A novel data condition and performance hybrid imputation method for energy efficient operations of marine systems

Michail Cheliotis[a,*], Christos Gkerekos[a], Iraklis Lazakis[a], Gerasimos Theotokatos[b]

[a]*Dept. of Naval Architecture, Ocean and Marine Engineering, University of Strathclyde, Glasgow/UK*
[b]*Maritime Safety Research Centre, Dept. of Naval Architecture, Ocean and Marine Engineering, University of Strathclyde, Glasgow/UK*

Abstract

Datasets with missing values can adversely affect the accuracy of any subsequent decision making, for instance in condition- and performance-monitoring for energy efficient operations of ship systems. Missing data imputation is therefore, a necessary step as it ensures that the data can reach their full knowledge-extracting potential. This paper aims at developing a novel hybrid imputation method, which can be employed to condition data acquired from marine machinery systems, thus increasing the quality of the original dataset and improving the decision making for ship efficient operations. The paper includes of all necessary imputation preparatory steps and further post-imputation processes. The developed method employs a hybrid k-NN and MICE imputation algorithm which combines data mining with first-principle knowledge. The proposed hybrid approach is compared with the individual performance of k-NN and MICE algorithms and is implemented in a dataset acquired from the main engine system of an oceangoing vessel. It is shown that the hybrid approach performs best, exhibiting an average error of 2.2% compared to the k-NN and MICE algorithms with errors 5.6% and 3.3%, respectively, highlighting that the small error of the proposed novel method improves the quality of data used in

*Corresponding author
Email address: michail.cheliotis@strath.ac.uk (Michail Cheliotis)

condition- and performance-monitoring.

Keywords: marine machinery, energy efficiency, data mining, data analytics, missing data, imputation

1. Introduction

Shipping is undoubtedly one of the major driving forces of the global economy, as demonstrated by the yearly volume and value of the seaborne trade (UNCTAD, 2017). Well maintained ships exhibit higher reliability, energy efficiency, profitability and safety. Modern condition- and performance monitoring methods, based on both probabilistic and machine learning models (Mobley et al., 2008; Kobbacy, 2008; Dikis et al., 2014; Beşikçi et al., 2016; Meng et al., 2016; Lazakis et al., 2016, 2018a; Raptodimos and Lazakis, 2018; Lazakis et al., 2018b), can help minimise ship machinery failures and promote energy efficient operation of marine systems (Lazakis and Ölçer, 2014; Trodden et al., 2015; Lepistö et al., 2016).

The accuracy of any ship condition- and performance monitoring model depends heavily, among other things, on the availability and quality of its input data (Dikis, 2017; Mohanty, 2015). Improving the quality of a dataset, prior to its use in a condition monitoring model, is a very important task, which includes treating the missing values (Kotsiantis et al., 2006; Tan et al., 2006; International Organization for Standardization, 2008; Han et al., 2012; Mohammed and Wagner, 2014; Bokde et al., 2018). Missing values befall in most data-driven research efforts and applications and involve the loss of relevant information. If they are not dealt with in a case-appropriate manner, they can reduce the power of models, skew results, and lead to inefficient machinery operations. As discussed by Banko and Brill (2001); Domingos (2012), reducing the number of missing values has a positive effect on the accuracy of any following models (e.g. decision making for energy efficiency).

In the maritime industry, there is a lack of a formalised approach for handling missing data (imputation), despite the increasing popularity of modern data

analytics (Pampaka et al., 2016). This is a concerning phenomenon, as datasets from marine machinery systems tend to contain from 4.4% to 26% missing values, depending on each case (Tsitsilonis and Theotokatos, 2018; Lazakis et al., 2018a). Also, the need for an imputation approach is becoming more prominent when considering the increased use of Machine Learning (ML) algorithms for maritime condition monitoring and other purposes. The need for the above can be seen from analogous efforts in other industries, including the offshore wind industry as proposed by Martinez-Luengo et al. (2019). In the maritime industry, a formalised and accurate approach for the imputation of missing data that includes all necessary imputation preparatory steps and any further post-imputation processes has not yet been suggested.

To tackle the previously mentioned gaps, the novelty of this study lies within the proposal of a new hybrid imputation method that combines data-driven solutions with valuable First Principles (FP) domain knowledge, applicable to the efficient operations of marine machinery. This approach is shown to yield more accurate results compared to traditional, application-agnostic, imputation methods, as it will be shown in the following sections of the paper. Moreover, alongside the hybrid imputation method, all the needed pre-imputation and post-imputation steps are covered. While this study is demonstrated on a marine main engine, it can be adapted to deal with any other physical system.

In the following paragraphs, section 2 provides the relative background, section 3 presents the proposed methodology, section 4 demonstrates the results of the methodology through a case-study of a marine main engine and finally section 5 shows the conclusions drawn.

2. Background

2.1. Peculiarities of shipping DAQ and data analysis

Advanced data analytics and ML tools have, so far, only limited applicability in the maritime industry due to the lack of reliable and complete datasets. This becomes even more prominent when considering methodologies for data pre-

processing and the imputation of missing data from marine machinery condition- and performance monitoring. Any attempts for formalised data analysis in the maritime industry, and especially for condition- and performance monitoring purposes should take into consideration the plethora of different sensors from different manufacturers, the fragile connection links, different sensor sampling rates, the frequent sensor malfunctions, and the inherent synchronisation issues. Also, the engineering interconnection of the different measured systems may be incorporated. The literature for data imputation in the maritime industry is very limited and covers very few of its facets. For example, Claramunt et al. (2017) proposed a methodology, which includes missing data handling, for the purpose of maritime traffic management under the scope of safety and security. Other maritime oriented studies that include missing data can be found in Dobrkovic et al. (2018); Fruth and Teuteberg (2017); Iphar et al. (2015) studies that deal with maritime barge logistics prediction, maritime logistics in general and the Automated Identification System (AIS) respectively. It is becoming evident that there is very little work that provides the steps required for a complete imputation methodology. Regarding the pre-processing of data from marine machinery systems, the available literature is also limited. Currently, the existing knowledge is limited to Original Equipment Manufacturer (OEM) manuals, academic publications and industrial empirical knowledge.

2.2. Imputation importance

Handling missing data is a very important part of data pre-processing as many ML tools used in condition monitoring are restrictive to missing data. More specifically, when unsupervised ML tools are used (e.g. clustering analysis) in datasets with missing values, the models' performance is adversely affected, as clustering algorithms do not have an internal way to manage missing data. In such cases, most clustering algorithms will produce incomplete, and therefore misguiding, results as instances with missing values will not get assigned into clusters. This will reduce the accuracy of the clustering analysis as the information in instance with missing data will be discarded and lost. Alternatively,

the clustering algorithms will simply fail when presented with missing values. Similarly, supervised learning algorithms (e.g. regression, classification) are adversely affected by missing data. For example, missing values in regression analysis can result in poorly fitted models. Also, in most supervised learning algorithms (e.g. classification) missing data can reduce the size of the training data and consequently hinder the models' predictive power. This becomes even more crucial during the deployment of supervised algorithms, as points with missing features cannot be used as input. The issue of missing data is amplified especially when data are limited.

In the maritime industry, and especially in the field of marine predictive maintenance, missing data can lead to inaccurate maintenance scheduling which can cause machinery failures, accidents and risk harm to human life and the environment. There are many reasons that many lead in missing values, especially in shipboard systems. Such factors can include the loss of calibration due to external influences, the loss of sensor connectivity and the inherent difficulty to replace failed sensors. In the context of ship operations, planning, and maintenance missing data can occur in datasets representing equipment condition, process or performance monitoring. Missing data can also be present in voyage related datasets which can include environmental information and general ship information (e.g. speed, propeller slip, currents speed, etc.).

2.3. Missingness mechanisms

In the maritime industry, there are many factors that can make identifying the missing mechanisms of data challenging. External and environmental disturbances and the inherent sensitivity of marine Data Acquisition (DAQ) systems can result in missing data with hybrid missing mechanisms (e.g both Missing Completely At Random (MCAR) and Missing At Random (MAR)). Nonetheless, understanding the underlying causes of missing data is an important step in data imputation, as it dictates the way the missing data can be handled. Little and Rubin (2002) specifies three mechanisms that affect how data are missing. The reader is also refered to (Rubin, 1976; Taylor and Rubin,

1996; Schafer, 1997).

The MCAR mechanism refers to cases where the missingness is independent of the data. In that case, there is no correlation between the missing data and the variables in the dataset, as the missingness is completely unsystematic. For example, in the maritime domain, a random failure of the fuel flowmeter will lead to data that are MCAR.

MAR is when missing data are related to other observations. In other words, the missingness is conditional on another variable. Even though MCAR and MAR seem similar, they should not be used interchangeably as the key difference lies in the condition of the missingness. For instance, if a Main Engine (M/E) is not operational, data from dependant systems may not be recorded, for example, the turbocharger's speed or the Exhaust Gas (EG) temperature.

Missing Not At Random (MNAR) refers to cases where the missingness of an observation depends only on the variable with the missing data; that is, the missingness is conditional to itself. This is inherently a very difficult mechanism to identify. For example, MNAR could result when missing data originate from a M/E that is known to be malfunctioning at the time of the recording.

2.4. Missing data handling methodologies

Understanding the cause of missing data can dictate the strategy that is used to impute them. Nonetheless, attributing all missing data in a dataset to a single missing data mechanism is a difficult task, as missing data do not only conform to one missing mechanism. Generally, there are two broad missing data management methodologies. The simplest methodology includes techniques that discard the missing data. These techniques are simple to implement and they allow for a hastened initiation of the analysis. Such techniques, reduce the size of the dataset which reduces the computational requirements, but can also affect the models' accuracy.

Data retention methodologies have also been developed as an alternative to avoid deleting data points and reducing the number of observations. There are many different data retention techniques which range from very simple to

rather complex, all of which replace the missing values with an estimate. In data science terminology all these approaches fall under the imputation umbrella.

2.5. Imputation techniques

Selecting the appropriate imputation technique is an important decision that has to be made prior to the initiation of the analysis. In theory, the choice of the technique depends on the missing mechanism. Since most missing data do not conform to only one missing mechanism, the selection of the best fitting technique can be case-specific. Lately, it is becoming more common to use hybrid approaches for the best predicting accuracy (Li et al., 2015), which will be presented, among other things, in the following section.

2.5.1. Complete-case and available-case analysis

Complete-case analysis is one of the simplest and easiest to use techniques, which is applied in cases of limited missing values and it is usually employed in purely MCAR data (Pigott, 2003). In complete-case analysis, instances with missing data are excluded from the analysis; this is also known as listwise deletion. Listwise deletion reduces the statistical power of the model and can skew the results significantly as shown in Marsh (1998); Wothke (2000); Graham (2009).

