Proceedings of the ASME 2019 38th International Conference on Ocean, Offshore and Arctic Engineering
OMAE2019
June 9-14, 2019, Glasgow, Scotland, UK

OMAE2019-95815

BIG DATA ANALYTICS AS A TOOL TO MONITOR HYDRODYNAMIC PERFORMANCE OF A SHIP

Prateek Gupta*, **Sverre Steen**
Department of Marine Technology
Norwegian University of Science and Technology
7052 Otto Nielsens veg. 10, Trondheim
Norway

Adil Rasheed
Department of Engineering Cybernetics
Norwegian University of Science and Technology
7034 Gløshaugen, Trondheim
Norway

ABSTRACT

A modern ship is fitted with numerous sensors and Data Acquisition Systems (DAQs) each of which can be viewed as a data collection source node. These source nodes transfer data to one another and to one or many centralized systems. The centralized systems or data interpreter nodes can be physically located onboard the vessel or onshore at the shipping data control center. The main purpose of a data interpreter node is to assimilate the collected data and present or relay it in a concise manner. The interpreted data can further be visualized and used as an integral part of a monitoring and decision support system. This paper presents a simple data processing framework based on big data analytics. The framework uses Principal Component Analysis (PCA) as a tool to process data gathered through in-service measurements onboard a ship during various operational conditions. Weather hindcast data is obtained from various sources to account for environmental loads on the ship. The proposed framework reduces the dimensionality of high dimensional data and determines the correlation between data variables. The accuracy of the model is evaluated based on the data recorded during the voyage of a ship.

Keywords: Big Data, Ship Hydrodynamics, Principal Component Analysis.

*Corresponding author: prateek.gupta@ntnu.no

INTRODUCTION

The influence of environment on the hydrodynamic performance of a ship is a long studied subject. Estimation of added wave resistance for a ship has always been a topic of research. Moreover, the introduction of Energy Efficiency Design Index (EEDI) and Energy Efficiency Operation Index (EEOI) proposed during 58th MEPC conference is an additional push towards improving the energy efficiency and reducing emissions from shipping industry. The performance of a ship in the absence of any environmental loads can be simply evaluated based on the calm water speed-power relation for the ship. As proposed by Boom & Hout (2008) [1], the speed-power relation or curve, for near calm water condition, can be established by means of speed trials. Alternatively, it is possible to establish such a curve by analyzing the in-service data collected on-board a newly built ship. Based on this curve a simple one-to-one mathematical relation can be formulated between the speed and power consumption of the ship.

As the environment becomes significant, large deviations are observed from the well-known parabolic calm water speed-power curve. In order to explain or predict these deviations, a lot of research has been done, for example, to create prediction models for speed loss of a ship. Prpić-Oršić & Faltinsen (2012) [2] estimated the speed loss of a ship due to ship motions and propeller ventilation. Feng et al. (2010) [3] presented a procedure to predict the speed reduction of a ship accounting only for added resistance in waves. Lu et al. (2018) [4] computed the speed loss of a ship using simulations based on different numerical and

mechanical approaches. Some researchers, on the other hand, predicted these deviations in terms of increased power or fuel consumption by the ship. Seo et al. (2013) [5] presented three different numerical approaches to predict the added resistance in waves and therefore, increased power consumption. Kim et al. (2017) [6] carried out the assessment of ship operating performance for a LNG ship using a power prediction model based on wave basin model test results, numerical computations and empirical formulations to account for environmental loads.

It can be clearly observed that all the above proposed procedures are either using simplified or approximated models of much more complex environmental loads or uses various different components, each one calculating an approximate correction for an individual environmental factor. Moreover, most of these methods neglect smaller influencing factors like engine-propeller degradation and fouling. The biggest advantage of a data-driven model is that it can be developed as a single component model which would be capable of including the effect of even the smallest influencing factor. The challenge here would be to correctly identify the variables which would appropriately quantify all the influencing factors which are responsible for the deviation from the basic speed-power characteristic of the ship.

