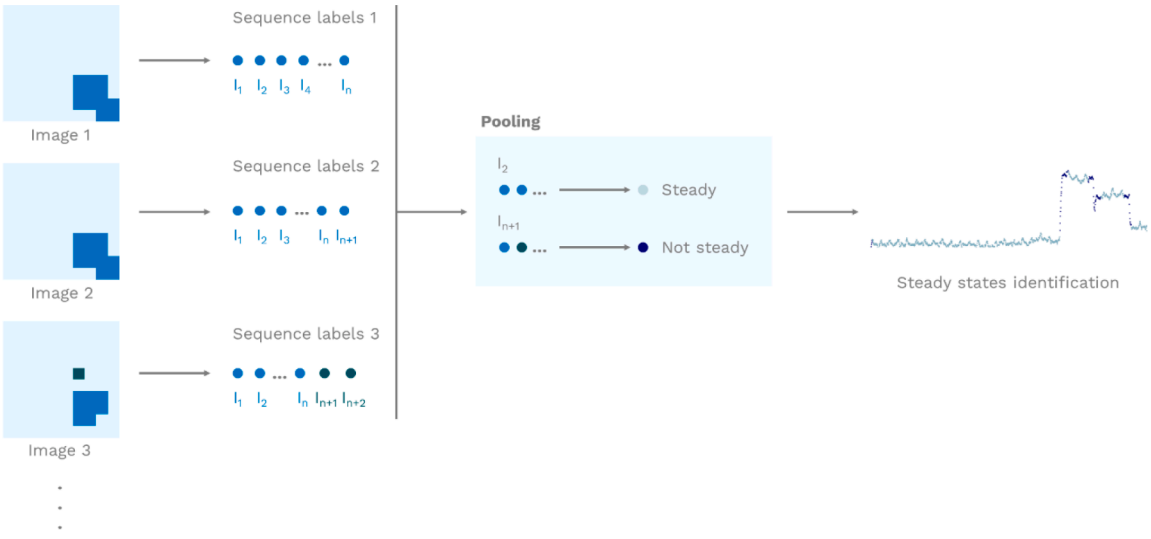

Fig. 4. Graphical representation of the post-processing phase.

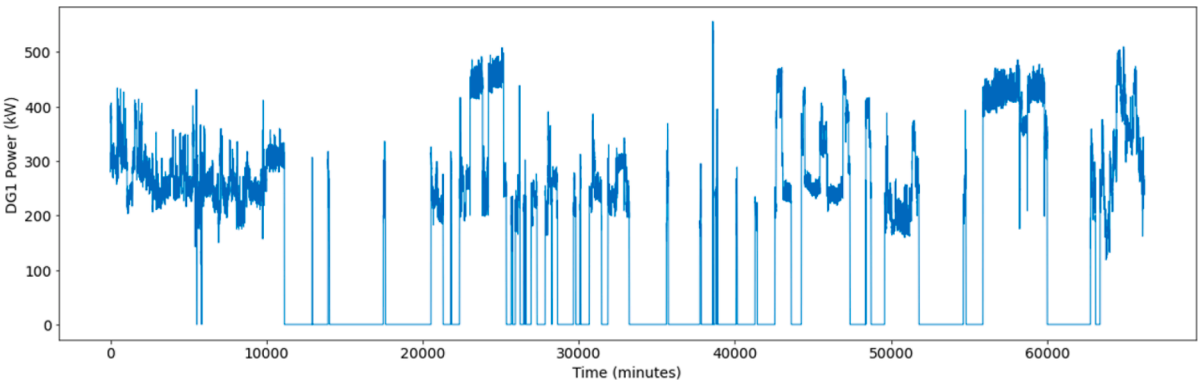

Fig. 5. DG1 power parameter time series plot.

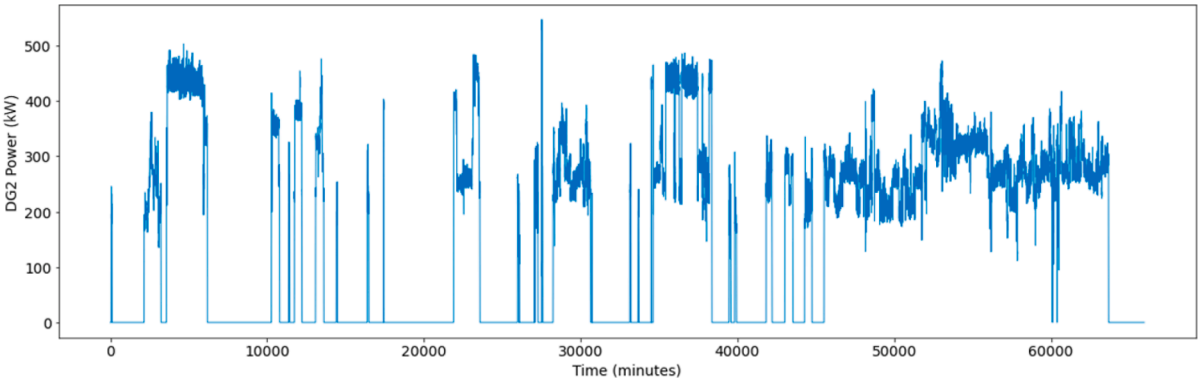

Fig. 6. DG2 power parameter time series plot.