Available-case analysis, or pairwise deletion, is an imputation technique based on the erasure of missing instances. It is an approach which is easy to use but, has decreasing popularity. In pairwise deletion, only the missing values are removed and not the entire instance. Pairwise deletion changes the sample size of each variable and thus makes comparison difficult; it also skews the results and reduces the efficiency of the model, as presented by Kim and Curry (1997); Myers (2011). Figure 1 summarises the difference between the two methods. This approach is incompatible with most ML algorithms as these require the same number of instances for all the variables.

Listwise and pairwise deletion are both valid techniques for managing missing data in many different applications. However, listwise and pairwise deletion

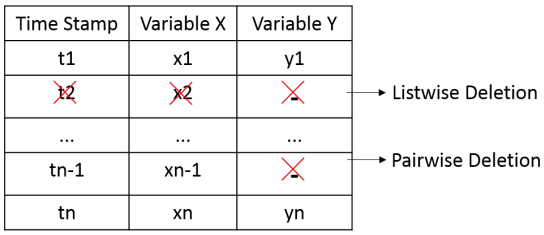

Figure 1: Visual representation of listwise and pairwise deletion. Missing values are depicted using the symbol –, and the × symbol represents observations to be deleted.

should not be used for machinery performance and condition data. In fact, listwise and pairwise deletion should not be used in the following situations. In multivariate datasets with synchronised sampling, pairwise deletion may cause synchronisation issues. For instance, in the case of examining the synchronised measurements from a motor coupled to a pump, applying pairwise deletion will cause the data synchronisation to be lost, as each instance of measurements for the two pieces of equipment should correspond to the same time-stamp. Also, listwise and pairwise deletion should not be used in datasets representing systems with different operational profiles. Both listwise and pairwise deletion will create an artificial operating profile. For example, removing data samples from a vector containing the power of the main engine will create an operating profile which is not representative of the M/E operating profile.

2.5.2. Vertical imputation

Vertical imputation is a group of easy deterministic techniques that use information from the same column as the one of the missing value (Pigott, 2003). One common deterministic strategy is to carry forward the last value. In that way, the last known value before a missing point is carried forward to replace it. Another approach is to impute using the mean of the observed values. This approach can be useful when it is used to impute isolated points in uncorrelated variables (e.g lubricating oil pressure of a turbocharger). Vertical imputation can also use other descriptive metrics such as the median and the mode. In general, vertical imputation is not suitable for successive missing points as it ar-

tificially increases the observations' frequency, underestimates the variability of the results and biases the results (Donders et al., 2006; Pigott, 2003). However, McKnight et al. (2007) states that vertical imputation is an intuitive method that can perform well if combined with other, more sophisticated, approaches.

2.5.3. Horizontal imputation

An alternative to vertical imputation is to impute the missing values with ones from similar parameters, or by using other logical rules. This is commonly known as horizontal imputation as information from the same record is used. Under the scope of the maritime sector, this approach could be used to treat missing values between two identical pieces of machinery (e.g. pumps). The main drawback of this approach is that there may not be two similar parameters in the available dataset, therefore this approach is not always viable (Longford, 2005; Gibert, 2014).

2.5.4. Hot-deck imputation

Hot-deck imputation is an approach that is based on the similarity of a missing instance with a complete one, as initially suggested by Ford (1983); Rizvi (1983); Roth (1994). Hot-deck imputation is more complicated compared with complete-case, available-case and vertical imputation. This approach matches donors (i.e., instances with observed values) with recipients (i.e., instances with missing values). A pool of possible donors is formed based on the similarity between the recipient and the complete instances. The similarity is quantified using a variety of different metrics, including the Euclidean distance, Manhattan distance, Mahalanobis distance and maximum deviation. Figure 2 demonstrates the working principle of hot-deck imputation. Depending on the donor selection, hot-deck can be considered both deterministic and stochastic, (Andridge and Little, 2010; Myers, 2011). Deterministic hot-deck imputation is a non-probabilistic approach. A single donor is selected from the donor pool based on the similarity criteria previously discussed. Alternatively, the imputation can be facilitated by using a summary value (mean, median, and mode) from the

donor pool. Stochastic hot-deck imputation is a probabilistic approach. From the donor pool, a single donor is selected at random and its value is parsed on the missing instance. The benefit of stochastic imputation is that it accounts for uncertainty in a more realistic way. The most common implementation of hot-deck is through the use of k-Nearest Neighbours (k-NN) (Batista and Monard, 2002), further discussed in section 3.

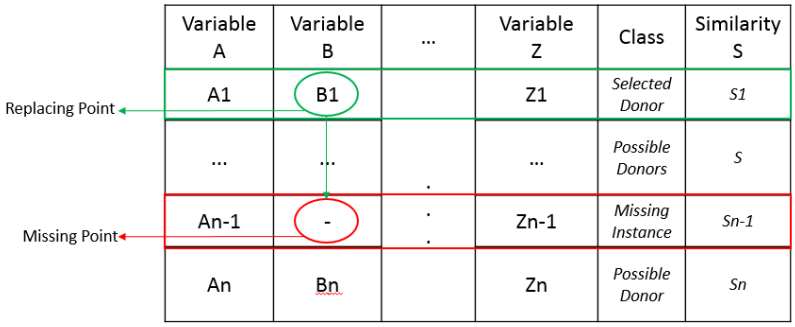

Figure 2: Visual representation of hot deck imputation, applied to a Z-dimensional dataset.

The main benefit of hot-deck imputation is that it does not rely on parameter specific models, hence the imputation is not influenced by any parameter selection. Also, as the imputation is based on actual values the dataset is not completed by artificial ones. However, even though hot-deck is suitable for different missing mechanisms (MCAR, MAR, MNAR) when it is used in single imputation, it fails to account for the uncertainty in an efficient way. More so, there is no explicit mathematical model behind the hot-deck methodology.

2.5.5. Regression-based imputation

Fitting a regression model to appropriate instances with recorded values is another widely used imputation method. Regression-based imputation is more complicated compared with complete-case, available-case and vertical imputation. A regression model is fitted between the target variable (i.e. variable with missing data) and the selected independent variables. The regression model can

be linear, polynomial, or of another type, depending on the dataset. The resulting regression equation is then used to impute instances with missing points in the target variable. In principle, regression-based imputation can be both stochastic and deterministic based on how uncertainty is factored in. Deterministic regression can lead to inaccuracies, as it does not account for uncertainty and it artificially reinforces linear relationships. Instead, exact predictions are used without considering the error term(Enders, 2001). On the other hand, stochastic regression is more common and realistic, as it accounts for the error. Nonetheless, stochastic regression can be more complicated as the distribution of the error term has to be taken into account (Lang and Little, 2016).

As shown by Longford (2005) the general form of stochastic imputation, between the associated variables Y (with missing instances) and W (complete data set), is provided by:

$$Y = f(W) + \epsilon \tag{1}$$

In the equation 1, $f(W)$ is some appropriate function (e.g. linear, polynomial, etc.) and ϵ is the error term which is used to account for the uncertainty.

2.5.6. Multiple imputation

Multiple Imputation (MI) represents a modern and more sophisticated approach to imputation. MI can increase the accuracy of the imputation, while reducing the bias. As discussed by Azur et al. (2011), using a MI approach allows for better accountability of the statistical uncertainty, as opposed to single imputations. The MI approach is based on the improved use of predominant imputation techniques. Assuming an incomplete dataset Y, MI follows the subsequent steps (Figure 3):

1. Impute the missing values of Y m times.
2. Analyse separately the m different datasets.
3. Merge the m different results into one dataset.

MI depends on certain user specified selection steps. Initially, the imputation method has to be specified; selection can include deterministic or stochastic

methods. Next, the number of imputation cycles must be identified. Increasing the number of cycles can increase the model's accuracy, however, this comes at a computational cost. The final selection involves the determination of the concluding missing values from m different datasets. The selection can be facilitated in either a deterministic (mean, median) or a stochastic (random or probabilistic selection) way. All of the above choices are case- and application-dependant. Through the recent literature, it is seen that the Multiple Imputation by Chained Equations (MICE) approach, further discussed in section 3, is one of the most promising and accurate implementation of multiple MI (Royston, 2004; Royston and White, 2015).

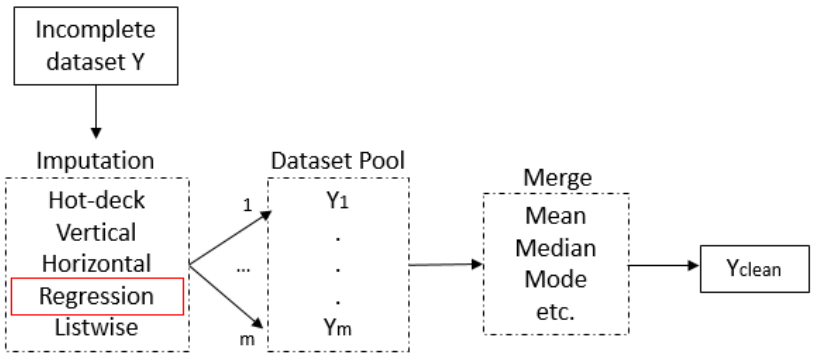

Figure 3: Visual representation of the generalised process of multiple imputation.

2.6. Comparison

Table 1 provides a qualitative summary of the above imputation techniques. From 1 it can be observed that both the hot-deck and regression imputation techniques exhibit the best overall quality (Hron et al., 2010; Templ et al., 2011; Srebotnjak et al., 2012; Sullivan and Andridge, 2015). Both approaches retain the size of the database which is a big advantage, especially when dealing with limited data. Similarly, both approaches can account for uncertainty in their predictions, which is not easily possible with other techniques and especially vertical imputation. Lastly, both hot-deck and regression imputation populate the dataset with plausible values, especially in hot-deck imputation where the

predictions are based on actual observations. When selecting imputation approaches, another consideration should be the ability to combine engineering knowledge with the imputation model. In hot-deck imputation, this is facilitated as the donors for the imputation can be specified by the researcher. This means that when there is a missing value in a specific variable, the donor is selected based on the variables' interconnection within the engineering system. The same philosophy can be applied to regression imputation when specifying the target-independent variables pair. Taking into account the merits of hot-deck imputation, it is worth examining in greater detail its most useful and widespread implementation, k-NN. Not only, k-NN is an effective imputation tool on its own, but it can easily be used in hybrid models (Batista and Monard, 2003; Armina et al., 2017). Similarly, considering the enhanced capabilities of MI (Schafer and Olsen, 1998; Ibrahim et al., 2005) and the merits of regression based approaches (Batista and Monard, 2003) it is worth examining MICE, which represents an ideal combination of the two approaches (Corben, 1954). MICE, like k-NN, is an approach that can be successfully used in hybrid models (Armina et al., 2017). Moreover, a novel hybrid methodology combining the merits of k-NN and MICE will be demonstrated, filling the gap of a formalised approach for handling missing data in the maritime industry.

Table 1: Summary of key findings, comparing different imputation techniques against four criteria.

