Ocean Engineering 212 (2020) 107730

Contents lists available at ScienceDirect

Ocean Engineering

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

Preparation of in-service measurement data for ship operation and performance analysis

Øyvind Øksnes Dalheim *, Sverre Steen

Department of Marine Technology, Norwegian University of Science and Technology (NTNU), Otto Nielsens vei 10, 7491 Trondheim, Norway
Kongsberg Maritime University Technology Centre (UTC) "Ship performance and cyber-physical systems", Trondheim, Norway

ARTICLE INFO

Keywords:
Data preparation
Data preprocessing
Outlier detection
Synchronization
Steady state detection
Data extraction

ABSTRACT

An increasing number of ships are being equipped with sensors and devices for monitoring of operational behavior, and the amount and access to operational data is gradually increasing. Due to various reasons described in this paper, the operational data may contain erroneous data points that are critical to assess prior to performing data analysis or building mathematical and statistical models. In this paper, a stepwise method for preparation of data for ship operation and performance analysis is presented. The method deals with removing jumps in the time series data, including loss of time synchronization between different measurement subsystems, outlier detection, including repeated samples, dropouts and spikes and data selection and extraction, including stationarity detection. The final result is a data set free from disturbances, distortions and undesired physical effects, that can be used to improve the quality of a ship operation and performance analysis.

1. Introduction

Ship operation and performance evaluation based on real ship operational data, is gradually becoming more relevant as the amount and access to operational data improves. The growing amount of operational data is facilitated by an increasing number of ships being equipped with sensors and devices for monitoring the operational behavior. The complexity of the systems varies from simple systems monitoring navigational variables such as position, speed, course and heading, to more sophisticated systems monitoring fuel consumption, propeller torque and/or thrust, propeller rpm, ship motions, rudder/azimuth angle, rotating machinery vibrations, heavy consumers and even some weather parameters. Increasing amount and access to large data sets of ships in operation is however not equivalent to a better understanding of real ship behavior. At least, the researcher should be aware of the data quality, but as the interest in data processing increases, a more systematic view on the methods to handle the data is required.

Over time, new techniques and tools for modeling have been developed, e.g. machine learning implementations such as deep learning, which has changed and improved the capabilities of mathematical and statistical modeling. However, what has not changed at all, being almost a law of nature, is the so-called GIGO — garbage in garbage out (Pyle, 1999, p. 23). That is, the quality of a data analysis will generally reflect the quality of the input data. Good data preparation is therefore essential to practical modeling in the real world (Pyle, 1999, p. 24). A related and widely used term is *data preprocessing*. (Pyle, 1999, p. 112) introduced eight steps of data preparation, from the initial activity of accessing the data to the final process of building the data model. The overall purpose of data preparation is to transform data sets in such a way that the information content is best exposed to the modeling tool (Pyle, 1999, p. 122). In this paper, it is assumed that data is available. This means that data is collected, but equally important that the data is accessible to the user in terms of legal ownership, data format and data connectivity. It is also assumed that the model strategy is selected, i.e. that the user has knowledge of the most suitable models to build for the ship operation and performance analysis. It should however be noted that issues concerning data accessibility are not so rare. Experience shows that lack of standards in terms of e.g. type of sensor, sensor quality, sensor naming schemes, signal meta data information, ship instrumentation and interfaces, data recorders, data communication and data transmission highly limits the accessibility of the data. Lack of standards or simply just ignoring available standards during preparation for in-service monitoring might become particularly challenging in a future signal interpretation, and will require a more careful adaption of methods and procedures to new installations. Scarce documentation of sensor quality with sensor descriptions and signal

* Corresponding author at: Department of Marine Technology, Norwegian University of Science and Technology (NTNU), Otto Nielsens veg 10, 7491 Trondheim, Norway.
E-mail addresses: oyvind.dalheim@ntnu.no (Ø.Ø. Dalheim), sverre.steen@ntnu.no (S. Steen).

https://doi.org/10.1016/j.oceaneng.2020.107730
Received 20 March 2020; Received in revised form 25 June 2020; Accepted 28 June 2020
Available online 2 July 2020

meta data information may lead to the final question whether or not it will be virtually possible to prepare the data for analysis. Along with entire fleets gradually becoming more digitized, standardization of ship instrumentation is fortunately getting more and more attention, including from ship classification societies and international federations with international standards such as ISO15926 (ISO, 2011) and ISO19848 (ISO, 2018). Other relevant tools that have been developed are the Functional Mock-up Interface (FMI), the SFI system and IMOs Common Maritime Data Structure (CMDS) initiative. To further improve the utilization of ship in-service data and to ease the process of preparing vessels for in-service monitoring as well as maintaining sensors and relevant instrumentation, a continued focus towards global standards for in-service monitoring should be maintained. This will further encourage the objectives and ambitions related to the development of a general data preparation toolbox.

Assuming data is collected and accessible, the data preparation starts with selecting the signals that, according to the selected model strategy, contain the required information for the analysis. In literature this is referred to as feature selection or reduction of data width, that is known to have a significant contribution to the overall computational effort. With a subset of features in hand, it is further critical to ensure that the data quality is sufficient, by identification and correction of signal distortions and disturbances. Going from physical ship behavior to a final data sample includes several steps, and the steps can introduce various kinds of disturbances and signal distortions to the data. First of all, each sensor has its individual quality in terms of a range and accuracy specification, that sets the premises for capturing the physical behavior and the level of sensitivity. From the physical sensor or measurement device, the signal enters the transducer, which converts the physical measurement into an electrical signal. The electrical signal is further amplified using a signal amplifier. The signal then goes through an analog-to-digital conversion (ADC), that samples the analog signal from the sensor and converts it to a stream of binary values. The stream of binary values finally enters the data logger. The transducers, amplifiers, ADC and data logger form the components of what is referred to as the data acquisition system (DAQ). All parts of a DAQ can potentially introduce distortions and disturbances to the data (Vaseghi, 2008). The most common distortions and disturbances are excessive instrumentation noise, signal clipping, intermittent noise spikes, temporary signal dropouts, power line pickup and spurious trends (Bendat and Piersol, 2010). In addition to the concern regarding the quality of each individual signal, unphysical relationship between the signals may also be present. This happens particularly if the analysis exploits data originating from multiple subsystems, and is seen as a missing time synchronization or time delay of the signal (Swider and Pedersen, 2017).

Use of operational data for modeling of ship behavior has to consider to what extent *all of the data, all of the time* is suitable for doing the intended analyses. Data depth, or just simply the length of the time series, does not have quite the same impact on computational effort as data width (Pyle, 1999, p. 120). Yet, more important, the subset of data should reflect the relationships that the model tries to analyze. In this paper, depth reduction is referred to as data extraction, indicating that proper time intervals of the time series data are extracted from the complete data set, to form intervals of data suited for the particular analysis of interest. Data extraction is also known as splitting of data. For a wide range of practical purposes for ship operation and performance analysis, the data extraction deals with identification of stationary time intervals. Other special cases can be extraction of port to port trips, extraction of specific operational modes, engine configurations, severity of weather parameters, etc.

It has been found that data preparation generally takes approximately 80% of the total data engineering effort (Zhang et al., 2010), and in the recent literature, particularly in data mining applications, there has been a more focused effort into data preparation. The research emphasizes development of practical techniques and methodologies for data preparation (Zhang et al., 2010). While literature puts effort into data preparation as a complete process in relation to general data mining approaches, there is limited literature on the details of preparing ship monitoring data for analysis. A review of recent literature shows that (Petersen, 2011) summarized some of the most important aspects of the data preparation process for mining of ship operation data, with a short presentation of asynchronously sampled data, missing values and outliers and feature extraction. For further improvements, the authors suggested to study the feature extraction process, as well as the data extraction process considering window size relative to the application, the selected models and features, in more detail.

Hansen (2011) presented an overview of the required steps from data acquisition to statistical analysis of ship performance, including discussion of sampling rate, stationarity, time synchronization and spike removal. The descriptions of sampling rate, stationarity and time synchronization were of a more general type. Regarding spike detection, a more detailed discussion was given, with instructions of two methods to detect spikes in monitored ship data. A considerable amount of material related to fault detection and fault tolerant systems has been published by the control community, for ships particularly related to control of general ship components (engine, propulsion system, rudder, etc.). A fault tolerant system is characterized by inherent fault detection procedures that initiate necessary decisions in order to prevent a further propagation of their effects (Blanke et al., 2015). A virtual example of fault detection developed for shipboard monitoring and decision support systems was presented by Lajic and Nielsen (2009). Nielsen et al. (2012) summarized important findings on fault-tolerant monitoring and decision support systems by using a frequency domain model to detect faults in ship motion variables used in sea state estimation. Rong et al. (2020) developed a method for probabilistic characterization of ship trajectories along a given route that enabled real-time ship trajectory anomaly detection. Fault detection in a control perspective is generally solved under real-time constraints, considering past information up to the current time instant. During analysis of historical data, e.g. in a ship performance analysis, one has the advantage of rather using sets of time series data, which enables delayless filtering of the sensor data.

Swider (2018) introduced the concept of data preprocessing in relation to analysis of power systems onboard ships using monitored data. A thorough introduction to the most common sources of distortions was given, as well as mathematical descriptions of methods for investigating data quality and relation between signals. For data cleaning the concept of digital filtering was presented, but no instructions or practical examples of spike removal, removal of repeated values or zero dropping were given. Methods for time synchronization were presented, and further elaborated in Swider and Pedersen (2017).

In this paper, we present a stepwise method for preparation of onboard monitored operational and navigational data for ship operation and performance analysis. The work is motivated by the increased activity in the use of operational data, combined with a missing general procedure for preparing the data for analysis along with detailed instructions for implementation of the methods. The present literature describes some of the available methods for data preparation. However, it has been found that the descriptions of the methods are incomplete and often exclude explicit guidelines on how the methods should be implemented. In addition, the important discussion considering more of a physical interpretation of the methods as well as how methodical input parameters relate to the data is usually ignored. Using a combination of methods from different sources poses additional challenges with regard to the order in which the methods should be implemented. It is also likely that the methods will be overlapping, or even worse that some parts of the preparation will be incomplete. The main contribution of this paper is therefore to unify the most physical interpretable methods for data preparation in a stepwise procedure, that is easy to implement, that considers in which order the methods should be applied and which assures that the data is compatible between all the relevant steps. The methods for data preparation that are presented are not

only applicable for ships exclusively, but more or less required for all kinds of operational analysis of real world data. Domain knowledge is however beneficial, due to better assessment of data quality and for better control of the input to the data analysis.

2. Preparing data for analysis

The concept of data preparation contains all steps to prepare and preprocess the operational data for its particular analysis. That is, selecting features for analysis, combining data quality assessment with data quality improvement that removes erroneous data regions, and running procedures for data extraction. The aim is to prepare a final data set free from disturbances, distortions and undesired physical effects, that improves the quality of the final results of a ship operation and performance analysis. A laborious quality assessment of input data will generally repay itself in terms of higher quality in the ship operation and performance analysis, and form a basis for building mathematical and statistical models with higher precision.

The description of data preparation is structured according to an order that is recommended by the authors. For the best result of data preparation it is recommended to follow this order during implementation of the methods. The first stage is to select all relevant features. This is followed by identification of missing data and loss of time synchronization for the selected features. After missing or unsynchronized data are identified, the still intact time series intervals are further prepared by identifying and removing erroneous data, referred to as outlier detection. Finally, a suitable practice of data extraction is applied.

2.1. Feature selection

The first stage in data preparation is to select the features that comprise all information required for doing the intended analyses, forming the total set of features $\{f_1, f_2 \dots f_F\}$. Features are the various measurement data that can be extracted from the monitoring system, for instance motion measurements, propeller RPM, wind speed, rudder angle, etc. By selecting particular features, the data analyst determines which data that should be presented to the model (Pyle, 1999). Feature selection is however not to be misinterpreted as feature extraction, whose intention is to map the useful information content into a lower dimensional feature space (Meyer-Bäse and Schmid, 2014). A thorough check of the available signals is advised, as to ensure that all the required information for further use is included, yet limited to the amount of information that actually will be used in order to reduce complexity and following computational effort. A list of available signals, including details of the variables and their units, should preferably be at hand. However, experience shows that such a systematic overview is generally missing and that the documentation of the measurements can be rather insufficient. In that case, the configuration file in the data logger can be informative with respect to descriptions of the signals. Care should always be taken regarding the origin of measurement, its unit and whether the signal is measured or calculated. It is during the stage of feature selection that misinterpretations of measurements are identified, and it is recommended to put effort into this stage.

2.2. Time vector jumps and synchronization

A ship typically has a variety of subsystems onboard, depending on size, complexity, primary purpose and usage etc. This can for example be systems for energy production, propulsion control, fire alarm, main engines, cranes, heating, ventilation and air conditioning (HVAC), navigation, cargo etc. When designing and setting up ship monitoring systems, an important design perspective is how to set up and fetch data from all the sensors installed on the ship. This applies to everything from routing of signals to the sampling rates and internal filters of the measurement variables. When available, a general approach is to fetch data from the main control unit of each system. In the system control unit, multiple signals are assembled together, and the unit may have ports and options for data export. In other cases, data has to be collected directly from each individual sensor.