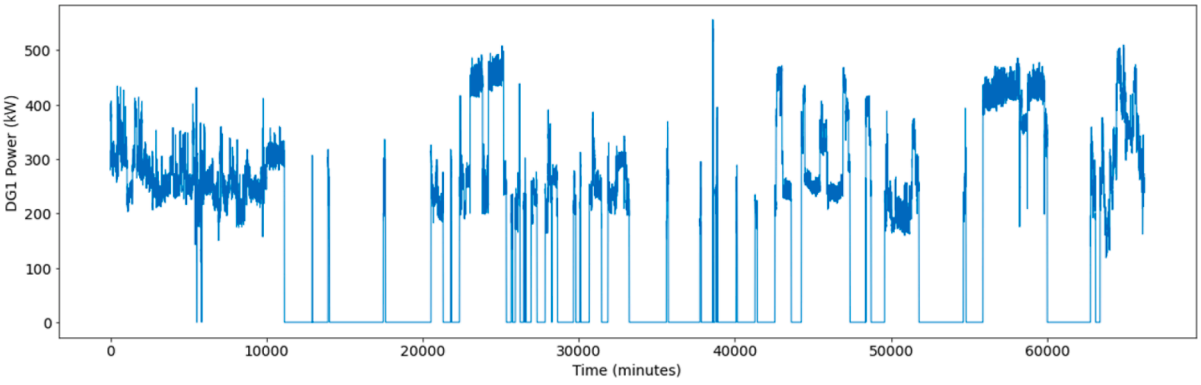

Fig. 7. DG3 power parameter time series plot.

Table 1
Descriptive statistics of the monitored parameters.

	DG1 Power	DG2 Power	DG3 Power
Count	66,207	65,947	65,943
Mean	151.67	151.41	227.24
Std.	159.15	157.56	176.99
Min.	0.00	0.00	0.00
25%	0.00	0.00	0.00
50%	177.95	183.95	261.22
75%	273.30	277.85	373.93
Max.	555.93	546.76	597.86

Prior to the estimation of the transition matrix by the implementation of the first-order Markov chain, the pre-processing step is applied. After analysing different configuration for the effective application of the sliding window algorithm, a time step of 1 and a sequence length of 60 is considered for this enquiry. EWMA is also implemented to reduce the high noise that time series data collected from marine machinery usually contains. Subsequent to the data pre-processing phase, the image generation, the connected component analysis, and the data post-processing steps are applied. Results that outline the identification of the steady states are expressed in Figs. 8–10.

Overall, various parameters and sequences were examined for DG1, DG2, and DG3. However, due to the journal paper extent limitations, the analysis of a total of three sequences including the DG power parameter for each case study is further presented below to visually evaluate and discuss how efficient the performance of the proposed methodology is. In this respect, Table 2 describes the different instances that each sequence contains.

As observed in Fig. 11, the proposed methodology identified

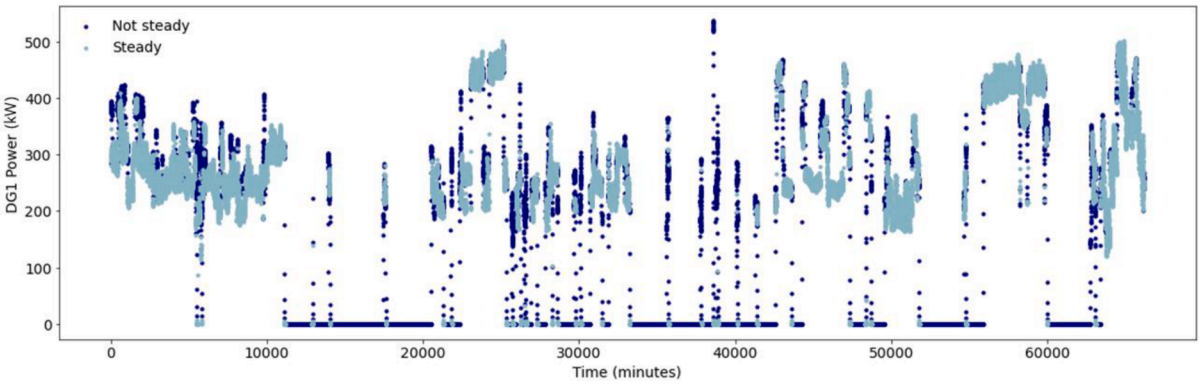

Fig. 8. Steady states identification for the DG1 power parameter.

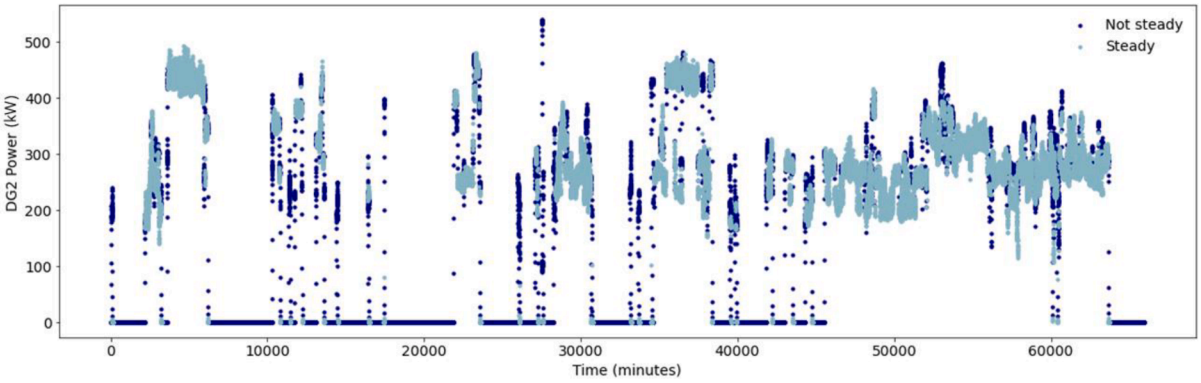

Fig. 9. Steady states identification for the DG2 power parameter.

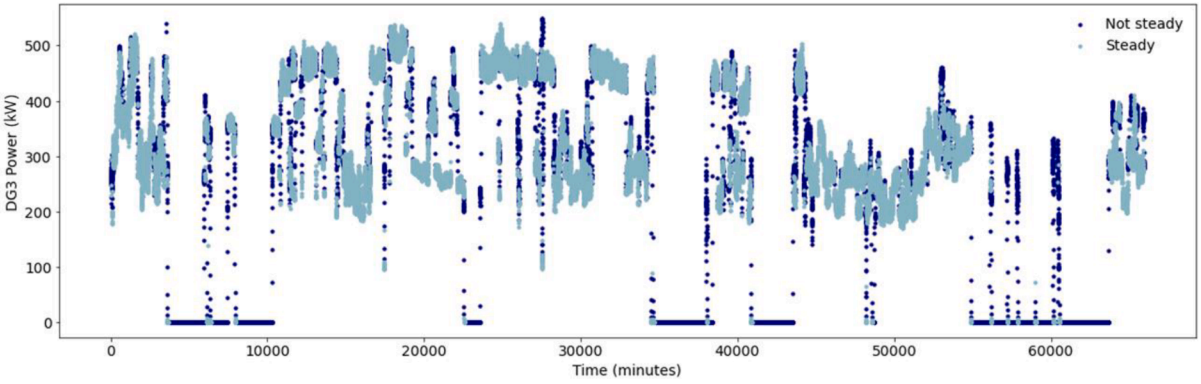

Fig. 10. Steady states identification for the DG3 power parameter.