Imputation method	**Easy to use**	**Retains dataset size**	**Populates with plausible values**	**Accounts for uncertainty**
Listwise/pairwise deletion	✓	✗	✗	✗
Vertical	✓	✓	✗	✗
Horizontal	✗	✓	✗	✗
Hot-deck	✗	✓	✓	✓
Regression	✗	✓	✓	✓

3. Methodology

The focus of this study is the development of a novel hybrid imputation methodology for the maritime industry, through the assessment and performance evaluation of different state-of-the-art imputation approaches. As mentioned above, this is the first step towards an advanced framework which will employ a novel data condition and performance imputation method for energy efficient operations of marine systems. The established approach is holistic and includes all the necessary steps for successful imputation, including pre- and post-imputation steps, which, as shown in the previous sections of this paper, is currently missing from the maritime industry. It has to be kept in mind that the selected imputation approach must cater to the specific needs of the maritime industry (i.e. subsequent use of data in condition or performance monitoring). The necessary steps to achieve this are included in the following distinctive processes:

1. Data collection: including the data gathering effort.
2. Preliminary analysis: including form handling, data synchronisation, data filtering and correlation examination.
3. Imputation process: including the implementation and assessment of the different imputation approaches by examining theAbsolute Percentage Error (APE), Mean Absolute Percentage Error (MAPE) and standard deviation of error (σ).
4. Operational analysis: including the correction of variables based on relevant manufacture's guidelines.

The proposed methodology is shown in Figure 4, in which the assessed imputation approaches including k-NN, MICE and the novel hybrid method are described in more detail together with the preliminary analysis and operational analysis.

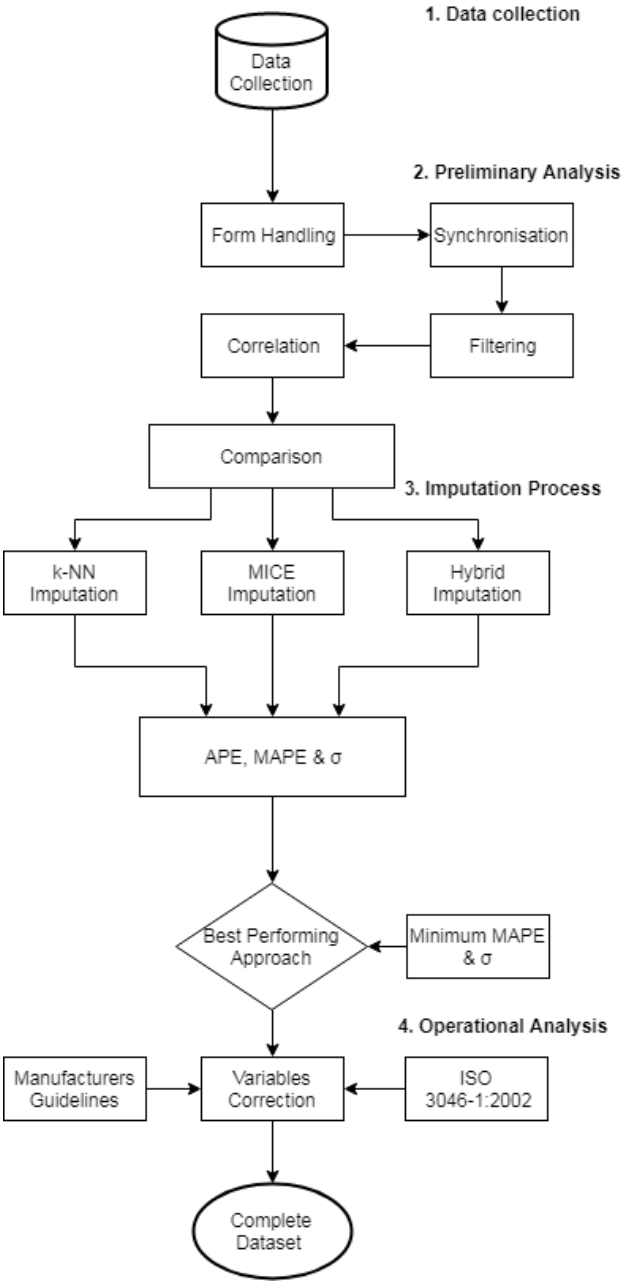

Figure 4: Flowchart of the novel proposed methodology.

3.1. Data collection

The data collection step includes the required activities and efforts to gather the data used in the methodology. The collected data originate from a marine commercial DAQ system installed onboard a merchant navy vessel. In general, data collection is not restricted to only commercial DAQ systems, as it can be facilitated from numerous other sources including log-books and other archives (e.g noon-reports, spread-sheets, etc.). However, commercial DAQ systems offer great advantages with regards to the ease of their use and their increased accuracy.

3.2. Preliminary analysis

Preliminary analysis is an essential preparatory step for the implementation and assessment of the hybrid imputation approach. It includes the form handling, synchronisation, filtering and correlation examination of the data and variables of the given dataset.

The preliminary analysis initiates with the form handling of data. This is a simple yet important step, as data are tabulated in the appropriate form for the next steps of the methodology.

Next, the data are synchronised using linear interpolation (Equation 2); a process that is necessary as the DAQ sensors do not have exactly the same sampling rate.

$$y_2 = \frac{(x_2 - x_1)(y_3 - y_1)}{(x_3 - x_1)} + y_1 \tag{2}$$

In equation 2, x represents the different time stamps and y the variables' values. Subscripts 1 refers to values before the selected timestamp, subscripts 2 refers to values at the synchronisation point and subscripts 3 refer to values after the selected timestamp. The synchronisation of the data ensures that there is consistency among them and they are harmonized over time. It is an especially important process when using similarity-based imputation approaches. This is necessary, as all the measurements from a potential imputation donor must correspond to a single time-stamp.

Data filtering is a very important process that prepares the dataset for imputation. This process determines the points that need to be imputed, by determining if a sensor reading is missing, or it has an illogical value. The assessment of whether a recorded value is logical or not depends on the engineering knowledge of the variable being measured. To determine if an instance has an illogical value (e.g. negative exhaust gas temperatures), sources from the equipment's manufacturer or the results from the commissioning tests (e.g sea trials) are used.

The last step of the preliminary analysis is the correlation examination of the variables. This is a common step in most data driven applications, as it gives a better understanding of the data. Identifying the correlation between the different variables makes imparting FP knowledge to the imputation process easier. The correlation of the available data was examined by using the Pearson correlation coefficient and cross-referencing the results with FP engineering knowledge. The Pearson correlation coefficient ranges between -1 (perfect negative linear correlation) and 1 (perfect positive linear correlation) while 0 denotes no linear correlation.

3.3. Imputation process

Following the preliminary analysis step, the imputation process takes place which includes the implementation of the novel hybrid method and its comparison with the identified state-of-the-art MICE and widely used k-NN algorithms. The assessment includes the application of the different imputation approaches to the required points (as identified during data filtering) in the dataset and the evaluation of the results using residual errors. For the hot-deck imputation approach the k-NN algorithm is used. This is a very popular tool for value prediction and it has widespread applicability in imputation. The k-NN algorithm is intuitively easy to understand and produces accurate results (Zhang, 2012; Huang et al., 2017; Zhang et al., 2018) For the regression based approach the MICE algorithm is selected. The MICE algorithm is a very effective tool in predicting values in multivariate datasets as it incorporates the benefits of

MI with regression. This algorithm is relatively new, compared to the more traditional imputation approaches (e.g. vertical imputation) (van Buuren and Oudshoorn, 1999; White et al., 2011). The hybrid approach that is presented is a novel tool that incorporates the benefits of k-NN and MICE while avoiding their respective shortcomings. In detail the hybrid approach has the following unique benefits:

1. Provides realistic imputations due to the use of the k-NN algorithm.
2. Provides easy incorporation of FP knowledge due to the use of the k-NN algorithm.
3. Is based on the widely recognized k-NN algorithm.
4. Takes advantage of the non-artificial replication of values offered by MICE.
5. Takes advantage of the flexible implementation of MICE.

MICE is a flexible and state-of-the-art, as already discussed, approach that avoids bias in the results by fitting a series of regression models in the data set (Shah et al., 2014). MICE is used to assess the effectiveness of the proposed novel hybrid imputation method. This approach is considered as an implementation of MI which uses linear regressions to help in the estimate of the missing values. For the rest of the section, let Y and K be variables with observed (Y_{obs}, K_{obs}) and missing points (Y_{miss}, K_{miss}) and Z a set of complete variables with Z_{obs} and Z_{miss} corresponding to the observed and missing points of Y and K. MICE calculates the missing points by using a Bayesian approach to update prior distributions of the variables. The Y and K variables are assumed random and prior distributions are assigned (usually an uninformative one). By taking into account the Z variables posterior distributions are obtained. The steps for the application of MICE are the following.

1. Impute all the missing points of Y and K with the means of the Y_{obs} and K_{obs} (Equations 3 and 4); n_{Yobs} and n_{Kobs} represent the total number of

observations for the Y and K variables respectively.

$$\hat{Y}_{miss,i} = \frac{\sum Y_{obs,i}}{n_{Yobs}} \tag{3}$$

$$\hat{K}_{miss,i} = \frac{\sum K_{obs,i}}{n_{Kobs}} \tag{4}$$

These initial estimates ($\hat{Y}_{miss,i}$, $\hat{K}_{miss,i}$) are placeholders and they only facilitate the initiation of the process.

2. Set the placeholders of one of the variables (e.g. $\hat{Y}_{miss,i}$) back to missing

3. Fit a linear regression model (Equation 5) between the observed points of the target variable (Y_{obs}) and the appropriate independent variables (either all, or a subset of Z).

$$\hat{Y}_{miss,i} = \theta^T Z \tag{5}$$

In equation 5, Z is a column vector of the independent variables and θ is row vector of the regression parameters. The $\hat{Y}_{miss,i}$ parameter represents the imputation estimates produced by the regression model.

4. Find the row vector θ by minimising the mean squared error (Equation 6).

$$MSE = \frac{1}{n_{Yobs}} \sum_{i=1}^{n_{Yobs}} (Y_{obs,i} - \hat{Y}_{obs,i})^2 \tag{6}$$

The row vector θ can be calculated based on two different approaches. If the dataset is large, then an optimasation approach can be used (e.g. gradient decent) to fit the regression model and find the row vector θ. However, due to the size of the dataset used, an algebraic method was employed to fit the regression model and find the row vector θ. Generally, if the dataset is relatively small algebraic methods can be used, as they offer greater simplicity. As the size of the dataset increases, optimisation approaches are used as they offer a significant reduction in the required computational time.

5. Use Equation 5 to impute the missing points of the Y variable.

6. Repeat steps 2, 3, 4, and 5 for every variable with missing points in the dataset

7. Reaching step 6 is one cycle. The entire process is repeated for a predetermined number of cycles (usually 10 repetitions is an empirically accepted number).

In summary, MICE uses linear regression in an iterative manner. The process initiates by using mean imputation. Every variable with missing values is used in a regression model to update the initial mean imputation.