For ship operation and performance monitoring, the systems and sensors must communicate with a server that samples and stores the data. Unfortunately, in various situations, delays, server latency or other interruptions may occur, that prevents the data samples to be logged in accordance with the system configuration. Other incidents can cause conflicts in the communication channels between the server and the measurement devices. If there is an interruption in the data stream to the data logger, the result can be an inconsistent time series of operational data. In the data set, this is typically revealed as temporary delays or dropouts giving nonuniform spacing of data in time, referred to as time vector jumps. For the individual signal, it is important to identify such jumps in the data prior to applying filters, making time averages and analyzing the frequency content of the signal. When combining several measurement variables for a ship operation and performance analysis, loss of time synchronization between the variables can also be a problem. This happens mainly if the data acquisition is separated into different modules or systems, and/or if various filter frequencies are applied before the data arrives at the data logger. Loss of time synchronization might also happen if a signal having a GPS-based time stamp, i.e. coordinated universal time (UTC), is to be merged with a signal getting its time stamp from the onboard DAQ. This is particularly a concern if the ship experiences shifts in time zones, for which the relative difference between two time vectors might change. It is generally recommended to avoid local time stamps for all kinds of in-service monitoring and data processing, and rather use a fixed reference time stamp such as the unix timestamp. In this way it is made explicit to the data analyst that the time reference is absolute and that no further considerations regarding shift in time zones are necessary. If the time stamp however is given by e.g. the DAQ in a local time format, it is recommended to transform the time vector to unix time prior to any further processing of the data. Complementary information regarding time zone should preferably be used for the transformation. If not available, the time zone can be established from GPS positional data. In the data set, the various losses of time synchronization show as corrupted correlations and unphysical relationships between the measured variables, which is further critical for the building of robust mathematical and statistical models.

Time vector jumps can be identified as outliers in the first order differenced series of the time vector. If the time vector $\mathbf{T} = (t_1, t_2 \dots t_N)$ consists of N discrete timestamps, the first order differenced series $\dot{\mathbf{T}}$ is found as

$$\dot{\mathbf{T}} = \mathbf{T}(i) - \mathbf{T}(i-1) = (t_2 - t_1, t_3 - t_2, \dots, t_N - t_{N-1}),\ i \in \{2 \dots N\} \tag{1}$$

where $\dot{\mathbf{T}}$ has a length of $N-1$. The standard deviation $\sigma_{\dot{\mathbf{T}}}$ of $\dot{\mathbf{T}}$ is expressed in Eq. (3). If a tolerance of k number of standard deviations from the ideal uniformly spaced time vector is accepted, the criteria for which an unintended jump in the time vector takes place between time index i and time index $i+1$ can be mathematically expressed as

$$|\dot{\mathbf{T}}(i) - 1/f_s| > k\sigma_{\dot{\mathbf{T}}} \tag{2}$$

where f_s is the sampling frequency in the DAQ.

$$\sigma_{\dot{\mathbf{T}}} = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N-1}\left(\dot{\mathbf{T}}(i) - \overline{\dot{\mathbf{T}}}\right)^2} \tag{3}$$

In Fig. 1, an example of time vector jump identification is presented using $k = 1$. The time vector used in the example originates from the dynamic positioning (DP) system, which is responsible for collecting signals from the global positioning system (GPS), gyro, wind sensor and the motion reference unit (MRU). Ten situations of time vector jumps larger than $\sigma_{\dot{\mathbf{T}}}$ are identified in this example, as illustrated in Fig. 1. For the ease of implementation, the tolerance of $\mathbf{T}$ relative to a time vector with uniform sampling may also be set as a constant number. A

general suggestion is to use half of the intended sampling interval as the fixed tolerance, e.g. such that $k\sigma_{\mathbf{T}} = \frac{1}{2f_s}$.

When time vector jumps are identified, the corresponding time t_i at which the jump takes place is saved to a matrix $\mathbf{J}$, regardless of whether the time vector leaves or enters a normal state with regular time data. New time vector jumps are appended to this matrix as they are identified, and the identification continues for all the measurement subsystems containing signals of interest. When the complete set of time vectors from all included subsystems are checked for jumps, the elements of $\mathbf{J} \in \mathbb{R}^{m\times 1}$ are sorted in ascending order. $\mathbf{J}$ is finally used to construct time intervals free from time vector jumps in all the signals, by looping through the elements of $\mathbf{J}$. The time intervals free from time vector jumps are saved to a matrix $\mathbf{Q} \in \mathbb{R}^{q\times 2}$, where q is the number of intervals, see Eq. (4). Each interval has a starting (t_1) and ending (t_2) time instant. By defining a minimum duration $t_{\min}$ that all time intervals should meet, a rejection of time intervals shorter than $t_{\min}$ can be included in the loop.

$$\mathbf{Q} = \begin{bmatrix} t_1^1 & t_2^1 \\ t_1^2 & t_2^2 \\ t_1^3 & t_2^3 \\ \vdots & \vdots \\ t_1^{i=q} & t_2^{i=q} \end{bmatrix} \tag{4}$$

After all time intervals are free from time vector jumps, the synchronization of the subsystems should be checked. Synchronization of subsystems means that all signals are mapped to a joint time reference, and that the time reference represents the actual time of each physical event. Prior to the synchronization, it should be known whether the recorded timestamps in the data represent the time of the actual measurement or if they represent the time when the measurement arrives the data logger. In both cases, the time vector of each subsystem can be used to synchronize the data. However, in measurement setups where the timestamps represent the arrival of the measurement to the data logger, a thorough investigation of signal and system delays should be carried out. Low-pass filter frequencies applied to the digital signal is also important to consider, since they induce a delay.

Going from multiple time vectors representative for each measurement subsystem to a joint time reference representative for the complete set of data, implies a selection of the most representative time reference. The following section presents a time reference selection technique for selecting the subsystem that best represents the complete set of systems. First, the summed difference between the time reference arrays $\mathbf{T}(s_i)$ and $\mathbf{T}(s_j)$ of two subsystems s_i and s_j are calculated for all M subsystems. Each sum is organized in a matrix, see Fig. 2. The total time difference between subsystem j and $i = \{1 \dots M\}$ is found by adding up all elements in each column, forming the column sum for subsystem j denoted CS^j shown in Eq. (5). The column sum CS^j is further weighted to a weighted column sum WCS^j by the number of applied signals in the subsystem relative to the total number of applied signals in all subsystems, represented by a count function as in Eq. (6). This step is to avoid excessive amount of interpolation in the data. The next step is to select the most representative time reference for all signals as the subsystem with lowest weighted sum of time differences. This time reference is finally used to establish a modified joint time reference $\hat{\mathbf{T}}$, that has uniform spacing in time, i.e. has a constant sampling interval t_s. This is an important step with reference to an upcoming frequency analysis or digital filtering in the ship performance analysis.

$$\mathrm{CS}^j = \sum_{i=1\dots M} \sum \left[\mathbf{T}(s_j) - \mathbf{T}(s_i)\right] \tag{5}$$

$$\mathrm{WCS}^j = \mathrm{CS}^j \cdot \frac{count(s_j)}{\sum_{i=1\dots M} count(s_i)} \tag{6}$$

The modified joint time reference is established from a first time instant t_{start}, followed by adding the constant sampling interval t_s to the previous time instant, up to the last time instant before a time vector jump takes place. This is repeated for each time interval $i = \{1 \dots q\}$ between the time vector jumps, as given by $\mathbf{Q}$ in Eq. (4). The uniform sampling interval t_s should preferably be the same sampling interval as is configured in the DAQ. The first time instant in each interval (t^i_{start}) needs a careful selection to avoid excessive amount of interpolation in the data. In short, this means that the final time reference should seek maximum overlap with its parent time vector. This can be done by identifying a time shift Δt^i that minimizes the squared difference (S^i) between the original time vector and the modified time vector for the jump free interval i. The minimization problem is

$$S^i = \sum_{k=1} [\mathbf{T}^i_k - \hat{\mathbf{T}}^i_k(\Delta t^i)]^2 \tag{7}$$

$$\frac{\partial S^i}{\partial \Delta t^i} = 0 \tag{8}$$

which gives

$$\Delta t^i = \frac{1}{p^i} \sum_{k=1}^{p^i} [\mathbf{T}^i_k - \mathbf{T}^i_1 - (k-1)t_s] \tag{9}$$

$$t^i_{start} = t^i_1 + \Delta t^i \tag{10}$$

where p^i is the length of the jump free time vector i that runs from t^i_1 to t^i_2, as given in Eq. (4). The first time instant (t^i_{start}) of the modified time vector for the jump free interval i is then given as in Eq. (10), and the corresponding modified time vector is found as

$$\hat{\mathbf{T}}^i = (t^i_{start}, t^i_{start} + t_s, t^i_{start} + 2t_s, \ \dots \ , t^i_{start} + nt_s) \tag{11}$$

$n \in \mathbb{N}$ is the number of elements in $\hat{\mathbf{T}}^i$ and should satisfy Eq. (12). This is to ensure that the elements of $\hat{\mathbf{T}}^i$ is within the jump free interval between t^i_1 and t^i_2. By resampling all the data (free from time vector jumps) to this modified joint time reference, the data set will be completely synchronized with a minimum amount of computational effort, and the introduction of interpolation errors is minimized.

$$n = \left\lfloor \frac{t^i_2 - t^i_{start}}{t_s} \right\rfloor \tag{12}$$

Swider (2018) described the available methods for data resampling as resampling using low-pass filters FIR, comb filter integrating CIC, Lagrange interpolation, spline function and resampling in frequency domain. Due to the optimized selection of the modified time vector $\hat{\mathbf{T}}$, a recommended practice for the data resampling is to simply use linear interpolation or a low order spline interpolation.

Due to various reasons, there are situations where a time vector can be distorted or even unavailable, which means that signal synchronization by use of time vectors is not feasible. In order to synchronize the subsystems without the use of time vectors, it is either required that some variables are measured by more than one subsystem, or that some variables can be combined to establish new complementary variables, which enables use of the cross-correlation function to check for system delays. In Fig. 3, a time series example that shows poor synchronization of two subsystems is given. The example shows the rotational speed of the propeller shaft of a ship (in % of maximum) measured by two different subsystems. By visual inspection, a distinct time delay between the signals is found. Swider and Pedersen (2017) presented a synchronization method that identifies time delays by maximizing the cross-correlation function R_{xy} between two signals x and y. The signals should measure the same physical property, but should be handled by two different subsystems. A straightforward implementation of this method identifies the average time delay for the input signals. Swider and Pedersen (2017) reported that the delay can change over time, for example due to gaps in the data logger, and recommended a split of data into smaller time windows for which the time delay is estimated. The approach of splitting signals into smaller time windows to account

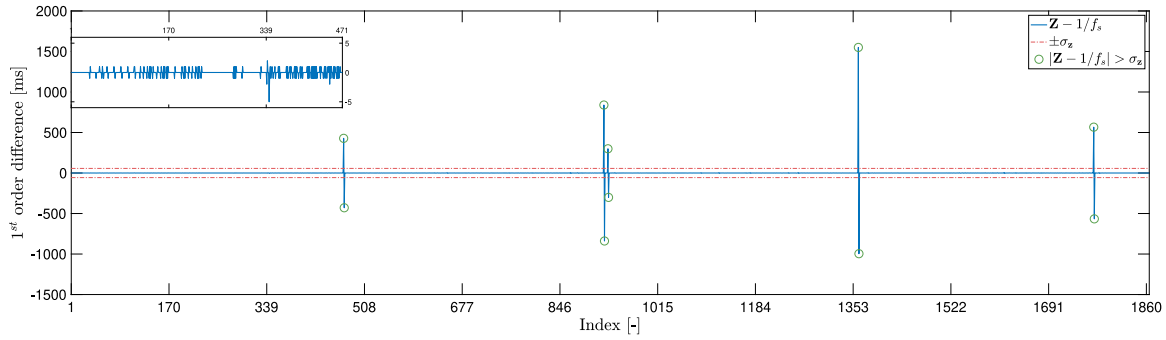


Fig. 1. Identification of data jumps by using the first order differenced series $\dot{\mathbf{T}}$ of the time vector $\mathbf{T}$. In the figure, the value of $1/f_s$ is subtracted from $\dot{\mathbf{T}}$ to show the relative difference between the ideal uniformly spaced time vector with a sample frequency of f_s, and the real sampled time vector $\mathbf{T}$. The upper inset plot shows a detailed view of the small variations in the sampling interval.

	$s_{j=1}$	$s_{j=2}$	$s_{j=3}$	⋯	$s_{j=M}$
$s_{i=1}$	0	$\sum[\mathbf{T}(s_2)-\mathbf{T}(s_1)]$	$\sum[\mathbf{T}(s_3)-\mathbf{T}(s_1)]$	…	$\sum[\mathbf{T}(s_M)-\mathbf{T}(s_1)]$
$s_{i=2}$	$\sum[\mathbf{T}(s_1)-\mathbf{T}(s_2)]$	0	$\sum[\mathbf{T}(s_3)-\mathbf{T}(s_2)]$	…	$\sum[\mathbf{T}(s_M)-\mathbf{T}(s_2)]$
$s_{i=3}$	$\sum[\mathbf{T}(s_1)-\mathbf{T}(s_3)]$	$\sum[\mathbf{T}(s_2)-\mathbf{T}(s_3)]$	0	…	$\sum[\mathbf{T}(s_M)-\mathbf{T}(s_3)]$
⋮	⋮	⋮	⋮	⋱	⋮
$s_{i=M}$	$\sum[\mathbf{T}(s_1)-\mathbf{T}(s_M)]$	$\sum[\mathbf{T}(s_2)-\mathbf{T}(s_M)]$	$\sum[\mathbf{T}(s_3)-\mathbf{T}(s_M)]$	…	0
	CS^1	CS^2	CS^3	…	CS^M
	WCS^1	WCS^2	WCS^3	…	WCS^M

Fig. 2. Selection of best time reference for time synchronization of subsystems.

for time varying delays, is however not necessary when following the time vector jump identification technique presented in this paper. By first identifying the time vector jumps, the complete data set is naturally divided into continuous time intervals where synchronization by the use of the average time delay is suitable.