Table 2
Sequence selection for visual analysis.

	Starting sequence instance	Ending sequence instance	Total of instances in sequence
DG1 sequence	46,000	47,400	1400
DG2 sequence	35,400	37,750	2350
DG3 sequence	24,000	26,500	2500

efficiently the total of four steady states perceived in DG1 sequence. There is a large steady state that initiates at the first instant and persists over half of the recorded time where the values are stabilised between 200 and 300 kW. Subsequently, an abrupt adjustment is observed to enable the transition from the first steady state to the second steady state. Such a transition is adequately identified and labelled as not steady. This state remains for roughly 200 min and then another transition, which is effectively identified, occurs. Subsequent to this transition the third state is achieved, the values of which are stabilised at approximately 300 kW. To achieve the last steady state, an abrupt adjustment is originated, which is adequately identified as not steady. The latter state refers to an idle state, which is usually also identified as not steady, as isolated pixels are also considered to be not steady to avoid repeated values in subsequent analysis. Thus, after steady states identification, a simple filter can be applied to avoid both idle states and states that are mainly constituted by repeated values.

Analogously, the steady states are effectively identified in the remaining sequences (see Fig. 12 and Fig. 13). However, unlike the

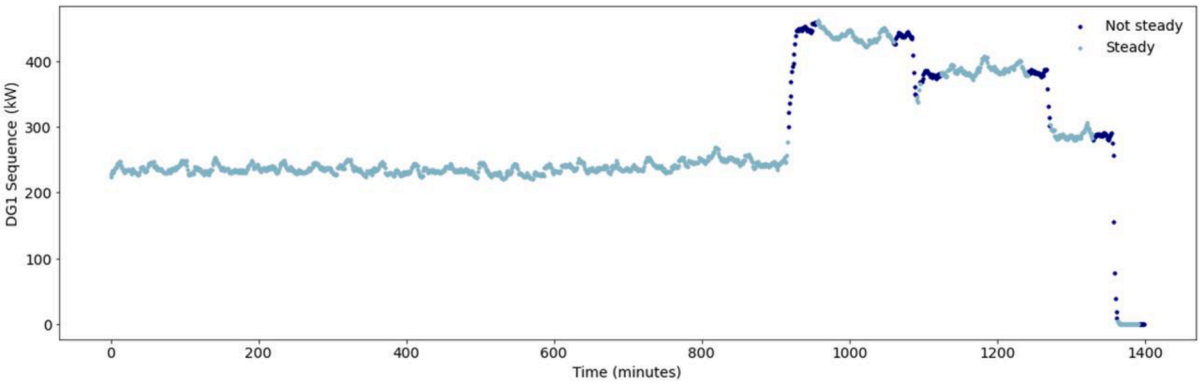

Fig. 11. Steady states identification for DG1 power parameter sequence based on the results obtained from the proposed methodology.

preceding sequence, the current ones present several instances, the typology of which is unclear. Despite their lack of clarity, it can be perceived that such instances are just prior or subsequent to a not steady state. Moreover, the length of these sequences is limited. Therefore, such unclear instances can be automatically identified subsequent to the steady states identification by applying a regular filter.

Additionally, to assess the effectiveness of the proposed methodology a comparative study is performed. To that end, *k*-means and GMMs with EM algorithm are implemented, which have been previously analysed to perform the steady states identification task within the maritime industry. *k*-means was successfully applied by Velasco-Gallego and Lazakis (2021) when dealing with short-term time series data collected from a main marine engine. Such a technique was applied under the assumption of identifying substantial groups of data, the records of each group being analogous to one another but differing significantly with the ones clustered in the remaining ones. To adequately select the optimal number of clusters, both the Silhouette and Davies-Boulding indices are applied to select the most appropriate number of clusters. The outlined results from the application of *k*-means are expressed in Figs. 14–17.

As perceived in Fig. 14, the number of clusters selected in all three scenarios is two, one of them referring to the idle states. The remaining states are clustered altogether in the remaining group. Thus, as observed in Fig. 15, the idle states are adequately differentiated from the remaining ones. However, the distinct operational states contained along the different sequences (see Figs. 15–17) cannot be differentiated amongst them, as these are not considered as particular substantial groups of data by the technique implemented. Moreover, the labelling approach differs from the proposed methodology, and thus the different steady states cannot be automatically detected. Expert knowledge is then required to relabel the outlined labels according to the labels initially defined (steady, and not steady). Due to all these preceding considerations, it is determined that *k*-means is unfeasible to perform the steady states identification task when dealing with long-term time series data. However, various modifications in the framework can be applied to enhance its performance by either implementing the sliding window algorithm to only consider short-term time series data or perform multiple iterations in each of the substantial groups identified.

Analogous results are perceived when implementing the GMMs with EM approach, as it is described in Figs. 18–21. To select the optimal number of components, four different types of covariance are assessed (full, tied, diagonal, and spherical), in accordance with Pedregosa et al. (2011), and a range between 1 and 10 (inclusive) of mixture models are also analysed. For this case, the total number of components selected is three in all cases. Thus, as perceived in Figs. 19–21, the different operational states are more effectively selected than the *k*-means approach. However, unlike the proposed approach, the steady and not steady states are not identified automatically, and thus expert knowledge is required to assess all the identified clusters and determine if they refer to either steady or not steady states. Moreover, the transition between states is not properly defined.