The second algorithm that is implemented and assessed isk-NN (k-Nearest Neighbours). k-NN is used to assess the effectiveness of the proposed novel hybrid imputation method. In k-NN, k stands for a specified number of instances (nearest neighbours) that will be considered. This is a non-parametric and lazy algorithm as it does not take into account the distribution of the data in the examined vectors and it has no explicit training phase (Zhang and Zhou, 2007). As with any hot-deck approach, k-NN is based on the similarity between features. There is no standard number for the k hyperparameter, it depends on the field of application and it selection lies with the researcher. In general, a small k will restrict the algorithm to a small region of the data and as a result it will produce results with low bias and high variance. A very small k (e.g. $k = 1$) creates models sensitive to outliers, noise and anomalous data, as the model is overfitted and not generalised enough for use in out-of-sample data. Conversely, a high k (e.g. $k = 20$) will create overgeneralised models, as it averages more possible donors, generating results with low variance and an increased bias. The similarity of the instances is assessed by the Minkowski distance (Equation 7).

$$D = (\sum_{i=1}^{n} \mid x_i - y_i \mid^p)^{1/p} \tag{7}$$

In equation 7 the p hyperparameter is set to 2 which transforms D to the Euclidean distance, the most common distance metric (Groenen and Jajuga,

2001). Similarly, x_i and y_i represent the examined instances. In addition to the distance metric, a weight is also assigned to each possible donor based on its distance. By doing so, closer neighbouring points (i.e. similar and most recent operating conditions) have greater influence over the instance to be predicted. This is a very important feature as it allows taking into account the actual operation of the system under examination.

Lastly, a hybrid approach combining k-NN and MICE is applied and assessed against the state-of-the-art MICE and widely used k-NN algorithms. The k-NN component is based on FP. This approach begins with the correlation analysis where the systemic correlation between the variables is specified. In the hybrid approach, each vector in the dataset is examined in turn. When an instance with a missing value is identified, the k-NN algorithm is deployed. However, the algorithm searches for possible donors only in correlated variables; as determined during the correlation analysis. By doing so, the entire process is executed quicker because only certain vectors of the dataset are examined. Also, the predictive power of the model is enhanced as only correlated variables are used to predict missing points. This process is repeated until no further changes occur to the data set. In many cases, the cessation of the k-NN algorithm signifies its incapacity to impute any more missing points as points from the correlated vectors may be missing simultaneously. In that case, the remaining missing points are predicted using the aforementioned MICE approach. The structure of the proposed novel hybrid methodology is demonstrated in Algorithm 1.

For the assessment of the different imputation tools the APE (Equation 8), MAPE (Equation 9) and the standard deviation of the error (σ) (equation 10) are calculated for each case. The goal is to select the imputation approach with minimum MAPE and σ. The APE and MAPE are selected as they are common and easy to understand metrics for the evaluation of the model's performance (Bryne, 2012). The standard deviation of the error was used to determine the possibility of introducing outliers within the predictions. In the following equations, x_i and $\hat{x_i}$ represent the actual value of the variable and the predicted

Algorithm 1 Hybrib novel imputation method using a combination of k-NN and MICE

Require: filtered dataset **x** of dimension $m \times n$

$modif\_flag \leftarrow 1$
while $modif\_flag == 1$ **do** ▷ Check **x** was updated in prev. loop
    $modif\_flag \leftarrow 0$
    for i = 1, 2, . . . , n **do**
        $temp\_column \leftarrow i$-th column of **x**
        $corr\_columns \leftarrow$ columns correlated with $temp\_column$
        for j = 1,2, . . . , m **do**
            if j-th element of $temp\_column$ does not exist **then**
                if j-th element of $corr\_columns$ exists **then**
                    j-th element of $temp\_column \leftarrow$ k-NN imputation
                    $modif\_flag \leftarrow 1$
                end if
                i-th column of **x** $\leftarrow temp\_column$
            end if
        end for
    end for
end while
x $\leftarrow$ MICE imputation
return x

value respectively.

$$APE = \mid \frac{x_i - \hat{x_i}}{x_i} \mid \tag{8}$$

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} APE \tag{9}$$

$$\sigma = \sqrt{\frac{\sum (APE_i - MAPE)^2}{n}} \tag{10}$$

In summary, a total of three different approaches are implemented and assessed. The first one is MICE, which is applied to the entire dataset until no missing points were left. The MICE approach uses the computational validness of regression with the similarity considerations of the hot-deck approaches and it was implemented in a data driven manner. The second approach that is tried is the k-NN. This approach is applied to the entire dataset, without taking into consideration any systemic correlations, until all the missing points in the dataset are imputed. Lastly, the hybrid approach is tested which combines k-NN with FP analysis and MICE. The k-NN algorithm is deployed by taking into account systemic interdependencies between variables. Then, the MICE algorithm is used to impute any missing points that the FP k-NN cannot predict.

3.4. Operational analysis

The operational analysis is the last step of the developed methodology and includes the correction of the variables to account for ambient conditions. This step ensures that the data are prepared for subsequent analysis, for example, condition monitoring. To account for the ambient conditions, various sources are taken into including international standards and the manufacturers' recommendations. Accounting for ambient conditions is a common step in many applications, as it ensures that the affected variables are adjusted accordingly.

4. Case study

4.1. Preliminary Analysis

The described methodology is applied in the case of a chemical tanker with length L_{OA}=183 m and deadweight 38,000 tonnes, in the Turbocharger (T/C) and M/E system. The selection of the system is based on its criticality and overall importance, as discussed by Harrington (1986); Taylor; Cheliotis and Lazakis (2018). The T/C in question is the TCA66-20032 by MAN B&W and has a maximum speed rating of 16000 rpm; it is connected to the M/E 6S50MC-C which has a Maximum Continuous Rating (MCR) of 9600 kW at 127 rpm.

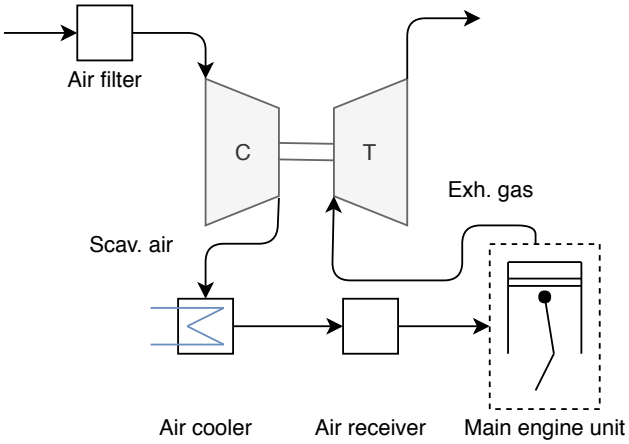

Figure 5: Diagram of a Main Engine system showing the physical interconnections between the measured parameters (the compressor is represented with C and the turbine with T).

The variables available for the analysis, as supplied by a marine DAQ, are: M/E power in kW, M/E speed in rpm, T/C inlet EG temperature in °C, T/C outlet EG temperature in °C, T/C speed in rpm, T/C Lubricating Oil (LO) inlet pressure in bar and T/C LO outlet temperature °C. Also, the engine room air and the air cooler cooling water temperatures are recorded.

As mentioned in the methodology, the first step of the preliminary analysis is the form handling of the dataset. The variables in the dataset are tabulated

and a uniform format is given. Then, the data are synchronised (Equation 2) to harmonize them over time. After the synchronisation, the dataset is filtered to determine the points that need to be imputed. Table 2 shows the limits used for data filtration. The entire dataset is scanned based on limits shown in Table 2; in any instance with values outside of the range specified in Table 2, the values are treated as missing points. The descriptive statistics of the

Table 2: M/E and T/C parameter limits used for the data filtering and identification of points for imputation.

Parameter	Units	Lower Limit	Upper Limit	Source
M/E Power	kW	0	10600	100% load from M/E shop test
M/E Speed	rev/min	0	131	100% load from M/E shop test
M/E Scav. Air Press.	bar	0	3.14	100% load from M/E shop test
M/E T/C EG Inlet Temp.	°C	35	650	From ambient temperature and 130% of the T/C OEM Limit
T/C EG Outlet Temp.	°C	35	650	From ambient temperature and 130% of the T/C OEM Limit
T/C LO Inlet Press.	bar	0	3.6	150% of the M/E shop test
T/C LO Outlet Temp.	°C	35	123	130% of the T/C OEM Limit
T/C Speed	rev/min	0	17600	110% of the T/C OEM Limit
Amb. Air Temp.	°C	-20	50	Ambient conditions range
Amb. Sea Temp.	°C	-5	40	Ambient conditions range

synchronised, uniform and filtered data are shown in Table 3. The final step

Table 3: Descriptive statistics of the recorded dataset.

	Main Engine parameters			T/C parameters				
	Power (kW)	Speed (rpm)	Scav. air press. (bar)	EG inlet temp. (°C)	EG outlet temp. (°C)	LO inlet press. (bar)	LO outlet temp. (°C)	Speed (rpm)
count	1336.0	1336.0	1336.0	1336.0	1336.0	1336.0	1336.00	1336.0
mean	4029.0	98.0	0.9	325.3	289.5	2.0	54.98	9493.6
std	1003.5	13.3	0.3	32.2	25.6	0.1	4.23	1994.7
min	9.0	5.0	0.0	53.6	142.2	1.6	35.40	67.9
25%	3809.5	96.8	0.7	318.9	278.6	1.9	53.00	9087.8
50%	4264.0	101.7	0.9	333.8	292.5	2.1	54.70	9878.8
75%	4680.2	104.5	1.2	339.5	304.5	2.2	58.90	10896.8
max	5876.0	126.1	1.6	364.8	345.5	2.8	63.00	12258.2

of the preliminary analysis is the correlation examination of the variables. For that reason, the Pearson correlation coefficient is initially calculated (Figure 6). The results of the table 6 are cross-referenced with the engineering knowledge of the M/E and T/C systems.