2.3. Outlier detection

Outlier detection is used to detect, and where appropriate, remove anomalous samples from data (Hodge and Austin, 2004). An outlier can briefly be described as a data point that is not based on a true physical value. Other commonly used notions are "spikes" or "drop-outs", but the main concept is that the values depart from the main modes of variability of the majority of the data (Gervini, 2012). Hereby, when referring to outliers, the term covers all data points that most probably are unphysical. This includes spikes, repeated values and drop-outs, where drop-outs appear as zero, a certain sensor dependent value, or NaN. Examples of spikes, drop-outs and repeated values are shown in a time series of wind anemometer data in Fig. 4, including detailed inset plots that illustrate the nature of outliers identified by visual inspection. For this particular signal, the events of drop-outs and repeated values take place at the same time, and last for approximately ten seconds. Some of the drop-outs appear as zeros, other as a certain (but non-persistent) negative value. Similarly as for the drop-outs, the spikes can appear interchangeably with repeated values, as for example identified in the inset plot corresponding to data around time index 6200 into the time series. In other situations, the spikes behave very different, such as shown in Fig. 5, where spikes in the propeller rpm data measured on the propeller shaft are identified. In this case the spike value, more or less, remains persistent for a short period of time, until the signal starts decreasing towards a value similar to what was the case prior to the spike incident. This behavior is probably caused by signal clipping. Yet, experience shows that the maximum value is not necessarily consistent, and certainly not across various measurements. This complicates the identification of such spike data. In Fig. 5, the inset plot is given to show the true variability of the propeller rpm that is concealed by the scaling of the y-axis.

Due to the various behavior of outliers, there are, unfortunately, no such thing as a universal outlier detector. Various approaches have their pros and cons depending on the data structure, dimensionality, parameter distribution etc. As a coarse classification, Hodge and Austin (2004) separated the fundamentals of outlier detection based on the amount of prior knowledge of data normality and abnormality. From

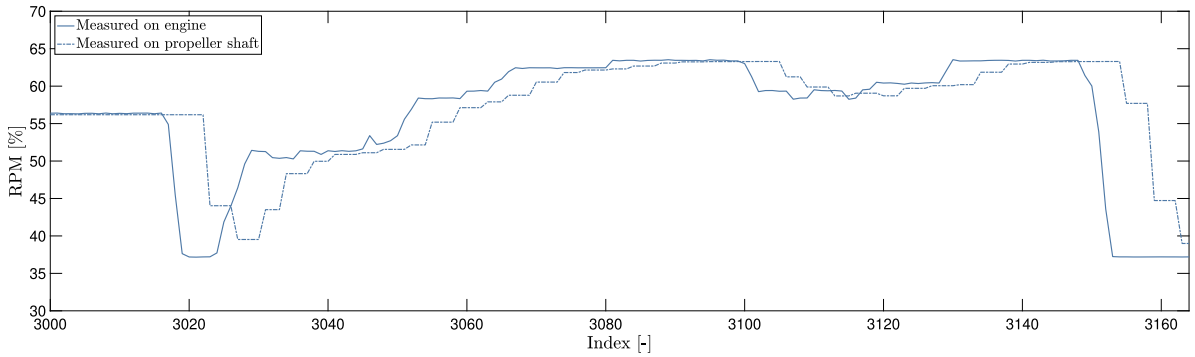


Fig. 3. Measurements of shaft RPM not synchronized.

outlier detection without any prior knowledge of the data, outlier detection having knowledge of normality of the data, to outlier detection having knowledge of both normality and abnormality. For most ship operation and navigational data, knowledge of data normality is generally available as the sensor installations measure predefined physical properties through systems usually having known system configurations. With relevance for onboard ship performance and navigation monitoring systems, Perera (2016) developed a fault detection method considering normality of data through linearization of ship performance and navigation conditions, by the use of principal component analysis (PCA). The method assumed single Gaussian type distributions only. As most performance and navigational data generally are not of single Gaussian type, the method depends on an appropriate pre-clustering of the performance and navigational data.

Detecting outliers without any prior knowledge of the data is less robust than incorporating either or both normality and abnormality. Without information of the normal state, the abnormal data can be hard to separate from the normal data, particularly in data sets having a considerable amount of outliers. One simple way of including information of data normality is by using physical laws and specific domain knowledge of the feature being measured. Based on such domain experience, the maximum and minimum values a process variable can take can be set, from which outliers can be identified. Such outliers are usually referred to as obvious outliers (Qin, 1997).

Yet, there are several types of outliers that obey the physical limits of the measurement variable. Spikes, repeated values and drop-outs might be present within the physical constraints of a variable, but still represent unphysical measurements. Spikes are recognized as data points representing sudden changes relative to the previous data points. In stationary conditions, sudden changes are more or less straightforward to identify. In more time varying conditions such as during heavy weather, ship maneuvering, acceleration and deceleration, course change etc., the spikes can be more troublesome to identify amongst all the data points representing physical variations. Outlier detection should however work under both stationary and non-stationary conditions.

To overcome the complexity of various types of outliers found in a wide range of signals, the method for outlier detection is separated into blocks that are specialized towards identifying a particular nature of outliers. First of all, this simplifies use of domain knowledge for including information on abnormality, as various types of outliers can be identified separately. Secondly, this simplifies adaption to alternative domains and new use of the methods. In each block, the identified outliers are added to an array $\mathbf{O}_{f_j}$, individually for each of the selected features $f_1, f_2 \dots f_F$. This is for assembling the complete set of outliers for each feature variable, as to prepare for an upcoming outlier replacement. The blocks are presented in the succeeding sections. Outlier replacement is described subsequently.

2.3.1. Block 1: Obvious outliers

The first block runs a detection of outliers based on physical constraints. That is, by simply setting a minimum and a maximum value for each measurement variable. The minimum and maximum values should be based on knowledge of the sensor and the physical process being measured. Sensor knowledge deals with the operating range of the sensor, for example the maximum torque a shaft torque sensor can measure. Physical knowledge deals with the possible values a source of measurement can generate, for example the maximum rpm of the engine. Naturally, the engine rpm cannot possibly take negative values, which means that the minimum value for the engine rpm should be zero.

The minimum and maximum values for each feature f_j are set to allow for the full range of possible values. To avoid introducing additional signal clipping, it is recommended to set the minimum and maximum values slightly lower and higher than the actual sensor/physical limits. The time index to the data points that exceed the maximum and minimum limits are saved to $\mathbf{O}_{f_j}$.

2.3.2. Block 2: Repeated values

The second block runs a check for repeated values. Remark that this is relevant for continuous variables only, which rules out measurements that either have been rounded off to a very few number of digits, or measurements that are pre-filtered using e.g. a median filter. Continuous variables can take all values, meaning that a certain value repeating itself might indicate a problem with either the sensor or the DAQ. For that reason, there is doubt whether or not these particular measurements relate to physical behavior.

Repeated values are identified using the first differenced series of each feature. For each synchronized time interval, given as the row element i of matrix $\mathbf{Q}$ in Eq. (4), the first differenced series $\dot{\mathbf{Y}}^i_{f_j}$ of the feature values $\mathbf{Y}^i_{f_j} = (y_1, y_2, y_3 \dots y_n)_{f_j}$ of feature f_j is found as

$$\dot{\mathbf{Y}}^i_{f_j} = (\mathbf{Y}^i_{f_j}(k) - \mathbf{Y}^i_{f_j}(k-1)),\ k \in \{2 \dots (t^i_2 - t^i_1 + 1)\} \tag{13}$$

The time index k to all repeated values for feature f_j are found by searching for $\dot{\mathbf{Y}}^i_{f_j} = 0$. When repeated values are identified, the time index is added to $\mathbf{O}_{f_j}$. Note that because repeated values are identified from the differenced series, each time index added to $\mathbf{O}_{f_j}$ must be increased by 1 before they are added to the outlier array.

Experience has shown that repeated values might as well occur interchangeably with drop-outs or spikes. Examples of this is found in the inset plots in Fig. 4, where a certain value repeats itself immediately after a spike or drop-out. To identify such behavior, the first differenced series, however now with time lag 2, is found for each synchronized time interval. The series is referred to as $\dot{\mathbf{Y}}^{i,k-2}_{f_j}$, and is expressed in Eq. (14).

The time index k to all repeated values for feature f_j are found by searching for $\dot{\mathbf{Y}}^{i,k-2}_{f_j} = 0$. When repeated values are identified, the time index is added to $\mathbf{O}_{f_j}$. Similar as for time lag 1, each time index added

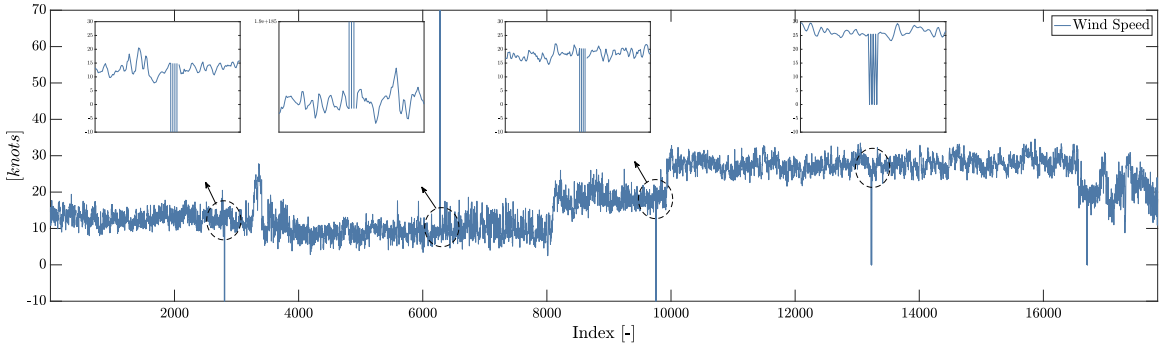


Fig. 4. Spikes, repeated values and drop-outs in wind speed measurements from wind anemometer onboard a platform supply vessel (PSV). Inset plots for detailed inspection for some of the outliers.

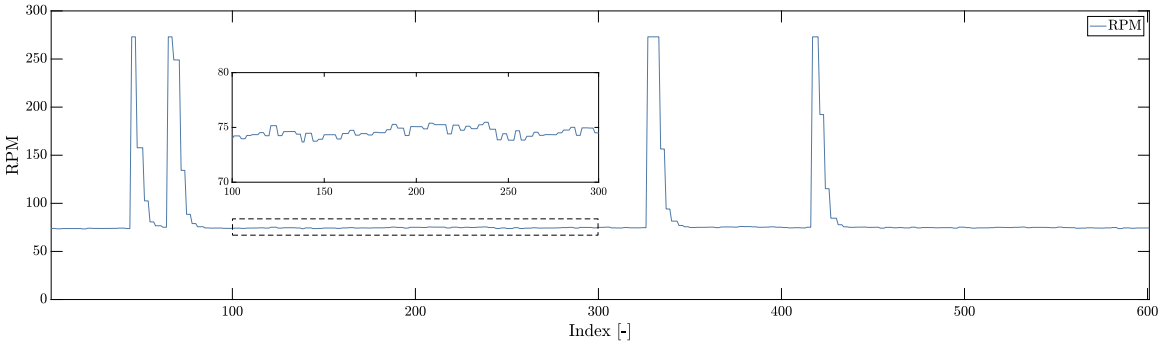


Fig. 5. Spikes in propeller rpm data from shaft sensor onboard the PSV. Inset plot to show the true variability of propeller rpm that is concealed by the large scale on the y-axis.

to $\mathbf{O}_{f_j}$ must be increased by 1 before they are added to the outlier array.

$$\dot{\mathbf{Y}}_{f_j}^{i,k-2} = (\mathbf{Y}_{f_j}^{i}(k) - \mathbf{Y}_{f_j}^{i}(k-2)),\ k \in \{3 \dots (t_2^i - t_1^i + 1)\} \tag{14}$$

2.3.3. Block 3: Drop-outs

Signal drop-outs appear as zero, a certain sensor dependent value, or NaN values. They can be identified by simply searching for a match between the feature values and 0 or NaN. Presence of certain sensor dependent values are not straightforward to identify, as they are individual for the particular feature. However, sensor dependent values taking place subsequently will nevertheless be identified in block 2. Similar as for the other outlier detection blocks, the time indices to all drop-outs should be added to $\mathbf{O}_{f_j}$.

When searching for zeros, one should be aware of the nature of the particular feature. If zero is a likely physical value, marking zeros as outliers should be avoided.