Therefore, the proposed methodology outperformed the other analysed approaches to apply the steady states identification tasks. Unlike the techniques utilised for comparative purposes, the proposed approach can be effectively applied for both long-term and short-term time series data. Furthermore, the steady states are automatically identified and differentiated from other states (e.g., idle states, and transient states), and thus it can be adequately deployed in real time. Accordingly, such an approach can be implemented in a real-time intelligent system for the application of smart maintenance strategies to adequately determine the pertinent instances to be further analysed, as raw data may contain non-operational states that may be irrelevant for further analysis. Hence, if such states are adequately identified and discarded, an increase in both the system performance and the computational efficiency are expected, which will promote an enhancement in

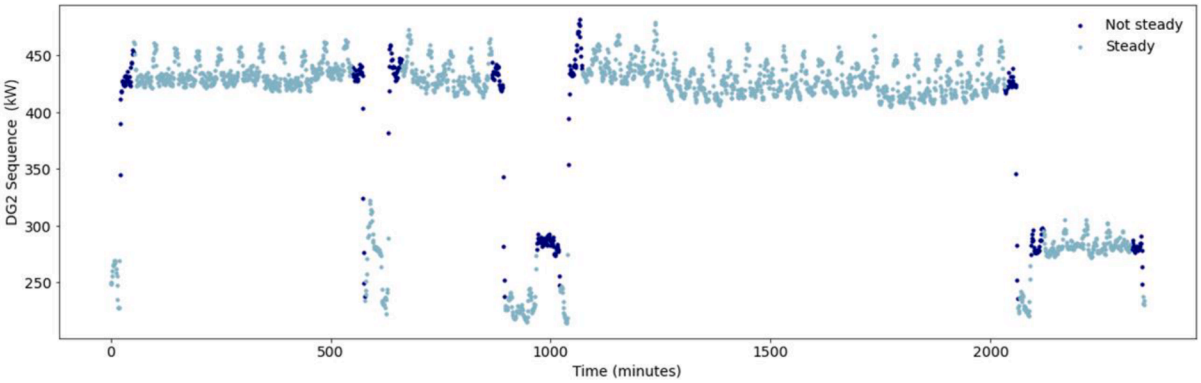

Fig. 12. Steady states identification for DG2 power parameter sequence based on the results obtained from the proposed methodology.

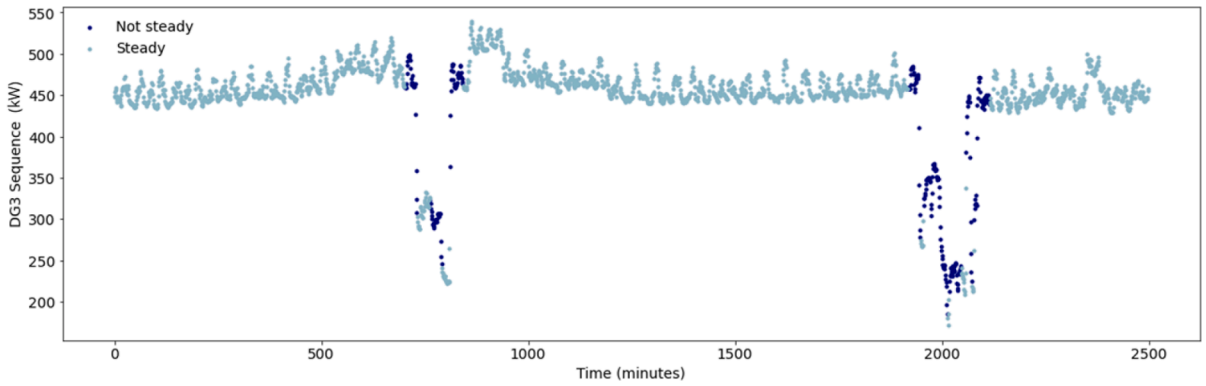

Fig. 13. Steady states identification for DG3 power parameter sequence based on the results obtained from the proposed methodology.

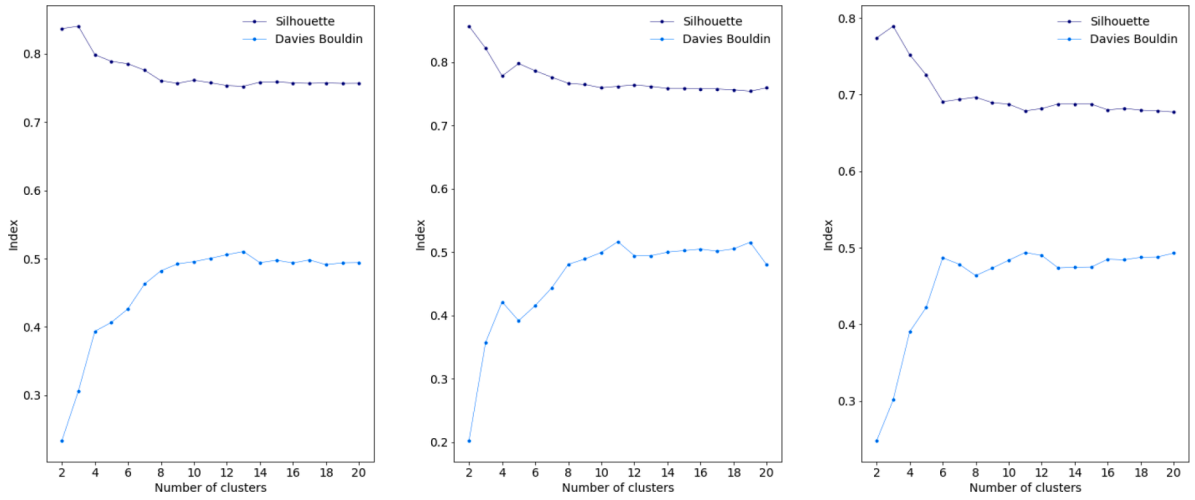

Fig. 14. Estimation of the Silhouette and Davies-Boulding indices for DG1, DG2, and DG3 power parameter, respectively.