For example, it is known, that the T/C LO inlet pressure and T/C LO outlet

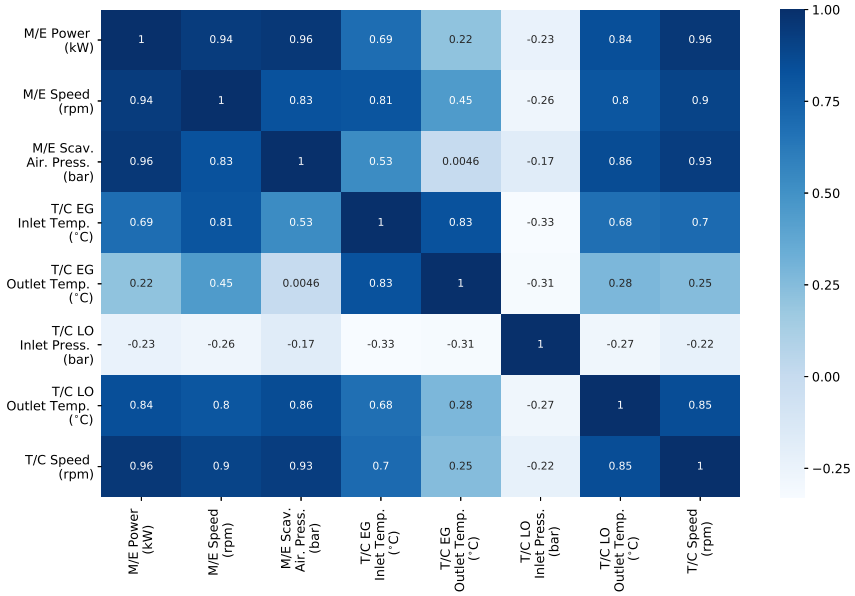

Figure 6: Heat-map showing the Pearson correlation coefficient of the M/E and T/C variables.

temperature are among the uncorrelated variables due to the fact that the T/C LO system is independent and does not come to contact with areas of the M/E or the T/C where the working processes occur. On the other hand, the T/C speed and the M/E power have the biggest correlation. The T/C speed is influenced by many factors. It is correlated with the temperature drop of the exhaust gases in the T/C. The T/C speed is also correlated with the M/E power and M/E speed, which in turn influences the temperature of the exhaust gases. It should be noted that there are other variables that influence the T/C speed, M/E power and the temperature of the exhaust gases. For example, such variables are the combustion pressure and the back-pressure of the T/C. The combustion pressure influences the power output and subsequently the temperature of the exhaust gases. On the other hand, the back-pressure of the T/C can affect (reduce) the T/C speed, as the back-pressure can restrict the flow of the gases (Hountalas et al., 2014; Guan et al., 2015). Even though these parameters are identified for their importance, they are not included in the selected dataset for the analysis.

This is due to the fact that the DAQ used for the measuring of the parameters is not capable of recording them. Comparing the results from Figure 6 and the engineering knowledge of the M/E and T/C systems, the final correlation between variables is obtained (Table 4). As it can be observed, Table 4 shows

Table 4: Resulting correlation based on the integration of the data-driven Pearson coefficient and the first-principle domain knowledge.

	Main Engine parameters			**T/C parameters**				
	Power (kW)	**Speed (rpm)**	**Scav. air press. (bar)**	**EG inlet temp. (°C)**	**EG outlet temp. (°C)**	**LO inlet press. (bar)**	**LO outlet temp. (°C)**	**Speed (rpm)**
M/E Power (kW)		✓	✓	✓	✗	✗	✗	✓
M/E Speed (rpm)	✓		✗	✓	✗	✗	✗	✓
M/E Scav. air press. (bar)	✓	✗		✗	✗	✗	✗	✓
T/C EG inlet temp. (°C)	✓	✓	✗		✓	✗	✗	✗
T/C EG outlet temp. (°C)	✗	✗	✗	✓		✗	✗	✓
T/C LO inlet press. (bar)	✗	✗	✗	✗	✗		✗	✗
T/C LO outlet temp. (°C)	✗	✗	✗	✗	✗	✗		✗
T/C Speed (rpm)	✓	✓	✓	✓	✓	✗	✗	

only the presence of correlation between variables and not the magnitude of the correlation, as required by the hybrid method (it requires only the presence of correlation between variables, and not the magnitude).

4.2. Imputation process

Following the completion of the preliminary analysis step, the implementation and assessment of the different imputation approaches take place. The results from the discussed imputation approaches are presented in Figures 7-14. For each case, a histogram depicting the APE, and a scatter plot, depicting each prediction are shown. In the scatter plot, the line $y = x$, representing perfect accuracy, is plotted to help the determination of the accuracy of each prediction.

Figure 7 shows the imputed values and APE for the T/C LO inlet pressure. Through the histogram, it is noted that all the approaches produce equally good predictions. Also from the histogram, it is observed that the approaches produce predictions with APE ranging from 0-35%. Through the scatter plot, it is observed that the k-NN algorithm produces the results with the biggest standard deviation (sparsity), which can introduce outliers in the dataset. As aforementioned, all three approaches produce results with errors. This behaviour is

attributed to the fact that T/C LO inlet pressure does not have a substantial correlation with the other variables. It has to be kept in mind that all of the imputation approaches are based, to some extent, on the correlation between the variables. he errors produced are justified, due to the T/C LO inlet pressure not following this assumption.

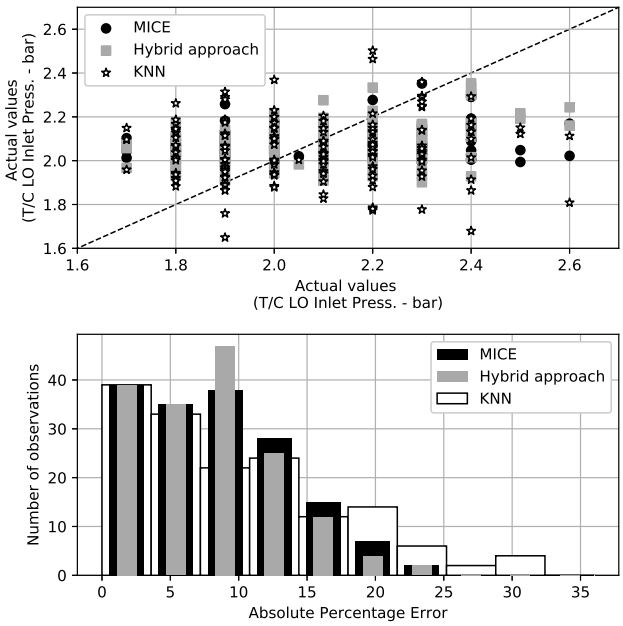

Figure 7: T/C LO inlet pressure imputation performance comparing the MICE, k-NN and hybrid methods.

Figure 8 shows the imputed values and APE for the T/C LO outlet temperature. Through the histogram, it is noted that the hybrid method has the best performance with the majority of the predictions having less than 1% APE. Also from the histogram, it is observed that the approaches produce predictions with APE ranging from 0-8%. Through the scatter plot, it is observed that both k-NN and MICE algorithms produce results with large sparsity, which can introduce outliers in the dataset. The hybrid approach performs the best, as it closely follows the $y = x$ line.

Figure 9 shows the imputed values and APE for the M/E power. Through

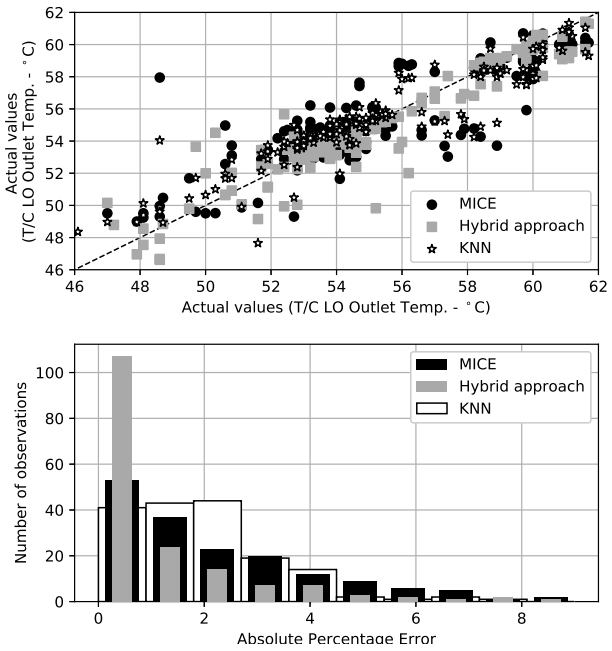

Figure 8: T/C LO outlet temperature imputation performance comparing the MICE, k-NN and hybrid methods.

the histogram, it is noted that the hybrid method has the best performance with the majority of the predictions having less than 1% APE. Also from the histogram, it is observed that the approaches produce predictions with APE ranging from 0-12%. Through the scatter plot, it is observed that the MICE algorithm produces results with large sparsity, which can introduce outliers in the dataset. As the M/E power is a highly correlated variable, the FP component of the hybrid method contributes to making it perform the best.

Figure 10 shows the imputed values and APE for the M/E speed. Through the histogram, it is noted that the hybrid method has the best performance with the majority of the predictions having less than 1% APE. Also from the histogram, it is observed that the approaches produce predictions with APE ranging from 0-8%. Through the scatter plot, we observe that the MICE algorithm produces results with large sparsity, which can introduce outliers in the dataset. It is observed that all the tools produce more accurate predictions from

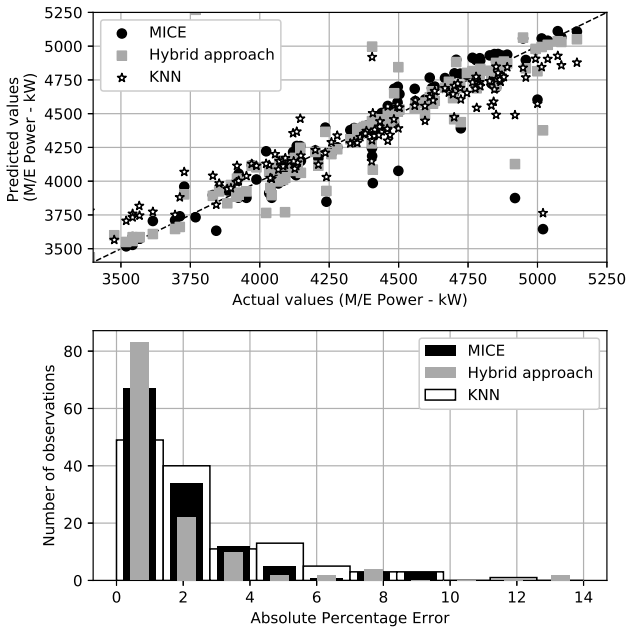

Figure 9: M/E Power imputation performance comparing the MICE, k-NN and hybrid methods.

100 rpm and above. At lower speeds, many of the predictions are relatively inaccurate, with the MICE tool predicting possible outliers. As with the previous cases, the hybrid approach follows the $y = x$ line the closest.

Figure 11 show the imputed values and APE for the M/E scavenging air pressure. Through the histogram, it is noted that the hybrid method has the best performance with the majority of the predictions having less than 1% APE. Also from the histogram, it is observed that the approaches produce predictions with APE ranging from 0-35%. Through the scatter plot, it is observed that the k-NN algorithm produces results with large sparsity, which can introduce outliers in the dataset. In this variable, all of the imputation methods produce results very close to the actual values. The variation of the predictions from MICE and the hybrid tool are quite low and they both follow closely the $y = x$ line.