2.3.4. Block 4: Spikes

Spikes are seen as sudden changes towards values either far outside the entirety of the feature data set, or values that significantly deviate from the rest of the data points in a similar context. With respect to time series data, a similar context generally refers to samples taking place in a temporal proximity.

Detection of spikes should work under both stationary and non-stationary conditions. Hansen (2011) presented use of two methods for detecting spikes in data collected by the ship performance monitoring system onboard a PostPanamax container vessel. The first method was a spike detection rule based on evaluating the running mean and the running standard deviation in a defined time frame, to compare a sample with previous and future values. Similar methods are straightforward to implement and mostly effective, but yet best suited for stationary conditions. The second method presented by Hansen (2011) was the *CUSUM* (cumulative sum) test, that incorporates a statistical framework into spike detection. The *CUSUM* test is generally effective in stationary conditions, however, experience with ship monitoring data has indicated some challenges with detection of false positives.

The suggested approach for spike detection, working in both stationary and non-stationary conditions, is by using digital filters. This was also recommended by Swider (2018). In most cases the linear filters are applicable, such as low-pass and band-pass filters. In other cases, nonlinear filters are more suitable such as the median filter. Detecting outliers using digital filters is similar to comparing sample values with a running mean and standard deviation, however, the running mean is replaced by a proper low-pass filtered version of the signal. The standard deviation $\sigma_{\widetilde{Y}}$, see Eq. (15), is evaluated from the frequency content above the low-pass cutoff frequency. That is, from the difference between the base signal $\mathbf{Y}$ and the filtered signal $\mathbf{F}_{\mathbf{Y}}$, expressed as $\widetilde{\mathbf{Y}}$ in Eq. (16). The standard deviation $\sigma_{\widetilde{Y}}$ is used to form a spike detection envelope around $\mathbf{F}_{\mathbf{Y}}$, for which exterior samples are marked as outliers. The width of this envelope controls the spike detection sensitivity. Further control is given to the spike detection by using some constant m representing the number of standard deviations from $\mathbf{F}_{\mathbf{Y}}$ the samples are allowed to vary. In total this forms a spike detection as mathematically expressed in Eq. (17). An example of wind speed data with an envelope of $5\sigma_{\widetilde{Y}}$ ($m = 5$) is shown in Fig. 7. The example is extracted from the same wind speed data shown in Fig. 4, and a low-pass cutoff frequency of 0.1Hz is used for the filtered signal $\mathbf{F}_{\mathbf{Y}}$.

$$\sigma_{\widetilde{Y}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\widetilde{\mathbf{Y}}(i) - \overline{\widetilde{\mathbf{Y}}}\right)^2} \tag{15}$$

$$\widetilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{F}_{\mathbf{Y}} \tag{16}$$

$$|\widetilde{\mathbf{Y}}| > m\sigma_{\widetilde{Y}} \tag{17}$$

The cutoff frequency and the envelope constant m control the spike detection. By lowering the cutoff frequency, more of the higher frequency variations are potentially identified as spikes. The standard deviation $\sigma_{\widetilde{Y}}$ will however naturally increase when the cutoff frequency

is lowered, causing an increased signal envelope, and by that, allowing larger variations relative to $\mathbf{F_Y}$. The envelope constant m has a direct influence on the spike detection in terms of adjusting the accepted distance from $\mathbf{F_Y}$ the signal is allowed to vary. A general recommendation is to set $m \in \{3 \dots 8\}$. More specifically, $m = 5$ is usually sufficient. For each feature variable, the cutoff frequency should be set relative to the expected frequencies of known system dynamics. A general recommendation is to set the cutoff frequency slightly below the highest frequency of known dynamics. Note however that the sampling frequency, and hence the Nyquist frequency, limits the highest frequency that can be observed in the data. To construct an example: a ship sailing at 10 knots in head sea waves with mean wave period of 10 s will get a frequency of encounter around 0.13 Hz. The encountering waves will cause the most high frequent, physical system dynamics in terms of vessel surging. If the DAQ samples speed over ground at a frequency $f_s > 2 \cdot 0.13$ Hz, the surging can be observed in the data. By following our recommendations for spike detection, the cutoff frequency for low-pass filtering should be set slightly below the highest frequency representing physical behavior. A cutoff around 0.10 Hz is therefore convenient. To illustrate how this relates to the ship monitoring data, an example of SOG data and low-pass filtered SOG data with an envelope of $5\sigma_{\tilde{Y}}$ ($m = 5$) is shown in Fig. 6. The filter noticeably removes the high frequency oscillations, while the envelope outlines the variability caused by vessel surging. By using a spike detection rule as expressed in Eq. (17), the envelope will effectively identify spikes as for example can be seen around time index 6290 in Fig. 6.

A second example of outlier detection is shown in Fig. 7 based on wind speed measurements. Fluctuations in a wind field are typically of a more low frequency type compared to waves, with periods reported to range from a couple of minutes up to several hours in the North Sea (Vincent et al., 2011). As wind and waves usually occur together, there might be some variability in the wind speed measurements originating from vessel surging. In situations where this is most prominent, the vessel surging is however usually negligible compared to the wind speeds, so the filtering may rather consider the typical frequencies in the wind field itself. Assuming fluctuations with periods down to three minutes, the cutoff frequency can be set around 5.0E-3 Hz. As shown in Fig. 7, this noticeably removes the high frequency oscillations, while the envelope (using $m = 5$) outlines the more frequent variability. Yet, the envelope will effectively work as a spike identifier as for example can be seen around time index 2500.

Filters require uniform sampled data. Above all, there should be no jumps in the time vector data. It is therefore essential to both run the check for time vector jumps and perform the time synchronization before initiating the block of spike detection.

2.3.5. Outlier replacement

As much as identification of outliers is an important topic in data preparation, it is just as important to consider how to deal with the outliers. For each feature f_j, outliers are identified and saved to $\mathbf{O}_{f_j}$. Because detection of outliers are separated into blocks, duplicates of outliers may be indexed and added to the outlier arrays. As a first step of dealing with the outliers, unique indices should be identified and extracted from $\mathbf{O}_{f_j}$, as well as sorted in ascending order. The unique and sorted indices are written to the modified outlier array $\hat{\mathbf{O}}_{f_j}$.

Outliers may come as individual occurrences and or in clusters of consecutive values (Pyle, 1999, p. 322). This is the reason for adding feature specific outliers to an array during outlier detection rather than performing a direct outlier replacement, as it gives more control of the outliers. $\hat{\mathbf{O}}_{f_j}$ are used to separate between individual outliers and clusters of consecutive outliers, for which data rejection may seem more reasonable than data replacement. Extracting the clusters of consecutive outliers follows a similar strategy as for time vector jumps, by identifying when the first differenced series of $\hat{\mathbf{O}}_{f_j}$, by consistent notation given as $\dot{\mathbf{O}}_{f_j}$, exceeds a certain limit. Because the features have been synchronized using an equally spaced time vector with interval t_s, this limit is nothing else than equal to t_s. Using mathematical notation, all individual outliers or clusters of outliers are separated into outlier intervals by searching for $\dot{\mathbf{O}}_{f_j} > t_s$. The end of the outlier interval (t_2^i) is the time instant in $\mathbf{O}_{f_j}$ for the index in $\dot{\mathbf{O}}_{f_j}$ at which $\dot{\mathbf{O}}_{f_j} > t_s$ is satisfied. The start of the next outlier interval (t_1^{i+1}) is then the subsequent time instant in $\mathbf{O}_{f_j}$. The time indices of the starting point (t_1^i) and the end (t_2^i) of each outlier interval i are saved to the two-column matrix $\mathbf{W}$, as shown in Eq. (18).

$$\mathbf{W} = \begin{bmatrix} t_1^1 & t_2^1 \\ t_1^2 & t_2^2 \\ t_1^3 & t_2^3 \\ \vdots & \vdots \\ t_1^{i=w} & t_2^{i=w} \end{bmatrix} \tag{18}$$

For individual occurrences of outliers or even for small clusters of outliers, a simple and effective outlier replacement seems the most reasonable. For longer clusters of consecutive outliers it might be more reasonable to reject rather than to replace the data. The decision between outlier replacement and rejection should be based on a maximum allowed number of consecutive outliers, that in combination with $\mathbf{W}$ decides which outlier intervals should be rejected and which outlier intervals should be replaced.

The maximum number of consecutive outliers that seem reasonable to replace depends on the application of the data. Using the original resolution of data should limit the accepted number of consecutive outliers compared to using windows of data for which features such as mean and standard deviation are extracted.

The outlier replacement values should not introduce a pattern into the data that is not actually present. The simplest form of outlier replacement is by using linear interpolation between two adjacent data points, next quadratic, cubic or spline interpolation. There should however be reasons to expect a nonlinear variation between consecutive samples before using a higher order interpolation method than linear interpolation. An even more sophisticated approach for outlier replacement is by using a digital filter, for which the pattern and variation in the data can be preserved up to a certain frequency. For individual outliers, a simple linear interpolation is however generally sufficient. For consecutive outliers, yet up to the maximum limit, a low-pass digital filter is recommended for outlier replacement.

Outlier replacement through either interpolation or by using a digital filter have to consider time vector jumps. Outliers should be replaced considering the adjacent measurements only, not by approximations running across a jump in time. Filtering or interpolating across jumps in time can possibly introduce non-existing patterns to the data. It is therefore highly recommended to use the time intervals free from time vector jumps, given by $\mathbf{Q}$ in Eq. (4), to construct the data basis for interpolation or digital filtering.

The replacement values are generated from information that is already present in other measurements, and regardless of replacement method, the estimated values are somewhat smoothed. To preserve some of the variability, noise can be added to the replacement values. In that case, the noise should be based on the existing variability, e.g. by sampling from a Gaussian distribution with zero mean and standard deviation equal to the upfront standard deviation of the measurement variable. For a ship operation and performance analysis, adding noise to the outlier replacement is however generally not important, as data will be averaged over a certain window.

2.4. Data validation

Data validation is a term that may refer to the general process of checking data quality, similar to data preparation, with the aim of detecting faulty measurements to finally arrive at a reliable data set. In this paper, data validation refers to the more stand-alone process

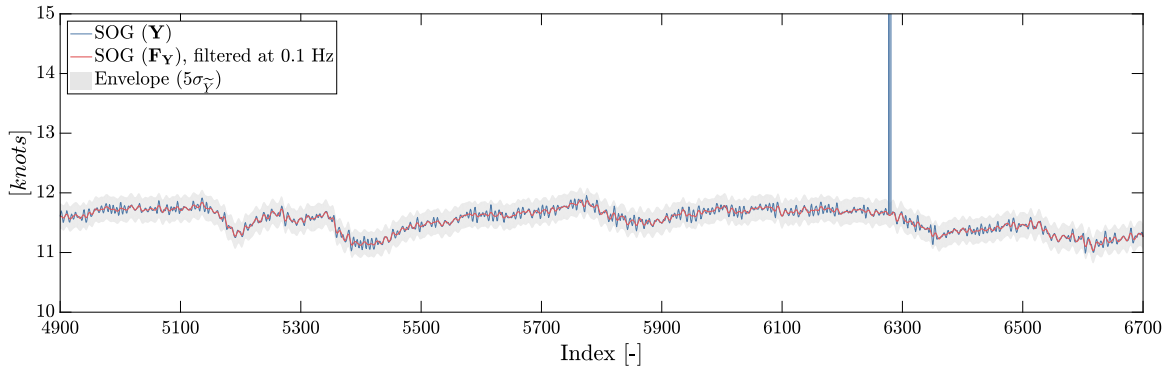


Fig. 6. Extract of speed over ground (SOG) data ($\mathbf{Y}$) with envelope of $\pm 5\sigma_{\tilde{Y}}$ relative to low-pass filtered SOG ($\mathbf{F}_{\mathbf{Y}}$) with cutoff frequency 0.1 Hz.

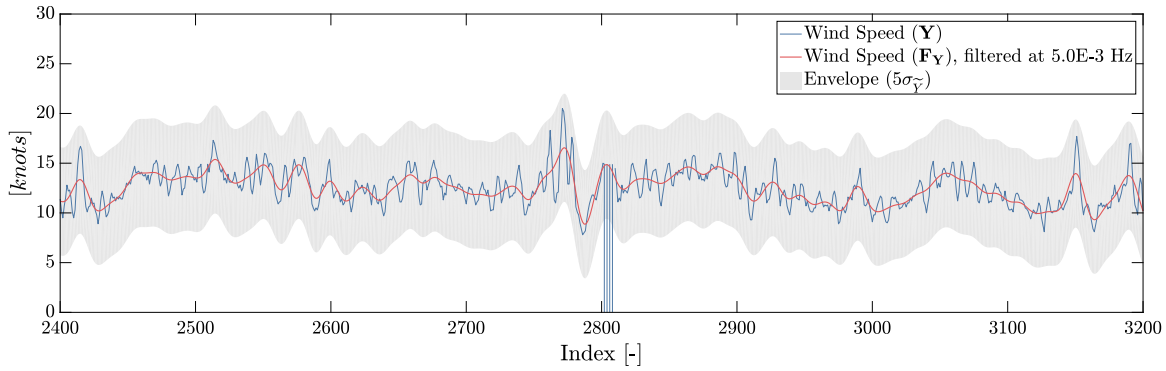


Fig. 7. Extract of wind speed data ($\mathbf{Y}$) with envelope of $\pm 5\sigma_{\tilde{Y}}$ relative to low-pass filtered wind speed ($\mathbf{F}_{\mathbf{Y}}$) with cutoff frequency 5.0E-3 Hz.

of checking whether the measurements are reasonable. Data integrity is another term often used in computer science. However, this is somewhat more related to the integrity in storage, like file system consistency, accessibility of data and avoiding corrupted files (Sivathanu et al., 2005). Data free from outliers, time vector jumps and with complete time synchronization may appear as good quality data during an individual examination. However, seen in relation to other measurement sources the data may be unphysical, for example a ship at service speed having zero rotational speed of the propeller, or a propeller having maximum rotational speed while the engine is standby. There are various reasons for the mismatch between signals, such as miscalibrated sensors, error in routing of signals, sensor drift, various signal disturbances and other hardware or software related malfunctions. Data validation is the final data preparation step to ensure that such unphysical data is rejected as input to the ship performance analysis.