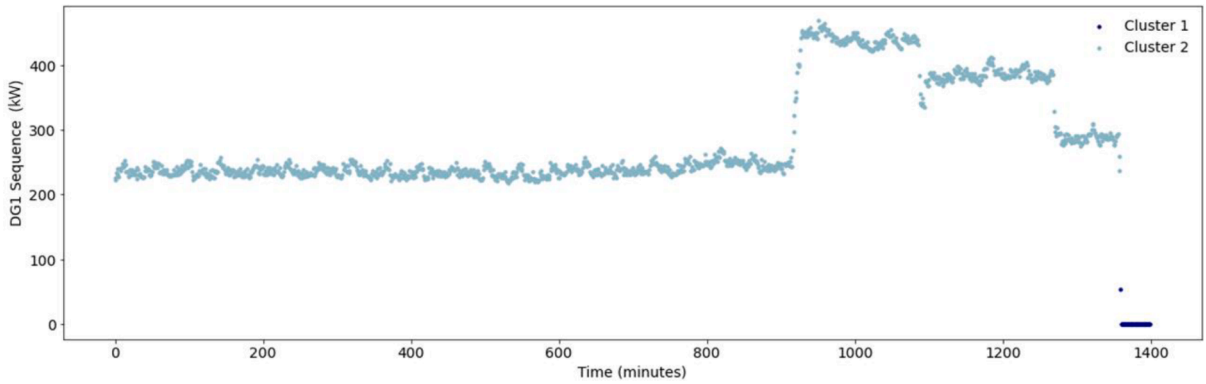

Fig. 15. Steady states identification for DG1 power parameter sequence based on the results obtained from the *k*-means application.

the transportation operations. To continue enhancing the identification of steady states, further research needs to be implemented. Thus, some of the aspects to be considered in the research agenda are listed hereunder.

- Consider optimisation techniques for the selection of the different states of the transition matrix. Although the selection of such states by implementing trial and error has been satisfactory, the authors considered that the identification task can be enhanced by optimally selecting such states.
- Evaluate the implication of different pre-processing steps prior to the implementation of the proposed methodology. For instance, it has been perceived that outliers, repeated values, and noise can have a negative impact in the adequate identification of the states, and thus these need to be adequately addressed precedingly.
- Consider the performance of multiple iterations and the addition of ensemble methods to enhance the outcome of the proposed methodology.
- Analyse a multivariate image generation approach. For this enquiry only the power parameter has been considered. However, the consideration of additional parameters, such as the speed the fuel

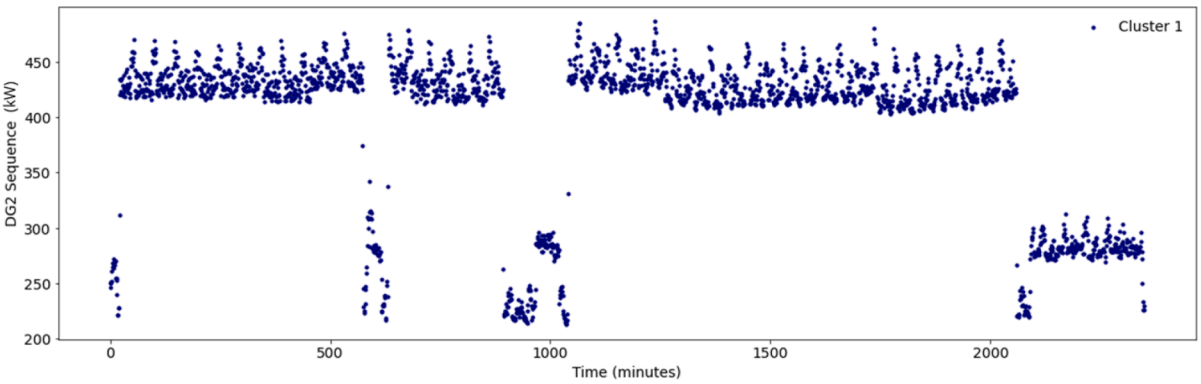

Fig. 16. Steady states identification for DG2 power parameter sequence based on the results obtained from the *k*-means application.

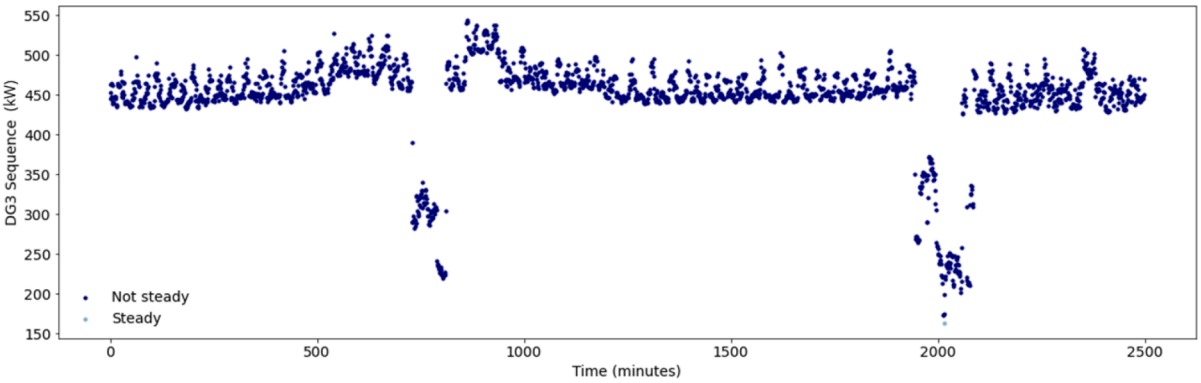

Fig. 17. Steady states identification for DG3 power parameter sequence based on the results obtained from the *k*-means application.