Figure 12 shows the imputed values and APE for the T/C EG inlet temper-

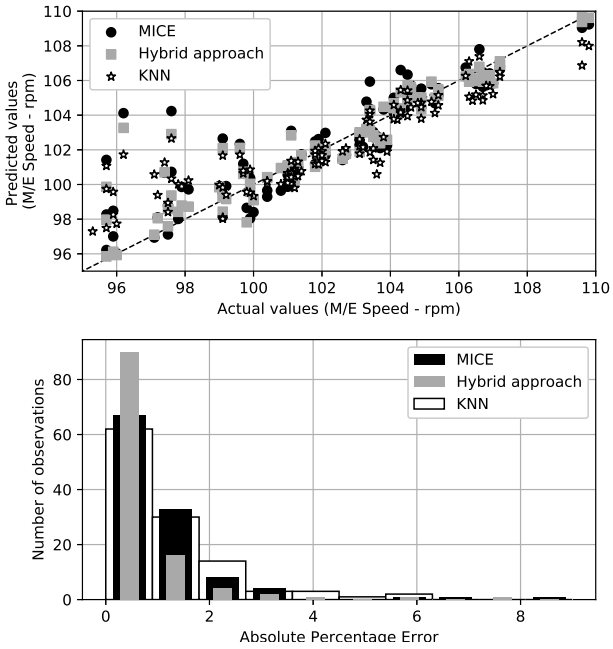

Figure 10: M/E Speed imputation performance comparing the MICE, k-NN and hybrid methods.

ature. Through the histogram, it is noted that the hybrid method has the best performance with the majority of the predictions having less than 1% APE. Also through the histogram, it is noted that the hybrid method has the best performance with the majority of the predictions having less than 1% APE Through the scatter plot, it is observed that the MICE algorithm produces results with large sparsity, which can introduce outliers in the dataset. The T/C EG inlet temperature exhibits a similar behaviour with the M/E Speed. The predictions of all the tools are relatively inaccurate at lower temperatures. However, this behaviour is reverted at higher temperatures, with the hybrid method following the $y = x$ line the closest.

Figure 13 shows the imputed values and APE for the T/C EG outlet temperature. Through the histogram, it is noted that the hybrid method has the best performance with the majority of the predictions having less than 1% APE. Also from the histogram, it is observed that the approaches produce predictions

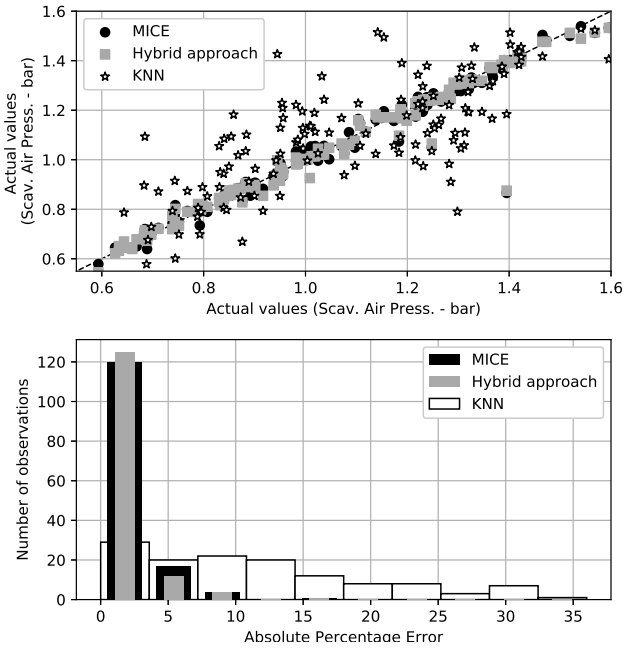

Figure 11: M/E scavenging air pressure imputation performance comparing the MICE, k-NN and hybrid methods.

with APE ranging from 0-10%. Through the scatter plot, it is observed that the MICE algorithm produces results with large sparsity, which can introduce outliers in the dataset. In general, the hybrid method displays consistently good predictions in the temperature range following the $y = x$ line the closest.

Finally, Figure 14 shows the imputed values and APE for the T/C speed. Through the histogram, it is noted that the hybrid method has the best performance with the majority of the predictions having less than 1% APE. Also from the histogram, it is observed that the approaches produce predictions with APE ranging from 0-18%. Through the scatter plot, it is observed that the MICE algorithm produces results with large sparsity, which can introduce outliers in the dataset. Similarly with the previous cases, the hybrid method displays consistently good predictions in the speed range following the $y = x$ line closely.

Summarising the above, Table 5 encapsulates the overall performance of the three tools for the examined variables. Table 5 shows the MAPE and mean

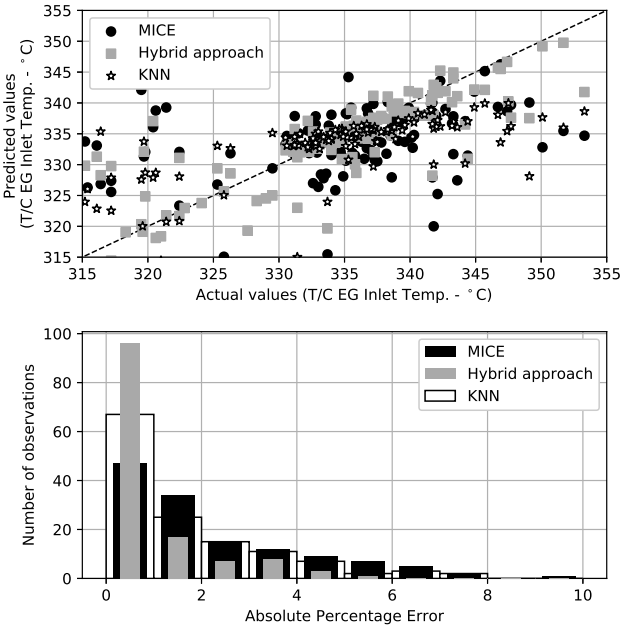

Figure 12: T/C EG inlet temperature imputation performance comparing the MICE, k-NN and hybrid methods.

standard deviation for each approach and for each variable. Also, an overall MAPE and mean standard deviation are shown to summarise the general performance of each approach. As it is observed, even though the two state-of-the-art approaches perform relatively well, the hybrid method outperforms them. It has the lowest overall mean error of 2.21% and the smallest overall standard deviation of 2.64%. The hybrid tool makes accurate predictions without running the risk of generating outliers. The worst performing tool is the k-NN with an overall mean error of 5.55% and an overall standard deviation of 8.9%. Observing the results, it becomes clear that in correlated variables (M/E power, M/E speed, T/C speed, T/C EG inlet temperature, T/C EG outlet temperature, M/E scavenging air pressure, T/C LO outlet temperature) the novel imputation method has superior performance It is observed that the FP component of the hybrid model makes a positive influence on the prediction. By understanding the systemic interdependencies of the system under examination

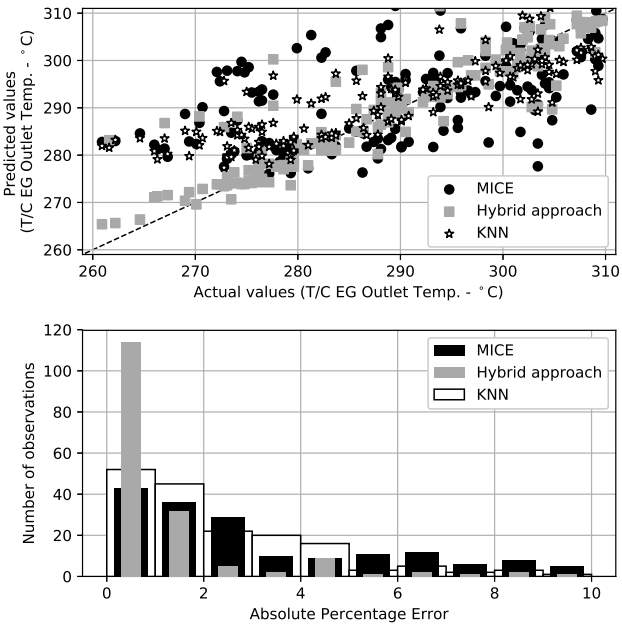

Figure 13: T/C EG outlet temperature imputation performance comparing the MICE, k-NN and hybrid method.

the performance of the predictions is enhanced. Therefore, the integration of the knowledge of the system to any predictive effort is encouraged and should be preferred to purely data driven approaches.

4.3. Operational analysis

Following the imputation process step, the resulting dataset is adjusted to account for the influence of the environmental conditions. For that reason, the T/C speed, M/E scavenging air pressure and the T/C EG inlet temperature were corrected (MAN B&W, 2014; Tsitsilonis and Theotokatos, 2018) according to the manufactures guides and the ISO 3046-1:2002 standards (International Organization for Standardization, 2008). The measured M/E scavenging air pressure, P_{scav}, was adjusted to its corrected figure, $P_{scav,corr}$ according to Equation 11. In Equations 11, 12 and 13 K, F_1 and F_2 are correction constants, while T_{air} and T_{sea} are the ambient temperatures of the air and the sea

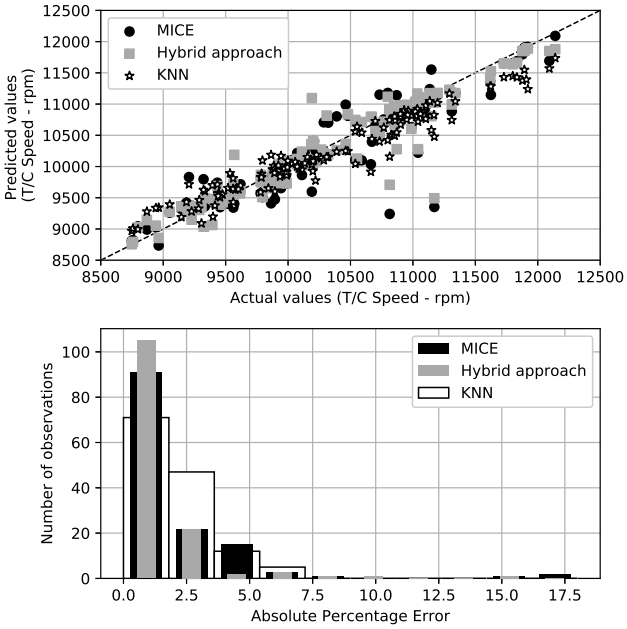

Figure 14: T/C Speed imputation performance comparing the MICE, k-NN and hybrid methods.

respectively.

$$P_{scav,corr} = P_{scva} + (T_{air} - 25)F_1(K + P_{scav}) + (T_{sea} - 25)F_2(K + P_{scav}) \quad (11)$$

The measured T/C speed, N, was adjusted to its corrected value, N_{corr} according to equation 12.

$$N_{corr} = \frac{N}{\sqrt{\frac{(K+T_{air})}{(K+25)}}} \quad (12)$$

The measured T/C EG inlet temperature, T_{egin}, was adjusted to its corrected value, $T_{egin,corr}$ according to equation 13.

$$T_{egin,corr} = T_{egin} + (T_{air} - 25)F_1(K + T_{egin}) + (T_{sea} - 25)F_2(K + T_{egin}) \quad (13)$$