In cases where historical measurement data is available and it is reasonable to believe that this data can be trusted, new measurements can be validated against regression models that are based on similar data. In other cases where no such prior reference to a proper and reliable signal behavior is at hand, there are generally two ways to perform data validation that are relevant for ship monitoring data. The most straightforward way is a direct validation between sensors measuring the same property. A ship may have, largely dependent on ship class, coexistent subsystems such as multiple GPS-units, compasses, motion reference units (MRU) and wind sensors, for which data validity can be checked by directly comparing the similarity of the data. A second way to perform data validation is by connecting measurements through various combinations, both linear and non-linear, to establish new variables from which time series similarity can be evaluated. That is, using physical knowledge and domain knowledge in particular, to calculate properties that additionally are measured by dedicated sensors. In the absence of such dedicated sensors, data validation may even be performed between two calculated properties. For ships in particular, there are numerous possible variable connections, and some examples are listed in Table 1.

The approach to data validation goes through a similarity measure, referred to as D with a complimentary subscript. A number of strategies for time series similarity measures exist, from absolute similarity (e.g. singular vector decomposition, canonical correlation analysis, regression and correlation analysis) to relative similarity measures such as distance measures (e.g. Euclidean distance), correlation measures (e.g. correlation coefficient), or principal component analysis (PCA), Fourier transform, and metric based measures (Lhermitte et al., 2011). The strategies have gained focus within domains such as pattern recognition, climatology and oceanography, serving as decision criterion in several time series clustering and classification techniques, each having their strengths and weaknesses as further discussed in Lhermitte et al. (2011). In data preparation, the aim is to validate the various data sources rather than performing a precise pattern recognition, so a simple strategy for similarity estimation can be accepted.

When measurements are either coexistent or combined to express the same physical property, a perfect time series similarity should have a linear correlation coefficient D_{CC} equal to one. This means that both strength and direction of the linear relationship between the two variables are intact. D_{CC} is however a measure of the linear relationship, and does not evaluate the direct difference in time series values, which means that it is unable to reveal amplitude scaling or amplitude translation. Therefore, the difference in, or preferably the ratio between, the arithmetic average of each corresponding time series should include as a complementary similarity measure. This ratio between averages is referred to as D_{RA}, and is mathematically expressed in Eq. (19). Amplitude scaling may be caused by e.g improper placement of a sensor, or by energy pickup in sensor wiring, while amplitude translation may be a typical result of improper sensor calibration. In Eq. (19), subscript $_1$ and $_2$ of $\mathbf{Y}$ represent feature 1 and 2 for

Table 1
Examples of variable connections for evaluation of time series similarity.

Measured property	Calculated property
Shaft power measured in frequency converter	Product of shaft rotational speed and torque
Propeller rpm measured on propeller shaft	Product of engine rpm and gear ratio
Surge velocity measured by MRU	Variance of speed over ground
Trim angle from depth measurements fore and aft	Average pitch angle measured by MRU
Ship course from GPS	Change of longitude and latitude position

which the ratio of arithmetic average is found. Superscript ^{i} represent the time vector interval index, as given in Eq. (4).

$$D_{RA} = \frac{\overline{\mathbf{Y}}_1^i}{\overline{\mathbf{Y}}_2^i} = \frac{\sum_{t=i_1^i}^{t_2^i} Y_1(t)}{\sum_{t=i_1^i}^{t_2^i} Y_2(t)} \tag{19}$$

If either the linear correlation coefficient or the ratio between the average levels is far from unity, the time series are considered not being similar. This generally indicates that at least one of the measurement sources is not suited for data analysis or mathematical modeling. The appropriate actions to take will depend on certain conditions. In situations where more than two coexistent sensors are available, each one of them can be validated against the others. If this results in a certain measurement source standing out in terms of low data similarity relative to the other coexistent sensors, the particular measurement should me ignored. Similar strategy can be used in cases where multiple combined measurements are available. If however the available coexistent or combined measurement sources limits to two, this kind of strategy is not feasible. The data validity procedure should then rather evaluate which measurement that seems the most reasonable. A simple way of considering reasonableness is to check for obvious errors, for example to check if either of the two measurements have no variation. If no obvious errors can be identified, the suggested approach is to construct a new variable that is the average of the two measurements.

In practice, there might be situations where special attention should be given either D_{CC} or D_{RA}. If the data analysis depends on the mean value of a signal rather than instantaneous values, similarity expressed through D_{RA} should be given more emphasis over D_{CC}. This is e.g. the case when analyzing ship performance, as the impact from individual waves on ship speed and propulsion power should be filtered using time averaging. If the data analysis rather depends on variability, such as standard deviation or a frequency analysis, D_{CC} should be given more emphasis over D_{RA}. A relevant type of analysis is e.g. analysis of vessel motions for shipboard sea state estimation, which recently has gained interest in the literature (Nielsen, 2017). Interpretation of *far* from unity is case specific, so whether a similarity should be accepted or not depends on the monitoring system and each individual sensor. The use of data validation through D_{CC} and D_{RA} may therefore be considered more informative rather than conclusive.

With respect to D_{CC}, there might be need for filtering the data before checking the correlation between measurements. The sensor noise typically varies among a collection of sensors, and this has a direct influence on D_{CC}. In a wide range of data analyses, the high frequency content such as sensor noise is, in any case, irrelevant for the result. Hence, it is not necessarily critical that two coexistent or combined measurements correlate in the sense of a complete frequency resolution. If high frequency content rather should be disregarded, low-pass filtered versions of the measurements should be used for calculating D_{CC}. However, note that identical cutoff frequency must be used for the low-pass filtered measurements entering the calculation of the correlation coefficient.

The validity of a measurement should preferably not be evaluated directly for the complete data set. As a signal disturbance may come and go, or a sensor may perform a self calibration, the data can be valid in parts of the data set. It is therefore recommended to do the data validation in batches of data, e.g. corresponding to the time intervals free from time vector jumps as given in matrix **Q**, see Eq. (4). The time intervals free from time vector jumps are even ready for filtering, as each interval is associated with a regularly spaced time vector.

2.5. Data extraction

Depending on the planned type of data analysis, there are various reasons for extracting parts of the data set. Data extraction is also known as splitting of data or data clustering. The general reason for doing data extraction is that inference of monitored data should consider the conditions during data acquisition, as this basically forms the assumptions for interpreting the data. This is particularly important when building data based mathematical models, as it helps choosing the proper modeling tools and methods, and to ensure that limitations and assumptions in the methods are fully met. Examples of data extraction methods can be port to port trips, which can be used for analyzing performance of weather routing services, fuel consumption and optimal speed in transit. Extraction of specific operational modes, which can be used for assessment of power system design (Swider, 2018), detailed analysis of maneuvers or other special cases of operations as for example dynamic positioning (DP). Extraction of data for which a particular thruster configuration is operative, as for example a thruster running with constant rotational speed (rpm) or constant power, or a propeller running with constant blade pitch.

In a wide range of data analyses, a general requirement is that the data is sampled under stationary conditions. That is, there are no transient behavior in the data, and inference can be made based on a single realization. If statistical parameters are estimated under the assumption of stationarity, while some non-stationary behavior is present, it is likely that the parameter estimates will be biased. For a ship, there are a number of ways that stationarity can be interpreted. Stationarity in ship speed, thrust admission, ship course, ship motions and weather conditions to mention some of them.

As the motivation behind data extraction is to get control of the input data to the various types of analyses, the choice of method to extract data should be selected thereafter. A combination of methods might also be the most reasonable choice for data extraction. In most practical applications, assessment of stationarity in the data is required, at least to some extent. Stationarity is also probably the most tricky requirement when it comes to data extraction. With respect to analysis of ship performance, the quality of the analysis is, among other factors, inversely connected to the amount of transient behavior in the data. A primitive, but fast and interpretable consideration is to study ship performance under a port to port trip. That is, analyzing consumed power or consumed fuel relative to the forward speed without including low speed maneuvering. A further improvement of this analysis is to filter out voluntary changes of operational control variables during the transit, such as forward thrust caused by changes in propeller rpm, or ship heading caused by azimuth or rudder control.

For the purpose of preparing data for a ship performance analysis, the following sections present two methods of data extraction. The first section presents a simple method to extract data acquired during transit between two ports, a so-called trip identifier. The subsequent section presents a computationally efficient method to extract stationary parts of the in-service measurement data, based on the work presented in Dalheim and Steen (2020).

2.5.1. Port to port trips

Many ships operate on a more or less scheduled route when it comes to which ports and destinations they are serving. This typically gives a circular pattern to the anchoring, which in combination with a geospatial mapping tool can be used to identify port departure and arrival. If no geospatial mapping tool is available, the time of arrival and departure can be identified using geofences, or virtual perimeters of a geographical area, around the relevant destinations in combination with the geographical position of the ship. A geofence can notify the point in time when leaving and entering a destination, from which data can labeled as a trip. There are however many ships that have more flexible schedules, which makes use of geofencing more cumbersome, less precise, and by that, less relevant. A more general procedure is therefore to identify trips by using the measured ship speed. During a trip, the ship will usually enter transit mode when low speed maneuvering is completed, and similarly end its transit when initiating low speed maneuvering. The start of a trip in transit mode can therefore be identified when the ship speed exceeds a certain limit, and correspondingly end when the speed goes below this limit. The speed limit should to some extent be set depending on ship type, size and the area of operation, but a forward speed of 4 knots is found to be useful. Exceptions are for ships typically operating at low speed, for example trawlers, for which it is necessary to combine the forward speed with measurements of the force in the trawl, or by using propeller rpm for identification instead of ship speed.

2.5.2. Stationary data

Identification of stationary parts of in-service measurement data is a more complex type of data extraction, yet an essential type of data extraction with respect to performing a ship performance analysis of high quality. It concerns splitting of data into time intervals for which one can assume that certain physical properties are free from transient behavior. In statistics, the strongest form of stationarity is referred to as *strong*, *strict* or *complete* stationarity. A weaker form of a stationarity is referred to as nth order *weakly* stationary, for which all joint moments up to order n exist and are time invariant (Box and Jenkins, 1976, p. 8). Time series analysis often consider second order weakly stationarity, which means that a time series has constant mean and variance. For many practical applications though, the requirements are even less strict, and the condition of *steady state* is rather used. Steady state refers to a condition where the process has a constant mean, and does not require the associated noise and disturbances to be stationary.

Steady state parts of time series data are formed by identifying time points at which certain properties of the time series data change. This is referred to as change-point detection. Dalheim and Steen (2020) developed a change-point detector based on hypothesis testing of the process value inside a moving window. The basic assumption was that the underlying process could be modeled by a deterministic linear trend model, as expressed in Eq. (20). In the equation, a_t refers to a zero mean white noise process with constant variance σ_a^2, b_0 represents the intercept of the linear model and $b_1 t$ the linear deterministic drift component formed by the slope b_1 and the relative time t within the moving window, starting from $t = 0$ for all windows. Under the assumption of independent normal innovations (a_t), the null-hypothesis that the process signal is stationary about the window sample intercept (b_0) can be tested using a two-tailed t-test on $\hat{b}_1$, based on the t-value in Eq. (21). The estimate of the linear slope ($\hat{b}_1$) is found by ordinary least squares estimation.

The actual interpretation of what *is* significant needs to be considered for the particular application of the steady state detector. In general, it concerns identifying the *critical* variables for what we aim to study using steady state models. A thorough interpretation of steady state relevant for a number of ship monitoring data analyses can be found in Dalheim and Steen (2020). The present work limits to identifying data sampled under a constant command of the ship. More specifically, the identification of time intervals at which the propeller rotational speed (rpm) and the propeller pitch angle are kept constant, which is required for the most common types of ship performance modeling.

$$z_t = b_0 + b_1 t + a_t \tag{20}$$

$$t_1 = \frac{\hat{b}_1}{\hat{\sigma}_{b_1}} \tag{21}$$

The steady state detector has two parameters that must be set; the significance level α and the window length n. The significance level controls the accepted slope in the window by representing the probability of rejecting a zero slope ($b_1 = 0$) when a zero slope in fact is true, i.e. the probability of conducting a type I error. If steady state in a process value is critical for the application of the time series data, the steady state detection must ensure a low probability of incorrectly accepting a zero slope. This will increase the reliability in the steady state data, but might at the same time cause a rejection of large parts of the data set, which is unfavorable for data completeness. Similarly, the less critical variables should accept a higher slope in the data window, which means accepting a higher probability of conducting a type II error. The second parameter, the window length, is the number of samples used to form the moving window. This is essential in the consideration of which effects to remove from the data. A long window is e.g well suited for detecting non-stationarity in slow processes, like sensor drift causing the sensor value to accumulate with time. A long window is however not suited for detecting unsteady behavior with short duration. In general, the window length should exceed the auto-correlation persistence of known system dynamics that are acceptable even for a condition of steady state, but yet short enough to detect undesirable changes of short duration.