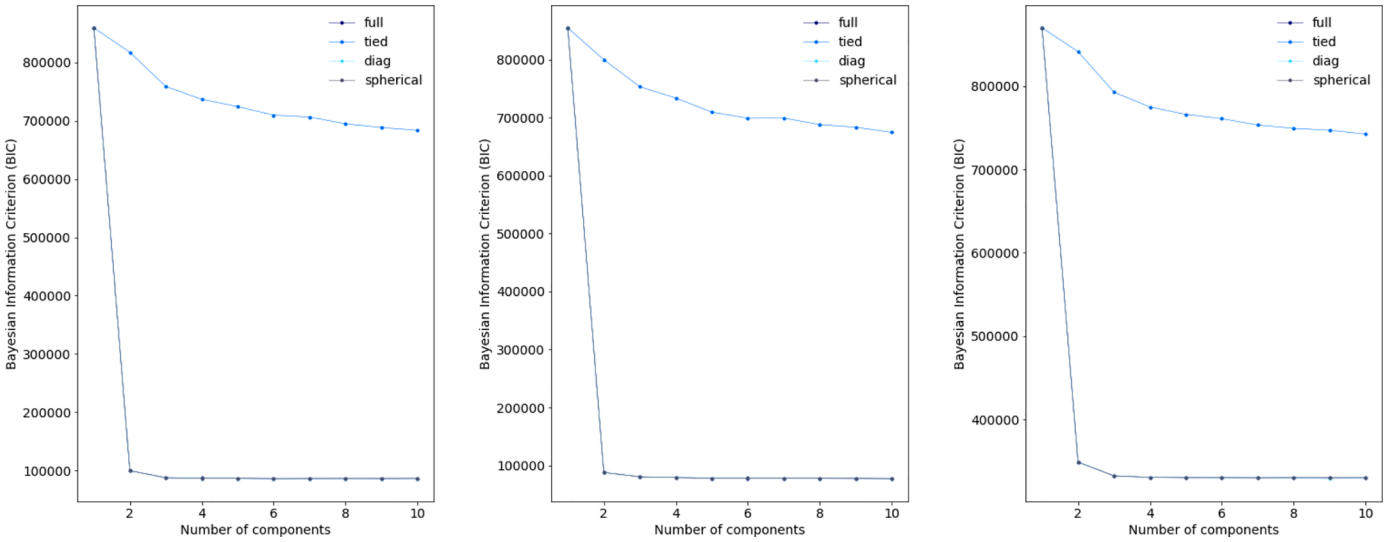

Fig. 18. Selection of the number of components for DG1, DG2, and DG3 power parameter, respectively.

flow rate, or exhaust gas temperature, may need to be included to assess if the identification task is enhanced.

- Apply additional metrics in the validation process. Due to the difficulties in utilising metrics for unsupervised approaches, further research needs to be performed to complement the visual analysis performed in this study with more tangible results.
- Consider more complex pooling methodologies to analyse if the performance effectiveness of the proposed methodology increases.

5. Conclusions

By enhancing data accessibility, the implementation of data-driven models has been possible to empower strategies in relation to O&M activities. However, although the number of studies that consider such models has increased expeditiously within the maritime industry, data pre-processing has not yet been considered as essential, due to a lack of analysis and formalisation in this industrial sector despite the fact several studies promoting the importance of this phase.

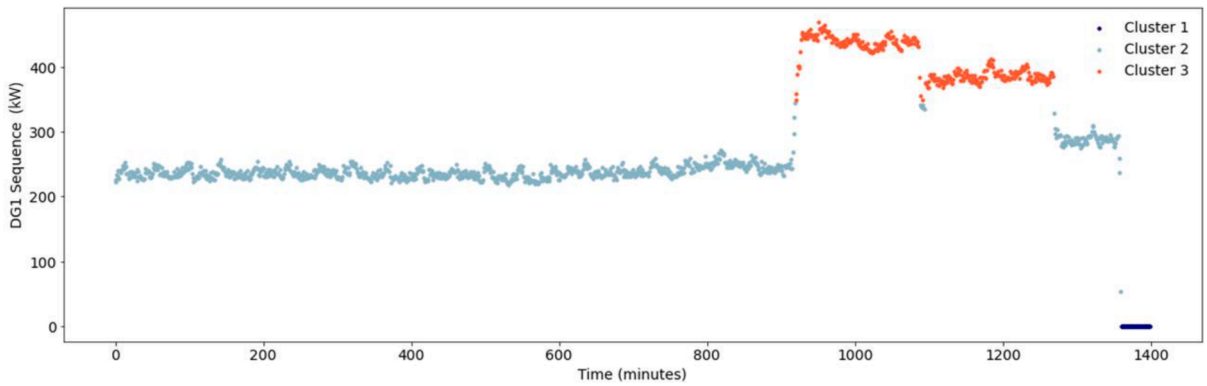

Fig. 19. Steady states identification for DG1 power parameter sequence based on the results obtained from the GMMs with EM algorithm application.

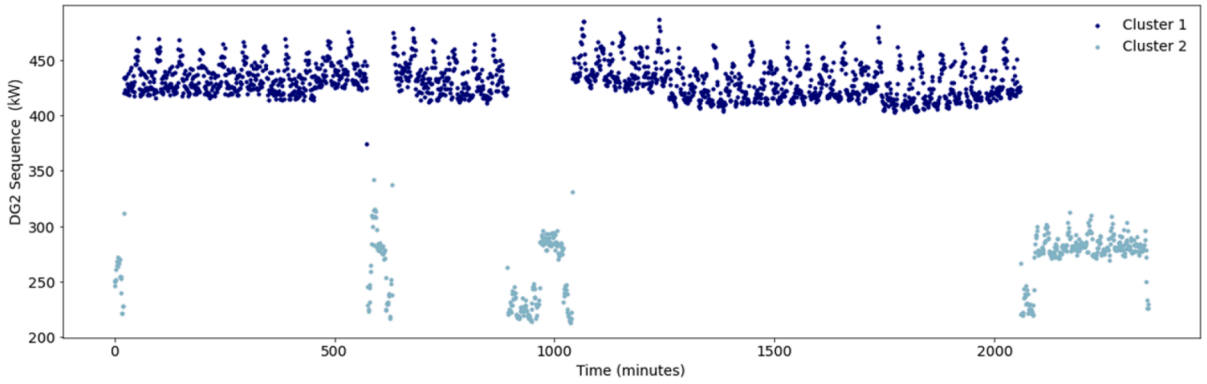

Fig. 20. Steady states identification for DG2 power parameter sequence based on the results obtained from the GMMs with EM algorithm application.