5. Conclusion

This paper develops a novel data condition and performance imputation method for energy efficient operations of marine systems. This study examined

Table 5: Summary of imputation approaches performance

Case	**MAPE**			**Mean σ**		
	k-NN	MICE	Hybrid	k-NN	MICE	Hybrid
M/E Power	3.16%	2.44%	2.29%	5.48%	4.06%	4.98%
M/E Speed	1.15%	1.02%	0.65%	1.12%	1.23%	1.05%
M/E Scav. Air Press.	17.15%	2.34%	1.92%	23.72%	3.63%	3.60%
M/E T/C EG Inlet Temp.	1.63%	2.15%	0.92%	1.69%	1.96%	1.16%
T/C EG Outlet Temp.	2.25%	3.08%	1.19%	1.96%	2.60%	1.64%
T/C LO Inlet Press.	14.92%	8.46%	7.97%	32.11%	5.24%	4.96%
T/C LO Outlet Temp.	2.25%	2.33%	1.29%	3.65%	2.31%	1.73%
T/C Speed	1.96%	1.92%	1.42%	1.49%	2.61%	1.97%
Average	**5.55%**	**2.97%**	**2.21%**	**8.9%**	**2.96%**	**2.64%**

imputations from a holistic view, including all the necessary pre- and post-imputation steps. At the same time, it is shown how a treated dataset with no missing values can lead to more accurate condition monitoring models and therefore, improve the efficient operation of ship systems. The superior performance of the proposed novel imputation method is compared against the existing k-NN and MICE methods and is demonstrated in the case of a M/E andT/C system of a 38,000 tonnes chemical tanker. In total, eight variables are examined including the M/E power, M/E speed, M/E scavenging air pressure, T/C EG inlet temperature, T/C EG outlet temperature, T/C LO outlet temperature, T/C LO inlet pressure and T/C speed. The examined variables are selected based on the availability of the accessible DAQ system. following discussions with the tanker operator. The key outcomes of this research work are the following:

- The development of a novel, hybrid imputation method for the specific needs of marine machinery condition and performance measurements, which streamlines all the pre- and post-imputation steps.
- The demonstration of data synchronisation, filtering, correlation analysis

and variable correction for imputation.

- The importance of treating missing condition monitoring data values, which leads to accurate models for improving the efficient operation of ship systems.
- The use of OEM thresholds and engineering knowledge for data filtering, prior to the imputation process.
- The investigative comparison between the k-NN, MICE and the proposed hybrid method for imputation purposes.
- The superior performance of the hybrid approach exhibiting a mean error of 2.21% compared to the MICE, k-NN algorithms with errors of 3.3% and 5.6%, respectively, highlighting that the small error of the proposed novel method improves the quality of data used in condition- and performance-monitoring.
- The highlight of using FP analysis in the prediction of measurements from an engineering system, compared to purely data driven approaches.
- In the case of uncorrelated variables (e.g. T/C LO inlet pressure), similarity-based imputation methods did not perform well, yielding errors above 8%. In those cases, time-series analysis should be preferred.

Following the above, future work may include the use of the treated dataset in developing an advanced, fault-detection and diagnostic tool for efficient ship operations. A treated dataset could be used to train advanced data-driven models for the accurate identification of developing faults while suggesting rectifying actions. In addition, future work may include the comparison of the novel hybrid imputation method with other imputation algorithms.

6. Acknowledgements

The work presented in this paper is partially funded by the Integrated Ship Energy and Maintenance Management System (ISEMMS) project. ISEMMS

project has received research funding from the Innovate UK Programme. This publication reflects only the authors views and Innovate UK is not liable for any use that may be made of the information contained within

References

Andridge, R.R., Little, R.J., 2010. A review of hot deck imputation for survey non-response. International Statistical Review 78, 40–64. doi:10.1111/j.1751-5823.2010.00103.x.

Armina, R., Mohd Zain, A., Ali, N.A., Sallehuddin, R., 2017. A Review on Missing Value Estimation Using Imputation Algorithm, in: Journal of Physics: Conference Series, IOP Publishing. p. 012004. URL: http://stacks.iop.org/1742-6596/892/i=1/a=012004?key=crossref.4ae7f5f8c76d6b171256642de95c6fcf, doi:10.1088/1742-6596/892/1/012004.

Azur, M.J., Stuart, E.A., Frangakis, C., Leaf, P.J., 2011. Multiple Imputation by Chained Equations: What is it and how does it work? International Journal of Methods in Psychiatric Research 20, 40–49. doi:10.1002/mpr.329.Multiple.

Banko, M., Brill, E., 2001. Scaling to very very large corpora for natural language disambiguation. Proceedings of the 39th Annual Meeting on Association for Computational Linguistics , 26–33URL: http://portal.acm.org/citation.cfm?doid=1073012.1073017, doi:10.3115/1073012.1073017.

Batista, G., Monard, M.C., 2002. A Study of K-Nearest Neighbour as an Imputation Method. HIS'02: 2nd International Conference on Hybrid Intelligent Systems , 251–260URL: http://conteudo.icmc.usp.br/pessoas/gbatista/files/his2002.pdf.

Batista, G.E.A.P.A., Monard, M.C., 2003. An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence 17, 519–533. URL: http://www.tandfonline.com/doi/abs/10.1080/713827181, doi:10.1080/713827181.

Beşikçi, E.B., Kececi, T., Arslan, O., Turan, O., 2016. An application of fuzzy-AHP to ship operational energy efficiency measures. Ocean Engineering 121, 392–402. URL: https://www.sciencedirect.com/science/article/pii/S0029801816301421, doi:10.1016/J.OCEANENG.2016.05.031.

Bokde, N., Feijóo, A., Kulat, K., 2018. Analysis of differencing and decomposition preprocessing methods for wind speed prediction. Applied Soft Computing Journal 71, 926–938. URL: https://doi.org/10.1016/j.asoc.2018.07.041, doi:10.1016/j.asoc.2018.07.041.

Bryne, R., 2012. Beyond Traditional Time-Series: Using Demand Sensing to Improve Forecasts in Volatile Times. Journal of Business Forecasting 31, 13–19. URL: http://eds.b.ebscohost.com/abstract?site=eds&scope=site&jrnl=1930126X&AN=79683755&h=AkHPu9WcokWKB2Swd1uGT%2FMgU%2FB7c6BpHixlg2Z5YbRf3AsCLaslv2Uty8MniLoKWAypGAVKVacvN3f8C2p8jQ%3D%3D&crl=c&resultLocal=ErrCrlNoResults&resultNs=Ehost&crlhashurl=login.aspx%3Fd.

van Buuren, S., Oudshoorn, K., 1999. Flexible multivariate imputation by MICE. 1st ed., TNO Institute Prevention and Health, Leiden.

Cheliotis, M., Lazakis, I., 2018. Ship Machinery Fuzzy Condition Based Maintenance, in: The Royal Institute of Naval Architects (Ed.), Smart Ships 2018, The Royal Institute of Naval Architects, London.

Claramunt, C., Ray, C., Salmon, L., Camossi, E., Hadzagic, M., Jousselme, A.L., Andrienko, G., Andrienko, N., Theodoridis, Y., Vouros, G., 2017. Maritime data integration and analysis: Recent progress and research challenges. Advances in Database Technology - EDBT 2017-March, 192–197. doi:10.5441/002/edbt.2017.18.

Corben, H.C., 1954. Much ado about nothing. Physics Today 7, 10–13. doi:10.1063/1.3061537.

Dikis, K., 2017. Establishment of a novel predictive reliability assessment strategy for ship machinery. Ph.D. thesis. University of Strathclyde. Glasgow. URL: http://digitool.lib.strath.ac.uk/webclient/StreamGate?folder_id=0&dvs=1542893816921~871&usePid1=true&usePid2=true.

Dikis, K., Lazakis, I., Turan, O., 2014. Probabilistic Risk Assessment of Condition Monitoring of Marine Diesel Engines. International Conference on Maritime Technology 2014, 7-9 July 2014, Glasgow, United Kingdom , 1–9.

Dobrkovic, A., Iacob, M.E., van Hillegersberg, J., 2018. Maritime pattern extraction and route reconstruction from incomplete AIS data. International Journal of Data Science and Analytics 5, 111–136. URL: http://link.springer.com/10.1007/s41060-017-0092-8, doi:10.1007/s41060-017-0092-8.

Domingos, P., 2012. A few useful things to know about machine learning. Communications of the ACM 55, 78. URL: http://dl.acm.org/citation.cfm?doid=2347736.2347755, doi:10.1145/2347736.2347755.

Donders, A.R.T., van der Heijden, G.J., Stijnen, T., Moons, K.G., 2006. Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology 59, 1087–1091. doi:10.1016/j.jclinepi.2006.01.014.

Enders, C.K., 2001. The Performance of the Full Information Maximum Likelihood Estimator in Multiple Regression Models with Missing Data. Educational and Psychological Measurement 61, 713–740. doi:10.1177/0013164401615001.

Ford, B.L., 1983. An overview of hot-deck procedures., in: Madow, W., Olkin, I., Rubin, D. (Eds.), Incomplete Data in Sample Surveys, New York Academic Press. pp. 185–207.

Fruth, M., Teuteberg, F., 2017. Digitization in maritime logisticsWhat is there and what is missing? Cogent Business and Management 4,

1–40. URL: http://doi.org/10.1080/23311975.2017.1411066, doi:10.1080/23311975.2017.1411066.

Gibert, K., 2014. Mixed intelligent-multivariate missing imputation. International Journal of Computer Mathematics 91, 85–96. URL: http://www.tandfonline.com/doi/abs/10.1080/00207160.2013.783209, doi:10.1080/00207160.2013.783209.

Graham, J.W., 2009. Missing Data Analysis: Making It Work in the Real World. Annual Review of Psychology 60, 549–576. URL: http://www.annualreviews.org/doi/10.1146/annurev.psych.58.110405.085530, doi:10.1146/annurev.psych.58.110405.085530.

Groenen, P.J.F., Jajuga, K., 2001. Fuzzy clustering with squared Minkowski distances. Technical Report. URL: www.elsevier.com/locate/fss.

Guan, C., Theotokatos, G., Chen, H., 2015. Analysis of two stroke marine diesel engine operation including turbocharger cut-out by using a zero-dimensional model. Energies 8, 5738–5764. doi:10.3390/en8065738.

Han, J., Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques. 3rd ed., Morgan Kaufmann, Waltham. URL: http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pe-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf%0Ahttp://scholar.google.com/schol, doi:10.1016/B978-0-12-381479-1.00001-0.

Harrington, R., 1986. Marine Engineering. The Society of Naval Architects and Marine Engineers, New York.

Hountalas, D.T., Sakellaridis, N.F., Pariotis, E., Antonopoulos, A.K., Zissimatos, L., Papadakis, N., 2014. Effect of turbocharger cut out on two-stroke marine diesel engine performance and NOx emissions at part load operation.