2.6. Overview of data preparation

The suggested procedures for preparing ship monitoring data for a ship operation and performance analysis have been presented. The entire process of data preparation, from initially possessing roughly raw historical time series data to the completion of a fully utilizable and functional data set ready to analyze, has been split into minor basic parts as a means to offer excellent overview, control and simple customization. The data preparation steps are summarized in the flow diagram given in Fig. 8.

3. Results

The data preparation tools have been developed and evaluated using two separate sets of times series data, each of them corresponding to one complete year of data (365 days). The data is sampled at 1 Hz by the in-service monitoring systems installed on two ships: a platform supply vessel (PSV) and a general cargo/multipurpose vessel (MPV), both designed by Kongsberg Maritime AS. The PSV is mainly serving offshore platforms in the North Sea. It has an overall length of almost 100 m, beam and max draft of 20 m and 7 m respectively, and a dead weight of about 5000 dwt. The vessel is equipped with diesel electric machinery, and azimuthing propulsion system consisting of twin azipull type AZP100CP. The typical service speed of the vessel is about 10–12 knots.

The MPV is mainly carrying cargo along the Norwegian coast. It has an overall length of almost 120 m, beam and max draft of 20 m and 5 m respectively, and a dead weight of about 5000 dwt. The vessel is purely run on liquefied natural gas (LNG) and is fitted with a hybrid shaft generator, rudder and a single screw controllable pitch propeller. The typical service speed of the vessel is about 15 knots.

The in-service monitoring system installed on each of the vessels collect sensor data from selected vessel equipment. The data acquisition is divided into different systems, that provides measurement data to the data logger. The data logger on each vessel is configured to sample at a frequency of 1 Hz. The data logger further provides time stamps to

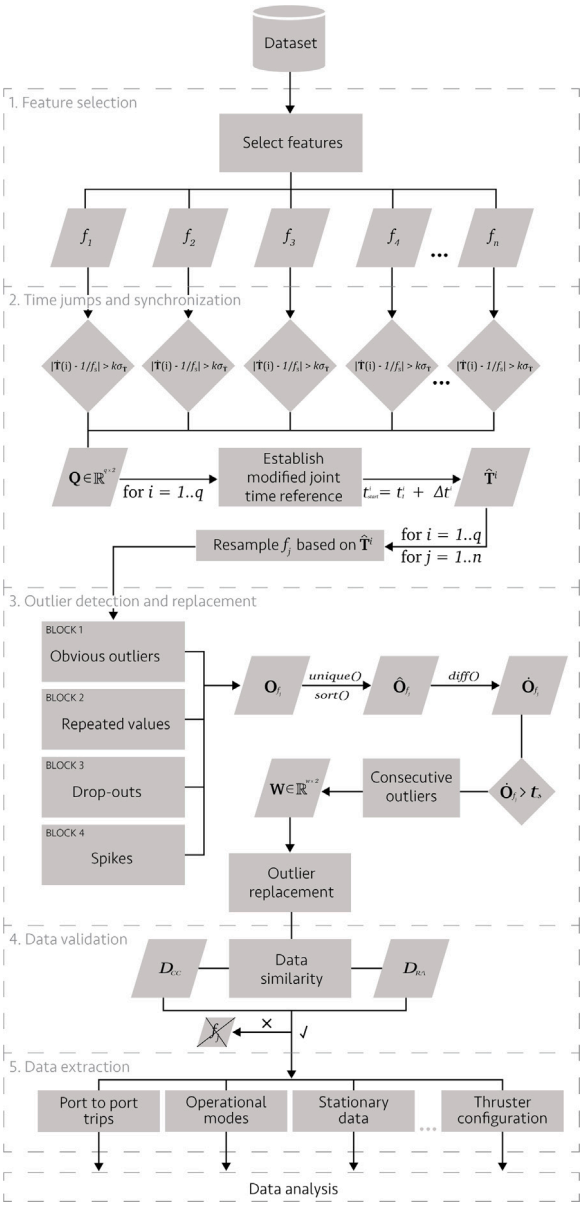


Fig. 8. Flow diagram showing all basic parts of the data preparation steps and how they are related, from data set to data analysis.

the incoming data, which means that each system-based collection of measurement data gets its own time stamp. The data is stored on a server running on the ship, and data transfer to shore is performed on a regular basis.

The following sections evaluate the methods developed for data preparation of ship in-service data, with regard to a ship operation and performance analysis. The evaluation of the methods is structured according to the recommended practice given in Section 2.

3.1. Feature selection

Each of the two case vessels has more than 100 sensors installed, from which measurement data can be extracted. For a ship operation and performance analysis there is a certain, but limited, set of measurement variables that is required for doing a proper analysis. In Table 2, a short list of features selected for doing a ship operation and performance analysis is given. The first column indicates from which system the measurements are acquired from. The second column of the table presents descriptive names of the variables that are measured.

Table 2
Short list of features selected for doing a ship operation and performance analysis.

System	Signals
T-sense optical torque measuring system	Torque, (Thrust) RPM
Motion reference unit	6 dof motion measurements
Global Positioning System (GPS) unit	Speed over ground Latitude, Longitude Course
Doppler SpeedLog	Speed through water
Power management	RPM/Propulsion power Azimuth/Rudder angle, Propeller pitch angle Depth stern, Depth bow
Gyrocompass	Heading
Anemometer	Wind speed Wind direction

3.2. Time vector jumps and synchronization

Evaluation of time vector jump identification is carried out by summing up the total number of time intervals free from time vector jumps, as well as the total number of samples that is removed due to jump identification. In the evaluation, the tolerance of $\mathbf{T}$ relative to a time vector with uniform sampling is set as half of the intended sampling interval (t_s = 1s), e.g. such that $k\sigma_{\dot{\mathbf{T}}}$ = 0.5s with reference to Eq. (2). To avoid unnecessary short time intervals, a minimum duration of 60s is set for the intervals free from time vector jumps.

For the PSV, the one year data set has to be split into 31 time intervals in order to avoid time vector jumps. In terms of amount of data, this means that 5.6% of the data has to be removed, corresponding to 494 h of data. For the MPV, the one year data set has to be split into 541 time intervals in order to avoid time vector jumps. Even though time vector jumps occur more frequently for the MPV, the jumps are generally shorter in length. In terms of amount of data, only 1.0% of the data has to be removed, corresponding to 86 h of data.

The established intervals free from time vector jumps are suited for checking the time synchronization. Because all jumps are removed, maximization of the cross-correlation function can be used directly to identify the average time delay between signals.

3.3. Outlier detection

Outlier detection is evaluated by summing up the total number of outliers identified in each of the two one year data sets. Results are given in Table 3 and Table 4 for the PSV and the MPV respectively. The amount of data identified as outliers relative to the total amount of data is given in the rightmost columns in the tables.

For the PSV, the amount of outliers in the data is generally below 1%, except for the longitude position of the vessel for which outliers constitutes 1.02% of the one year data set. Similar amounts of outliers are detected for the MPV, i.e. mostly less than 1%. The propeller shaft thrust measurement however reveals a large amount of outliers, amounting to nearly 20% of the data set. The main cause of this is found to be drop-outs to negative values, where each drop-out has a significant time persistence. Next, the rudder angle measurement has a considerable amount of outliers with nearly 8% of the data set identified as outliers. Similar as for the thrust measurement, the outliers are mainly drop-outs to either a large or to a small rudder angle, with significant time persistence.

During the recent outlier detection, no limit for the maximum number of consecutive outliers that decides upon data replacement or data rejection was set. The quantities listed in Tables 3 and 4 are hence referring to the amount of data that should either be replaced or rejected for further use.

Table 3
Outliers identified in the one year data set of the PSV.

Measurement variable	Number of outliers [−]	Out of one year [%]
Speed over ground	41 593	.136
Speed through water	32 759	.107
Latitude	227 584	.745
Longitude	311 040	1.019
Course	23 602	.077
Heading	43 645	.143
Wind speed	98 411	.322
Wind direction	264 534	.866
Pitch motion	41 525	.136
Roll motion	43 716	.143
Azimuth angle port side	564	.002
Azimuth angle starboard	600	.002
Shaft rpm port side	645 863	2.115
Shaft rpm starboard	574 308	1.881
Shaft torque port side	305 105	.999
Shaft torque starboard	293 210	.960
Pitch angle port side	6 828	.022
Pitch angle starboard	6 326	.021
Depth bow	2 916	.010
Depth stern	4 278	.014

Table 4
Outliers identified in the one year data set of the MPV.

Measurement variable	Number of outliers [−]	Out of one year [%]
Speed over ground	110 148	.353
Speed through water	132 030	.423
Latitude	4 254	.014
Longitude	62 778	.201
Course	49 527	.159
Wind speed	579 008	1.854
Wind direction	161 928	.518
Pitch motion	654	.002
Roll motion	3 238	.010
Rudder angle	2 480 484	7.942
Shaft rpm	41 955	.134
Shaft torque	63 630	.204
Shaft thrust	6 146 287	19.678
Pitch angle	2 311	.007
Depth bow	299 719	.960
Depth stern	73 096	.234

3.4. Data validation

To demonstrate and evaluate the method for data validation, two examples of time series data showing poor data similarity in terms of either the linear correlation coefficient (D_{CC}) or the ratio between averages (D_{RA}) are shown. The corresponding similarity measures are given in Table 5 for each of the two examples. The first example (Fig. 9) shows the wind speed measured by two coexistent wind anemometers onboard the PSV. The value of D_{RA} indicates that the two coexistent measurements are not equal in terms of their mean value ($D_{RA} = 78\%$). In addition, D_{CC} indicates a linear correlation below unity ($D_{CC} = 92\%$). Considering the time series data in Fig. 9, the reduced linear correlation is apparently caused by a stronger drop in wind speed measured by sensor 1 compared to sensor 2 about halfway into the time series. The second example (Fig. 10) shows the roll angle measured by two coexistent MRUs onboard the PSV. The value of D_{RA} indicates that the two coexistent measurements are far from equal in terms of their mean value ($D_{RA} = 330\%$). The linear correlation is however very close to unity ($D_{CC} = 98\%$). This behavior is also evident in the figure, showing that the two coexistent measurements follow each other, however at two different mean levels.

3.5. Data extraction

Evaluation of data extraction is carried out by presenting various examples of time series intervals formed by applying the data extraction methods as presented in Section 2.5. First, a simple port to port trip identification is shown for each of the two case vessels. Then, examples of steady state identification are shown using various data from the two case vessels.

Table 5
Calculated time series similarity based on coexistent sensors onboard the PSV.

Measurement	D_{CC}	D_{RA}
Wind speed (anemometer)	0.920	0.780
Roll angle (MRU)	0.982	3.302

3.5.1. Port to port trips

An example of trip identification using speed over ground measured on the PSV is shown in Fig. 11. The speed limit is set to 4 knots. The identified trip has a duration of approximately 8.5 h, and consists mainly of transit operation at service speed of 12–14 knots. In Fig. 12 a similar trip identification for the MPV is shown. The identified trip has a duration of approximately 19.5 h, and consists mainly of transit operation at service speed of 13–15 knots. In both cases, the time series data before and after the identified trip is found to be mainly low speed maneuvering of short duration. Trip identification based on speed over ground measurements is a very primitive tool to extract data of transit operation. Yet, it is straightforward to implement and gives a fast and robust splitting of the data into trips.

3.5.2. Stationary data

Data extraction based on identification of steady state time intervals has been tested and evaluated using time series data from the PSV. Fig. 13 presents three examples of steady state identification on ship control variables that are critical for a ship performance analysis, i.e. the propeller rpm, the propeller pitch and the ship heading. The time series data of the propeller rpm and the propeller pitch are extracted from the same time interval. This interval takes place in between two time vector jumps and lasts for about 100 min. The ship heading data is extracted from a separate time interval, lasting for about 210 min. The reason for presenting a separate time interval is simply because the ship heading was constant throughout the time interval used for the propeller rpm and pitch. The color in the figures indicates the local state identification, with green color representing the time intervals where the particular variable probably is at steady state. The three measurement variables are all checked for steady state using a significance level $\alpha = 1\%$, but with individually selected window lengths. For the propeller rpm measurement a window of 5 min ($n = 300$) is used. Evaluating the result presented in Fig. 13(a) shows that the method successfully rejects the two parts of the time series that by visual inspection clearly is not at steady state, more specifically the rpm drop and rebuild between time ≈ 18 and 32 min and between time ≈ 40 and 46 min into the time series. The propeller pitch angle is checked for steady state using a window of 30 s ($n = 30$). The short window length is set because changes in propeller pitch angle generally occur more frequently and are of shorter duration compared to the propeller rpm. The result is presented in Fig. 13(b), showing how the propeller pitch angle (measured in % of maximum pitch angle) is reduced and increased between time ≈18 and 48 min into the time series. The steady state detection rejects most parts of this data. However, due to the small window, there are even some short time intervals in between that are accepted as steady state, e.g. between time ≈ 22 and 27 min. The third example of steady state identification considers the ship heading using $\alpha = 1\%$ and a window of 20 min ($n = 1200$). The long window is set because a change in ship heading generally is a slow process, expected to occur infrequently during a traditional transit operation. The result is shown in Fig. 13(c), and agrees well with visual inspection. Short intervals of apparently steady state data have been rejected due to non-sufficient length compared to the window size. The small but distinct changes in heading are successfully rejected as steady state, for example as shown between time ≈ 30 and 50 min.