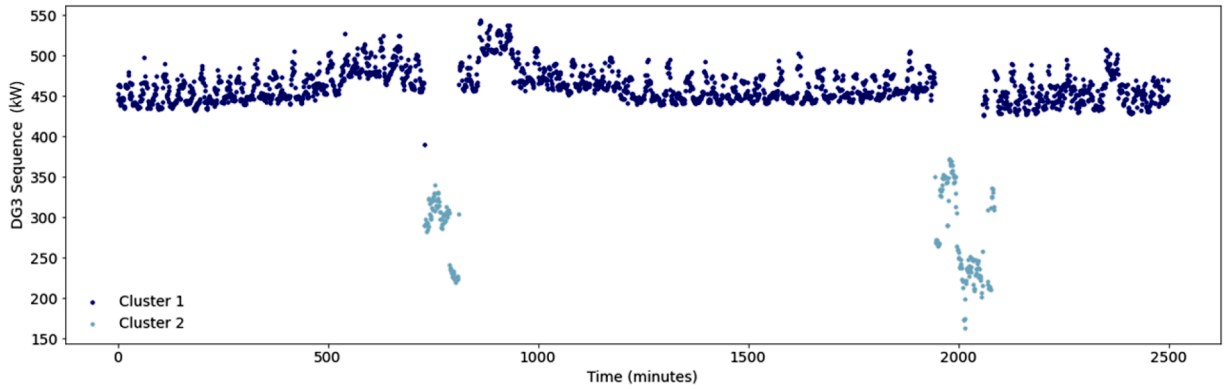

Fig. 21. Steady states identification for DG3 power parameter sequence based on the results obtained from the GMMs with EM algorithm application.

Accordingly, this study presents a novel framework to perform the steady states identification task, which is one of the most fundamental steps of the data pre-processing phase when dealing with critical marine machinery. Such an approach is constituted by two main phases: image generation, which facilitates the transformation of the sequences into images through the estimation of their respective transition matrices by implementing the first-order Markov chain, and connected component analysis in order to determine the different regions identified in each of the images. Both data pre-processing and post-processing were comprehensive in describing the importance of applying such phases.

To highlight the performance of the proposed methodology, a case study was presented, which referred to the power parameter collected from a total of three diesel generators of a tanker ship. Results demonstrated that the proposed methodology outperformed other techniques already analysed for identifying the distinct steady states within the maritime industry. GMMs with the EM algorithm approach presented a more effective performance than the *k*-means technique, although the results of which cannot be utilised in a smart maintenance tool deployed in real time. Moreover, multiple iterations may need to be performed to adequately determine the different steady states. The proposed methodology has demonstrated both its capability of identifying main steady states and its capability of being deployed in real time. However, its performance may decrease when time series contain high noise. Moreover, the optimal selection of the number of states can be critical in ensuring a satisfactory performance of the proposed methodology for such a task. Therefore, to advance towards the implementation of smart maintenance within the shipping sector, some aspects need to be considered in the research agenda, including the analysis of multivariate image generation or the application of additional metrics in the validation process.

CRediT authorship contribution statement

Christian Velasco-Gallego: Conceptualization, Methodology, Software, Formal analysis, Visualization, Validation, Writing – original draft, Writing – review & editing. **Iraklis Lazakis:** Conceptualization, Methodology, Resources, Validation, Writing – review & editing, Supervision.

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.

References

Aslam, S., Michaelides, M.P., Herodotou, H., 2020. Internet of Ships: a Survey on Architectures, Emerging Applications, and Challenges. IEEE Internet of Things J. 7 (10), 9714–9727. https://doi.org/10.1109/JIOT.2020.2993411.

Balcombe, P., Staffel, I., Garcia Kerdan, I., Speirs, J.F., Brandon, N.P., 2021. How can LNG-fuelled ships meet decarbonisation targets? an environmental and economic analysis. Energy 227, 1–12. https://doi.org/10.1016/j.energy.2021.120462.

Brandsæter, A., Vanem, E., Glad, I.K., 2017. Cluster Based Anomaly Detection with Applications in the Maritime Industry. In: 2017 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC), pp. 328–333. https://doi.org/10.1109/SDPC.2017.69.

Brandsæter, A., Vanem, E., Glad, I.K., 2019. Efficient on-line anomaly detection for ship systems in operation. Expert Syst. Appl. 121, 418–437. https://doi.org/10.1016/j.eswa.2018.12.040.

Cheliotis, M., Lazakis, I., 2018. Ship machinery fuzzy condition-based maintenance. In: Proceedings of the 2018 Smart Ship Technology Conference. Royal Institution of Naval Architects.

Cheliotis, M., Gkerekos, C., Lazakis, I., Theotokatos, G., 2019. A novel data condition and performance hybrid imputation method for energy efficient operations of marine systems. Ocean Eng. 188, 1–14. https://doi.org/10.1016/j.oceaneng.2019.106220.

Cheliotis M., Lazakis I., Theotokatos G., 2020. Machine learning and data-driven fault detection for ship systems operations. Ocean Engineering 216, pp. 1–17, doi: https://doi.org/10.1016/j.oceaneng.2020.107968.

Coraddu A., Lim S., Oneto L., Pazouki K., Norman R., Murphy A.J., 2019. A novelty detection approach to diagnosing hull and propeller fouling. Ocean Engineering 176, pp. 65–73, doi: https://doi.org/10.1016/j.oceaneng.2019.01.054.

Dalheim Ø.Ø., Steen S., 2020. Preparation of in-service measurement data for ship operation and performance analysis. Ocean Engineering 212, pp. 1–17, doi: https://doi.org/10.1016/j.oceaneng.2020.107730.

Dalheim, Ø.Ø., Steen, S., 2020b. A computationally efficient method for identification of steady state in time series data from ship monitoring. J. Ocean Eng. Sci. 5, 333–345. https://doi.org/10.1016/j.joes.2020.01.003.