Hron, K., Templ, M., Filzmoser, P., 2010. Imputation of missing values for compositional data using classical and robust methods. Computational Statistics and Data Analysis 54, 3095–3107. URL: http://dx.doi.org/10.1016/j.csda.2009.11.023, doi:10.1016/j.csda.2009.11.023.

Huang, J., Keung, J.W., Sarro, F., Li, Y.F., Yu, Y.T., Chan, W.K., Sun, H., 2017. Cross-validation based K nearest neighbor imputation for software quality datasets: An empirical study. Journal of Systems and Software 132, 226–252. doi:10.1016/j.jss.2017.07.012.

Ibrahim, J.G., Chen, M.H., Lipsitz, S.R., Herring, A.H., 2005. Missing-Data Methods for Generalized Linear Models. Journal of the American Statistical Association 100, 332–346. doi:10.1198/016214504000001844.

International Organization for Standardization, 2008. ISO 3046-1:2002 Reciprocating internal combustion engines: Performance: Part 1: Declarations of power, fuel and lubricating oil consumptions, and test methods: Additional requirements for engines for general use. URL: https://www.iso.org/standard/28330.html.

Iphar, C., Napoli, A., Ray, C., 2015. Data Quality Assessment for Maritime Situation Awareness. Isprs Geospatial Week 2015 II-3, 291–296. doi:10.5194/isprsannals-II-3-W5-291-2015.

Kim, J.o., Curry, J., 1997. The treatment of missing data in multivariate analysis. Sociological Methods and Research 6, 215–240.

Kobbacy, K.A., 2008. Complex System Maintenance Handbook. 1st ed., Springer, New Jersey.

Kotsiantis, S.B., Kanellopoulos, D., Pintelas, P.E., 2006. Data preprocessing for supervised learning. International Journal of Computer Science 1, 111–117. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.132.5127&rep=rep1&type=pdf, doi:10.1080/02331931003692557.

Lang, K.M., Little, T.D., 2016. Principled Missing Data Treatments. Prevention Science , 1–11doi:10.1007/s11121-016-0644-5.

Lazakis, I., Dikis, K., Michala, A.L., Theotokatos, G., 2016. Advanced Ship Systems Condition Monitoring for Enhanced Inspection, Maintenance and Decision Making in Ship Operations. Transportation Research Procedia 14, 1679–1688. URL: https://www.sciencedirect.com/science/article/pii/S235214651630134X?via%3Dihub, doi:10.1016/J.TRPRO.2016.05.133.

Lazakis, I., Gkerekos, C., Theotokatos, G., 2018a. Investigating an SVM-driven, one-class approach to estimating ship systems condition. Ships and Offshore Structures 0, 1–10. URL: https://doi.org/10.1080/17445302.2018.1500189, doi:10.1080/17445302.2018.1500189.

Lazakis, I., Ölçer, A., 2014. Selection of the best maintenance approach in the maritime industry under fuzzy multiple attributive group decision-making environment. Proceedings of the Institution of Mechanical Engineers Part M: Journal of Engineering for the Maritime Environment 230, 297–309. doi:10.1177/1475090215569819.

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

Lepistö, V., Lappalainen, J., Sillanpää, K., Ahtila, P., 2016. Dynamic process simulation promotes energy efficient ship design. Ocean Engineering 111, 43–55. URL: http://dx.doi.org/10.1016/j.oceaneng.2015.10.043, doi:10.1016/j.oceaneng.2015.10.043.

Li, H., Zhao, C., Shao, F., Li, G.Z., Wang, X., 2015. A hybrid imputation approach for microarray missing value estimation. Technical Report. URL: http://www.biomedcentral.com/1471-2164/16/S9/S1, doi:10.1186/1471-2164-16-S9-S1.

Little, R., Rubin, D., 2002. Statistical Analysis with Missing Data. 2nd ed., John Wiley & Sons, New Jersey.

Longford, N.T., 2005. Missing Data and Small-Area Estimation. 1st ed., Springer, New York.

MAN B&W, 2014. Influence of Ambient Temperature Conditions. Technical Report. MAN Diesel and Turbo. Copenhagen.

Marsh, H.W., 1998. Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. Structural Equation Modeling 5, 22–36. doi:10.1080/10705519809540087.

Martinez-Luengo, M., Shafiee, M., Kolios, A., 2019. Data management for structural integrity assessment of offshore wind turbine support structures: data cleansing and missing data imputation. Ocean Engineering 173, 867–883. URL: https://www.sciencedirect.com/science/article/pii/S0029801818308576#pdfjs.action=download, doi:10.1016/j.oceaneng.2019.01.003.

McKnight, P., McKnight, K., Sidani, S., Figueredo, A., 2007. Missing Data: A Gentle Introduction. Guildford Press, New York.

Meng, Q., Du, Y., Wang, Y., 2016. Shipping log data based container ship fuel efficiency modeling. Transportation Research Part B: Methodological 83, 207–229. URL: https://ac.els-cdn.com/S0191261515002386/1-s2.0-S0191261515002386-main.pdf?_tid=edd5cf81-2c9d-474b-8918-7d92b23c8a55&acdnat=1524580041_130121dc52c2959fe7ff7741b8684c32, doi:10.1016/j.trb.2015.11.007.

Mobley, R.K., Higgins, L.R., Wikoff, D.J., 2008. Maintenance Engineering Handbook. 7th ed., McGraw Hill.

Mohammed, J.Z., Wagner, M.J., 2014. Data Mining and Analysis: Fundamental Concepts and Algorithms. 1st ed., Cambridge

University Press, New York. URL: http://www.worldcat.org/oclc/890721171%5Cnhttps://www.chapters.indigo.ca/en-ca/books/data-classification-algorithms-and-applications/9781466586758-item.html%0Ahttp://dx.doi.org/10.1007/s10115-014-0808-1%0Ahttp://dl.acm.org/citation.cfm?doid=3071073.305492, doi:10.1145/3054925.

Mohanty, A.R., 2015. Machinery Condtion Monitoring. 1st ed., Taylor & Francis.

Myers, T.A., 2011. Goodbye, Listwise Deletion: Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data. Communication Methods and Measures 5, 297–310. doi:10.1080/19312458.2011.624490.

Pampaka, M., Hutcheson, G., Williams, J., 2016. Handling missing data: analysis of a challenging data set using multiple imputation. International Journal of Research and Method in Education 39, 19–37. doi:10.1080/1743727X.2014.979146.

Pigott, T.D., 2003. A Review of Methods for Missing Data. Educational Research and Evaluation 7, 353–383. doi:10.1076/edre.7.4.353.8937.

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

Rizvi, M.H., 1983. Hot-deck procedures: Introduction.

Roth, P.L., 1994. Missing data: A conceptual review for applied psychologists. Personnel Psychology 47, 537–560. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1744-6570.1994.tb01736.x/abstract, doi:10.1111/J.1744-6570.1994.TB01736.X.

Royston, P., 2004. Multiple imputation of missing values. The Stata Journal 4, 227–241.

Royston, P., White, I., 2015. Multiple Imputation by Chained Equations (MICE): Implementation in Stata . Journal of Statistical Software 45. doi:10.18637/jss.v045.i04.

Rubin, D.B., 1976. Inference and missing data (with discussion). Biometrika 63, 581–592.

Schafer, J.L., 1997. Analysis of Incomplete Multivariete Data. 1st ed., Chapmn & Hall/CRC, London.

Schafer, J.L., Olsen, M.K., 1998. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. Multivariate Behavioral Research 33, 545–571. doi:10.1207/s15327906mbr3304.

Shah, A.D., Bartlett, J.W., Carpenter, J., Nicholas, O., Hemingway, H., 2014. Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. American Journal of Epidemiology 179, 764–774. URL: https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwt312, doi:10.1093/aje/kwt312.

Srebotnjak, T., Carr, G., De Sherbinin, A., Rickwood, C., 2012. A global Water Quality Index and hot-deck imputation of missing data. Ecological Indicators 17, 108–119. URL: http://dx.doi.org/10.1016/j.ecolind.2011.04.023, doi:10.1016/j.ecolind.2011.04.023.

Sullivan, D., Andridge, R., 2015. A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck. Computational Statistics and Data Analysis 82, 173–185. URL: http://dx.doi.org/10.1016/j.csda.2014.09.008, doi:10.1016/j.csda.2014.09.008.

Tan, P.N., Steinbach, M., Vipin Kumar, 2006. Introduction to data mining. 1st ed., Pearson Education. doi:10.1016/0022-4405(81)90007-8.

Taylor, D.A., . Introduction to Marine Engineering , 1–4doi:10.15713/ins.mmj.3.

Taylor, P., Rubin, D.B., 1996. Multiple Imputation after 18 + Years. Journal of the American Statistical Association 91, 473–489.

Templ, M., Kowarik, A., Filzmoser, P., 2011. Iterative stepwise regression imputation using standard and robust methods. Computational Statistics and Data Analysis 55, 2793–2806. URL: http://dx.doi.org/10.1016/j.csda.2011.04.012, doi:10.1016/j.csda.2011.04.012.

Trodden, D.G., Murphy, A.J., Pazouki, K., Sargeant, J., 2015. Fuel usage data analysis for efficient shipping operations. Ocean Engineering 110, 75–84. URL: http://dx.doi.org/10.1016/j.oceaneng.2015.09.028, doi:10.1016/j.oceaneng.2015.09.028.

Tsitsilonis, K.M., Theotokatos, G., 2018. A novel systematic methodology for ship propulsion engines energy management. Journal of Cleaner Production 204, 212–236. URL: https://linkinghub.elsevier.com/retrieve/pii/S0959652618324934, doi:10.1016/j.jclepro.2018.08.154.

UNCTAD, 2017. Review of Maritime Transport. United Nations, Geneva.

White, I.R., Royston, P., Wood, A.M., 2011. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine 30, 377–399. doi:10.1002/sim.4067.

Wothke, W., 2000. Longitudinal and multi-group modeling with missing data Reprinted. Technical Report. Small Waters Corp. URL: http://www.mpib-berlin.mpg.de/research_resources/index.htmlandhttp://www.smallwaters.com/books/mpi_modeling_code.html.

Zhang, M.L., Zhou, Z.H., 2007. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition 40, 2038–2048. doi:10.1016/j.patcog.2006.12.019.

Zhang, S., 2012. Nearest neighbor selection for iteratively kNN imputation. Journal of Systems and Software 85, 2541–2552. URL: http://dx.doi.org/10.1016/j.jss.2012.05.073, doi:10.1016/j.jss.2012.05.073.

Zhang, S., Cheng, D., Deng, Z., Zong, M., Deng, X., 2018. A novel kNN algorithm with data-driven k parameter computation. Pattern Recognition Letters 109, 44–54. URL: https://doi.org/10.1016/j.patrec.2017.09.036, doi:10.1016/j.patrec.2017.09.036.