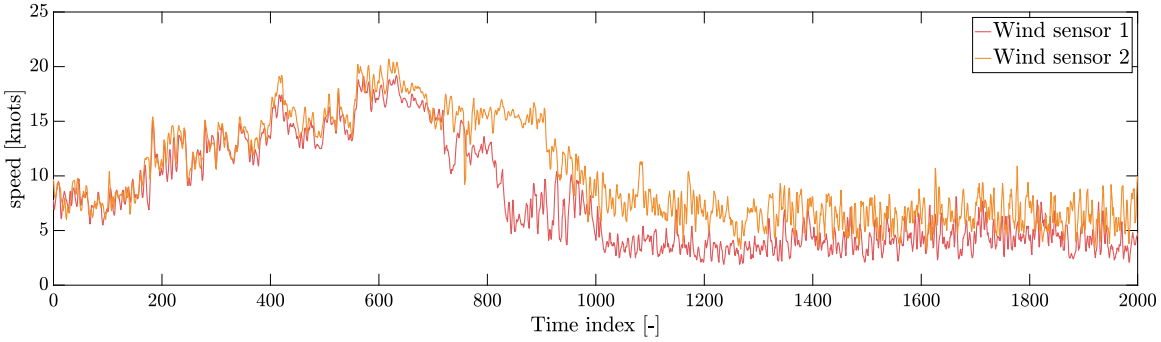


Fig. 9. Data validation of wind speed measured by two coexistent anemometers onboard the PSV.

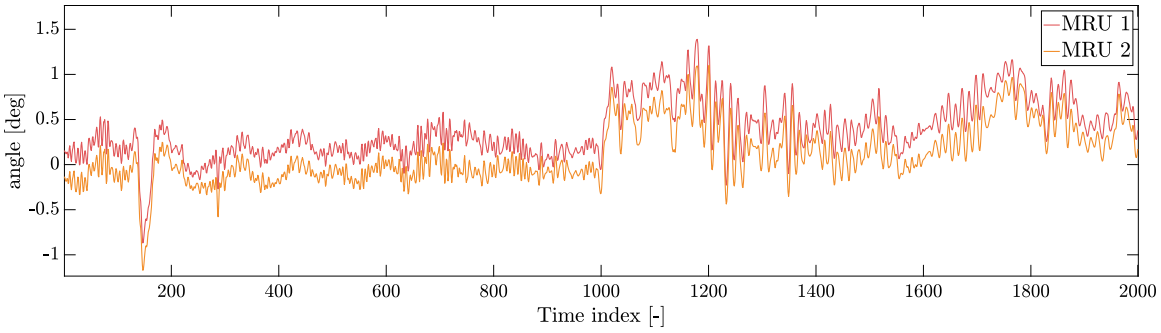


Fig. 10. Data validation of roll angle measured by two coexistent motion reference units (MRU) onboard the PSV.

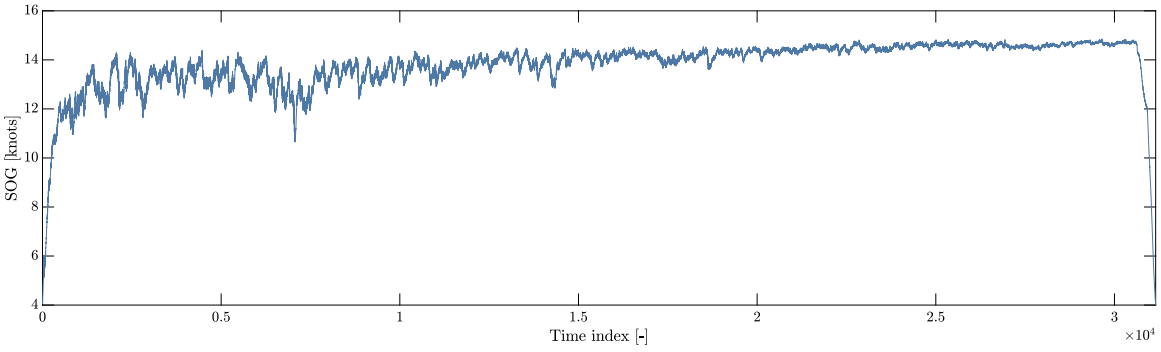


Fig. 11. Trip identified for the PSV using port to port extraction. Low speed maneuvering set to be initiated at a forward speed below 4 knots.

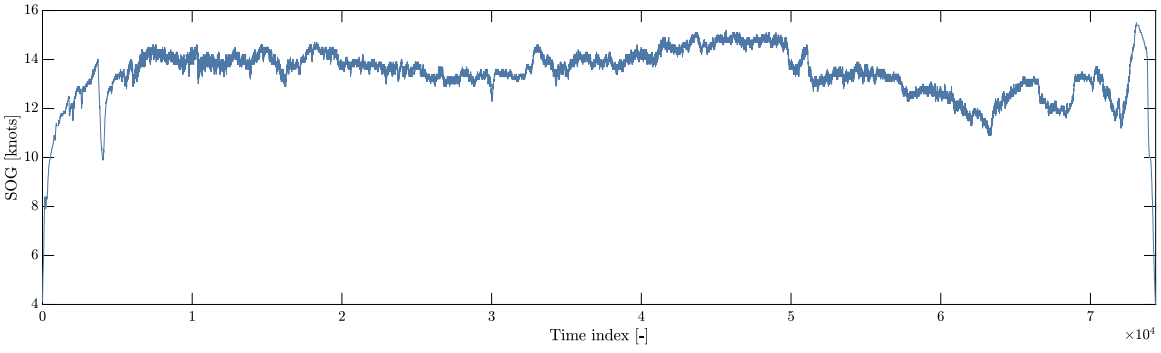
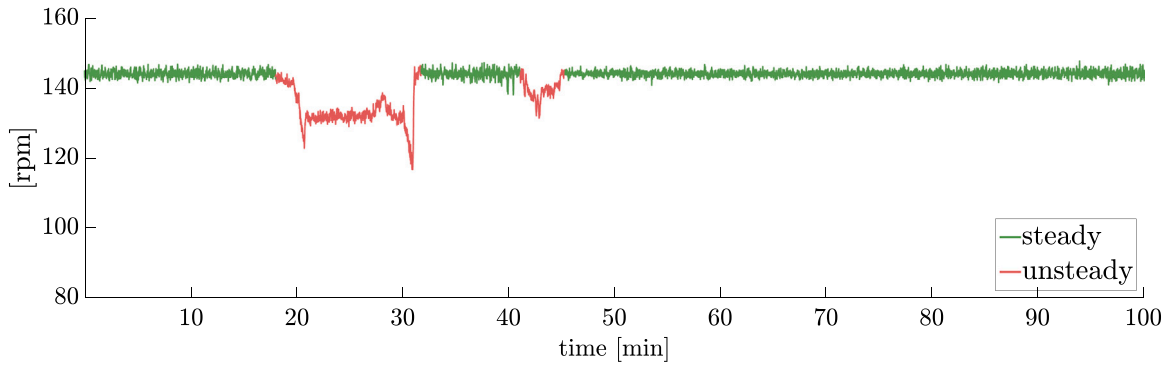


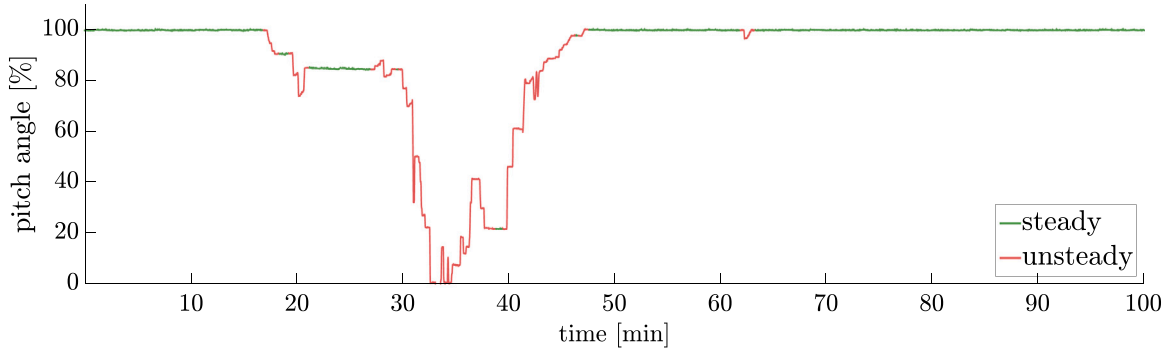
Fig. 12. Trip identified for the MPV using port to port extraction. Low speed maneuvering set to be initiated at a forward speed below 4 knots.

3.6. Data preparation for ship performance analysis

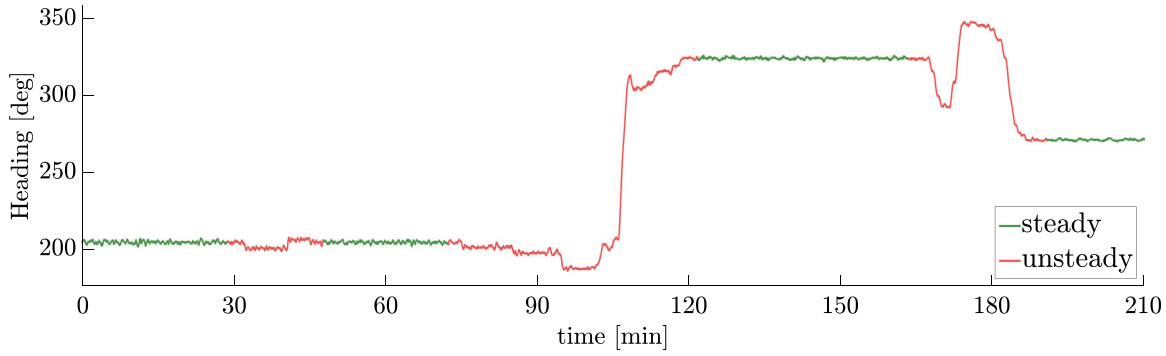
The presented data preparation tools are evaluated with respect to a ship performance analysis of a platform supply vessel (PSV) and a general cargo/multipurpose vessel (MPV). The result of a thorough data preparation is high quality data fully customized for the particular type of analysis. High quality means that errors, distortions and unphysical relationships are omitted as input to the data analysis. Fully customized means that only data relevant for the particular analysis is extracted from the data set.



(a) Propeller rpm data. $\alpha = 1\%$, $n = 300$ (5 minutes).



(b) Propeller pitch data. $\alpha = 1\%$, $n = 30$ (30 seconds).



(c) Ship heading data. $\alpha = 1\%$, $n = 1200$ (20 minutes).

Fig. 13. Time series of propeller rpm data, propeller pitch data and ship heading data showing local steady state (green) based on a t-test of the estimated slope inside a moving window. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To evaluate the final result of the data preparation tools applied on ship monitoring data, a ship performance evaluation in terms of a speed–power analysis is given for each of the two case vessels. Except from the data preparation, no additional corrections are made to the speed and propulsion power data. A plot of the propulsion power vs the forward speed is descriptive in that it presents the variation in consumed power relative to the variation in speed. The plot is particularly descriptive for the ship performance if shown together with the calm water speed–power relation, as it provides a more precise impression of both the amount of operational data not matching the expected calm water relation as well as how far the operational data deviates from the expected calm water relation.

Fig. 14 shows the speed–power data from the one year data set of the PSV, along with curves for the calm water speed–power relation at three different vessel draughts. The blue colored markers represent data prior to application of the data preparation tools, corresponding to 30 537 481 samples. Note however that data exceeding the physical constraints of either the propulsion power or the forward speed are left out, as the most extreme values explode the dimensions of the axes, making the majority of data unreadable. For the PSV the amount of data exceeding the physical constraints, and therefore not shown in the figure, amounts to approximately 150 000 samples, corresponding to 32 h of operation.

The speed–power plot illustrates that the data preparation tools removes unphysical data as well as data that is likely to originate from unsteady behavior. It shows that the prepared data becomes more fitted to the calm water curves as a lower baseline, and that the data for which the forward speed of the vessel is low while the propulsion power is high is removed from the prepared data set. Out of the one year data set from the PSV, 59.7% of the data was removed during data preparation in combination with port to port trips data extraction, while 73.3% of the data was removed during data preparation in combination with steady state identification of propulsion power, propeller pitch angle and ship heading.