Dong, C., Huang, G., Cheng, G., 2021. Offshore wind can power Canada. Energy. https://doi.org/10.1016/j.energy.2021.121422 (In Press).

Ellefsen, A.L., Bjørlykhaug, E., Æsøy, V., Zhang, H., 2019. An unsupervised reconstruction-based fault detection algorithm for maritime components. IEEE Access 7, 16101–16109. https://doi.org/10.1109/ACCESS.2019.2895394.

Ellefsen, A.L., Han, P., Cheng, X., Holmeset, F.T., Æsøy, V., Zhang, H., 2020. Online Fault Detection in Autonomous Ferries: Using Fault-Type Independent Spectral Anomaly Detection. IEEE Transactions on Instrumentation and Measurement 69, 8216–8225. https://doi.org/10.1109/TIM.2020.2994012.

Emovon, I., Norman, R.A., Murphy, A.J., 2018. Hybrid MCDM based methodology for selecting the optimum maintenance strategy for ship machinery systems. J. Intell. Manuf. 29, 519–531. https://doi.org/10.1007/s10845-015-1133-6.

International Maritime Organization (IMO), 2021. Fourth IMO Greenhouse Gas (GHG) Study 2020 - Full Report. International Maritime Organization, London url. https://www.imo.org/en/OurWork/Environment/Pages/Fourth-IMO-Greenhouse-Gas-Study-2020.aspx.

Jamshidi A., Haijizadeh S., Su Z., Naeimi M., Núñez A., Dollevoet R., De Schutter B., Li Z., 2018. A decision support approach for condition-based maintenance of rails based on big data analysis. Transportation Research Part C: emerging Technologies 95, pp. 185–206, doi: https://doi.org/10.1016/j.trc.2018.07.007.

Lazakis, I., Ölçer, A., 2016. Selection of the best maintenance approach in the maritime industry under fuzzy multiple attributive group decision-making environment. J. Eng. Maritime Environ. 230, 297–309 https://doi.org/10.1177%2F1475090215569819.

Lazakis, I., Raptodimos, Y., Varelas, T., 2018. Predicting ship machinery system condition through analytical reliability tools and artificial neural networks. Ocean Eng. 152, 404–415. https://doi.org/10.1016/j.oceaneng.2017.11.017.

Lazakis, I., Gkerekos, C., Theotokatos, G., 2019. Investigating an SVM-driven, one-class approach to estimating ship systems condition. Ships Offshore Struct. 14 (5), 432–441. https://doi.org/10.1080/17445302.2018.1500189.

Li, X., Kang, Y., Li, F., 2020. Forecasting with time series imaging. Expert Syst. Appl. 160, 1–13. https://doi.org/10.1016/j.eswa.2020.113680.

Lu, Y., Sun, L., Zhang, X., Feng, F., Kang, J., Fu, G., 2018. Condition based maintenance optimization for offshore wind turbine considering opportunities based on neural network approach. Appl. Ocean Res. 74, 69–79. https://doi.org/10.1016/j.apor.2018.02.016.

Pedregosa, F., et al., 2011. Scikit-learn: machine Learning in Python. JMLR 12, 2825–2830 url. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

Perera, L.P., Mo, B., 2016. Data analysis on marine engine operating regions in relation to ship navigation. Ocean Eng. 128, 163–172. https://doi.org/10.1016/j.oceaneng.2016.10.029.

Raptodimos Y., Lazakis I., 2018. Using artificial neural network self-organising map for data clustering of marine engine condition monitoring applications. Ships and Offshore Structures 13, pp. 649–656, doi: https://doi.org/10.1080/17445302.2018.1443694.

Raptodimos Y., Lazakis I., 2019. Application of NARX neural network for predicting marine engine performance parameters. Ships and Offshore Structures 15, pp. 443–452, doi: https://doi.org/10.1080/17445302.2019.1661619.

Su, Z., Jamshidi, A., Núñez, A., Baldi, S., De Schutter, B., 2019. Integrated condition-based track maintenance planning and crew scheduling of railway networks. Transportation Research Part C: Emerging Technologies 105, pp. 359–384. https://doi.org/10.1016/j.trc.2019.05.045.

Tacar Z., Sasaki N., Atlar M., Korkut E., 2020. An investigation into effects of Gate Rudder® system on ship performance as a novel energy-saving and manoeuvring device. Ocean Engineering 218, pp. 1–12, doi: https://doi.org/10.1016/j.oceaneng.2020.108250.

Tan Y., Tian H., Jiang R., Lin Y., Zhang J., 2020. A comparative investigation of data-driven approaches based on one-class classifiers for condition monitoring of marine machinery system. Ocean Engineering 201, pp. 1–12, doi: https://doi.org/10.1016/j.oceaneng.2020.107174.

Taskar, B., Andersen, P., 2020. Benefit of speed reduction for ships in different weather conditions. Transp. Res. Part D: Transport Environ. 85, 1–17. https://doi.org/10.1016/j.trd.2020.102337.

Theotokatos, G., Stoumpos, S., Bolbot, V., Boulougouris, E., 2020. Simulation-based investigation of a marine dual-fuel engine. J. Marine Eng. Technol. 19, 1–13. https://doi.org/10.1080/20464177.2020.1717266.

Velasco-Gallego, C., Lazakis, I., 2020. Real-time data-driven missing data imputation for short-term sensor data of marine systems. A comparative study. Ocean Engineering 218, 1–23. https://doi.org/10.1016/j.oceaneng.2020.108261.

Velasco-Gallego, C., Lazakis, I., 2021. A novel framework for imputing large gaps of missing values from time series sensor data of marine machinery systems. Ships Offshore Struct. https://doi.org/10.1080/17445302.2021.1943850.

Zhu, L., Yu, F.R., Wang, Y., Ning, B., Tang, T., 2019. Big Data Analytics in Intelligent Transportation Systems: a Survey. IEEE Trans. Intell. Transp. Syst. 20, 383–398. https://doi.org/10.1109/TITS.2018.2815678.