Similarly as for the PSV, the speed–power data from the one year data set of the MPV is shown in Fig. 15. The blue colored markers represent data prior to application of the data preparation tools, corresponding to 31 234 048 samples. For this vessel the amount of data

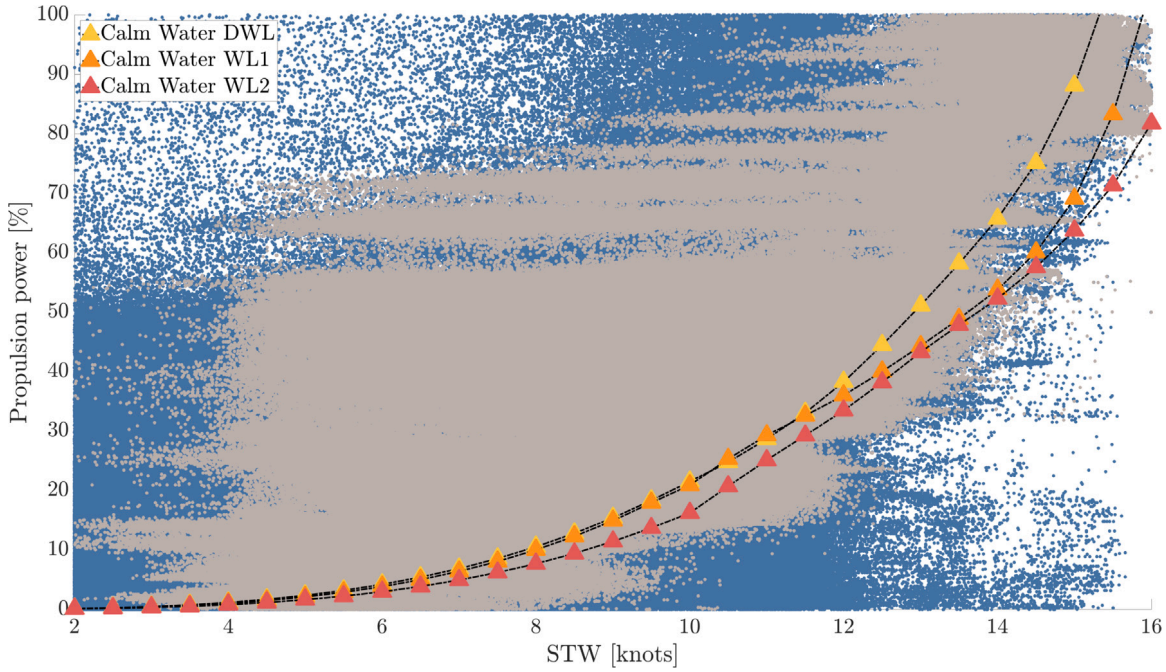


Fig. 14. Speed–power relationship for the platform supply case vessel, showing data prior to (blue colored markers) and post data preparation (gray colored markers). The triangular markers represent the scaled model test data of the calm water performance at three vessel draughts. Except from the current data preparation, no additional corrections are made to the speed and shaft power data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

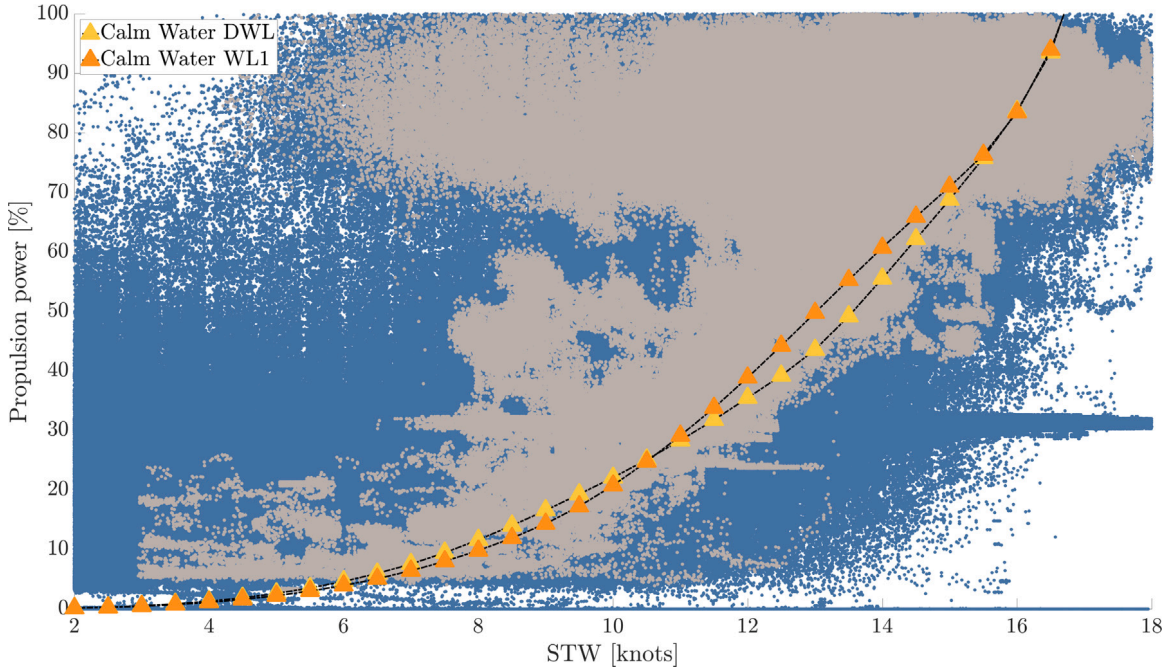


Fig. 15. Speed–power relationship of the general cargo/multipurpose case vessel, showing data prior to (blue colored markers) and post data preparation (gray colored markers). The triangular markers represent the scaled model test data of the calm water performance at two vessel draughts. Except from the current data preparation, no additional corrections are made to the speed and shaft power data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

exceeding the physical constraints of either the propulsion power or the ship speed corresponds to approximately 1 h of operation. The speed–power plot from the MPV data illustrates the performance of the data preparation tools very well. The prepared data fits nicely to the calm water curves as a lower baseline, and the majority of the data showing a combination of low forward speed and high propulsion power is removed from the prepared data set. Out of the one year data set from the MPV, 70.0% of the data was removed during data preparation in combination with port to port trips data extraction, while 80.8% of the data was removed during data preparation in combination with steady state identification of propeller rpm, propeller pitch angle and ship heading.

The most prominent difference between the two speed–power plots based on the PSV and the MPV ship monitoring data sets is the lower

left region of the scattered data, representing low forward speed and medium propulsion power. For the PSV, a major part of this data persists in the prepared data set. For the MPV, most of this data is removed in the prepared data set. This difference can be explained by domain knowledge and logical reasoning. The PSV operates in a more exposed wave and wind environment compared to the MPV, which means that a larger speed loss is expected to take place. The PSV also has a more unfavorable ship length to wave length ratio, which also affects the speed loss. The MPV is therefore expected to have less amount of data in the low forward speed medium propulsion power region, corresponding to large speed loss, compared to the PSV.

4. Conclusion

A stepwise recommended practice for preparation of in-service measurement data for ship operation and performance analysis has been presented and evaluated. The presented methods for preparation of in-service measurement data have been demonstrated and shown to be efficient tools for obtaining high quality in-service data. It is shown how the data preparation improves a ship performance analysis, by presenting a speed–propulsion power relation for two case vessels having different vessel designs.

It is generally recommended to follow this procedure for data preparation concerning most kinds of time series data from continuous monitoring of physical processes. However, it is still encouraged to use specific domain knowledge during the implementation. This is particularly relevant during data extraction, both in terms of selecting variables and in setting the necessary parameters for steady state identification. By following this practice, more focus can be given to the actual data analysis compared to the data preparation, still preserving that only high quality data is used as input to the data analysis.

CRediT authorship contribution statement

Øyvind Øksnes Dalheim: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Sverre Steen:** Conceptualization, Supervision, Writing - review & editing, Funding acquisition, Project administration.

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.

Acknowledgments

The research presented in this paper was carried out in the Kongsberg Maritime University Technology Centre at the Norwegian University of Science and Technology in Trondheim. It was partly financed directly by Kongsberg Maritime through the grant to the University Technology Centre and partly by the project InnoCurrent, which is financed by the the Research Council of Norway grant number 282385 and Kongsberg Maritime. The authors gratefully acknowledge Kongsberg Maritime for providing the full-scale data used in the research.

References

Bendat, J.S., Piersol, A.G., 2010. Random Data: Analysis and Measurement Procedures, fourth ed. Wiley, New Jersey, p. 604.

Blanke, M., Kinnaert, M., Lunze, J., Staroswiecki, M., 2015. Diagnosis and Fault-Tolerant Control, third ed. Springer, http://dx.doi.org/10.1007/978-3-662-47943-8.

Box, G.E.P., Jenkins, G.M., 1976. Time Series Analysis: Forecasting and Control, second ed. Holden-Day, San Franscisco.

Dalheim, Ø.Ø., Steen, S., 2020. A computationally efficient method for identification of steady state in time series data from ship monitoring. J. Ocean Eng. Sci. http://dx.doi.org/10.1016/j.joes.2020.0.

Gervini, D., 2012. Outlier detection and trimmed estimation for general functional data. Statist. Sinica 22 (4), 1639–1660. http://dx.doi.org/10.2307/24310190.

Hansen, S.V., 2011. Performance Monitoring of Ships. (Ph.D. thesis). Technical University of Denmark, Lyngby, p. 201.

Hodge, V.J., Austin, J., 2004. A survey of outlier detection methodologies. Artif. Intell. Rev. 22 (2), 85–126. http://dx.doi.org/10.1023/B:AIRE.0000045502.10941.a9.

ISO, 2011. ISO 15926 Industrial Automation Systems and Integration. Technical Report, International Organization for Standardization.

ISO, 2018. ISO 19848:2018 ships and marine technology - standard data for shipboard machinery and equipment.

Lajic, Z., Nielsen, U.D., 2009. Fault detection for shipboard monitoring and decision support systems. In: Proceedings of the International Conference on Offshore Mechanics and Arctic Engineering, Vol. 6. OMAE, pp. 679–686. http://dx.doi.org/10.1115/OMAE2009-79367.

Lhermitte, S., Verbesselt, J., Verstraeten, W.W., Coppin, P., 2011. A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. Remote Sens. Environ. 115 (12), 3129–3152. http://dx.doi.org/10.1016/j.rse.2011.06.020.

Meyer-Bäse, A., Schmid, V., 2014. Pattern Recognition and Signal Analysis in Medical Imaging, second ed. Academic Press, Kidlington, Oxford.

Nielsen, U.D., 2017. A concise account of techniques available for shipboard sea state estimation. Ocean Eng. 129 (October 2016), 352–362. http://dx.doi.org/10.1016/j.oceaneng.2016.11.035.

Nielsen, U.D., Lajic, Z., Jensen, J.J., 2012. Towards fault-tolerant decision support systems for ship operator guidance. Reliab. Eng. Syst. Saf. 104, 1–14. http://dx.doi.org/10.1016/j.ress.2012.04.009.

Perera, L.P., 2016. Statistical filter based sensor and DAQ fault detection for on-board ship performance and navigation monitoring systems. IFAC-PapersOnLine 49 (23), 323–328. http://dx.doi.org/10.1016/j.ifacol.2016.10.362, URL https://www.sciencedirect.com/science/article/pii/S2405896316319498.

Petersen, J.P., 2011. Mining of ship operation data for energy conservation.

Pyle, D., 1999. In: Cerra, D.D. (Ed.), Data Preparation for Data Mining. Academic Press, San Francisco.

Qin, S., 1997. Neural networks for intelligent sensors and control — Practical issues and some solutions. In: Omidvar, O., Elliot, D.L. (Eds.), Neural Systems for Control, first ed. Academic Press, pp. 215–236. http://dx.doi.org/10.1016/B978-0-12-526430-3.X5000-4.

Rong, H., Teixeira, A.P., Guedes Soares, C., 2020. Data mining approach to shipping route characterization and anomaly detection based on AIS data. Ocean Eng. 198, 106936. http://dx.doi.org/10.1016/j.oceaneng.2020.106936.

Sivathanu, G., Wright, C.P., Zadok, E., 2005. Ensuring data integrity in storage: Techniques and applications. In: Proceedings of the 2005 ACM Workshop on Storage Security and Survivability. pp. 26–36. http://dx.doi.org/10.1145/1103780.1103784.

Swider, A., 2018. Data Mining Methods for the Analysis of Power Systems of Vessels. (Ph.D. thesis). Norwegian University of Science and Technology, Trondheim, p. 412.

Swider, A., Pedersen, E., 2017. Influence of loss of synchronisation between signals from various marine systems on robustness of predictive models algorithms. In: Smart Ship Technology International Conference. London, pp. 37–45.

Vaseghi, S.V., 2008. Advanced Digital Signal Processing and Noise Reduction, fourth ed. Wiley, Chichester, West Sussex.

Vincent, C.L., Pinson, P., Giebel, G., 2011. Wind fluctuations over the North Sea. Int. J. Climatol. 31 (11), 1584–1595. http://dx.doi.org/10.1002/joc.2175, URL http://doi.wiley.com/10.1002/joc.2175.

Zhang, S., Zhang, C., Yang, Q., 2010. Data preparation for data mining. Appl. Artif. Intell. 17 (5–6), 375–381. http://dx.doi.org/10.1080/713827180.