With the advancement of technology, data-driven approaches have become nearly ubiquitous. Some researchers used this opportunity to develop ship performance evaluation or speed prediction models based on pure statistical or mathematical approach. Mao et al. (2016) [7] tested three different statistics based regression models, using limited/indirect information, for predicting the speed of a small size containership. Pedersen (2014) [8] illustrated the possibility of using purely data-driven method, based on Artificial Neural Networks (ANN) or Gausian Process Regression (GPR), for predicting power consumption of a ship. Gjølme (2017) [9] developed a data-driven machine learning model to predict the speed loss of a ship due to current, wind and waves. Bal Beşikçi et al. (2016) [10] presented an ANN based model to predict fuel consumption of a ship and based on that, developed a Decision Support System (DSS) for energy efficient ship operations. Perera (2017) [11] presented a study to illustrate the use of big data analytics as a data handling framework to process the large volume of data recorded onboard a ship.

The aim of this publication is to develop a data-driven mathematical model which can be used to monitor and assess the hydrodynamic performance of a ship during a sea voyage. The mathematical model is based on a selected set of variables obtained directly or indirectly via onboard measurements and weather hindcast data. The variable selection process is based on the engineering knowledge available to us as well as the results obtained from a preliminary Principal Component Analysis (PCA) model[1]. The preliminary PCA model is also used to detect and understand potential outlier sample data points. Finally, a better fit PCA model is developed with minimum number of Principal Components to statistically explain the variance in the dataset.

BIG DATA ANALYTICS

Big data analytics is a data handling framework which can be used to extract meaningful information from large datasets, often termed as big data [12]. It can be used to perform tasks like data exploration, feature selection, pattern recognition, etc. Big data analytics can be implemented using various Machine Learning (ML) based methods to perform data analysis. As in the case of ML, big data analysis methodologies can also be classified as: supervised, unsupervised and semi-supervised learning.

With the advent of modern technology and automation, Machine Learning (ML) methods are becoming increasingly popular in the field of data science. ML can be said to be a subfield of Artificial Intelligence (AI), which itself is a subfield of computer science. The primary concern about such methods is that they are becoming increasingly opaque and difficult to explain. Holzinger (2018) [13] presented a discussion about the increasing need for explainable AI (XAI) instead. It is well known that a complex Artificial Neural Network (ANN) can be challenging to comprehend. The methodology used in the current work uses a simple Principal Component Analysis (PCA) model which can be quite explainable as demonstrated by Brinton (2017) [14]. The PCA model presented here is developed using a commercial application, The Unscrambler X[2].

Principal Component Analysis (PCA)

An analysis involving more than one variable, often known as multivariate analysis, is generally characterized by a number of correlated variables. Principal Component Analysis (PCA) [15] is an unsupervised machine learning or big data analysis method based on statistics that transforms the correlated multivariate data into a small number of independent and uncorrelated variables, known as Principal Components (PCs). These PCs accounts for variability in the dataset. In general, the first PC accounts for maximum variability and the succeeding PCs accounts for as much of the remaining variability as possible. PCA is also viewed as a method to reduce the dimensionality of a high dimensional dataset that retains most of the information contained in the large dataset. PCA splits the dataset matrix (X) into a modelled part (X_M) and a residual error part (E), with X_M and E having the same dimensions as X:

$$X^{m\times n} = X_M^{m\times n} + E^{m\times n} \tag{1}$$

[1] Further explained in Results section.

[2] https://en.wikipedia.org/wiki/The_Unscrambler

Here, superscript $m \times n$ is the dimension of the original dataset X, i.e., X has m rows and n columns. Generally, in case of a time series data, m is the number of samples and n is the number of variables in the recorded dataset. The modelled part, X_M, is expressed as a subspace with a certain complexity or dimensionality. The model dimensionality (A) represents the number of PCs used to create the model. It should be noted that the residual error E changes with varying model dimensionality. Thus, the primary aim of a PCA model is to, ideally, retain all the information in $X_{M,\,A}$ and discard the remaining noise in E_A.

Pre-processing. Before establishing the model $X_{M,\,A}$, it is customary to pre-process the original dataset. The pre-processing involves two main steps: scaling and mean-centering the data. If variables are recorded on different scales, for example, they have different units of measurement, it is mathematically advantageous to scale these variables to the same scale. This is usually done by multiplying the recorded data matrix (X_{rec}) with a diagonal matrix (S_0) containing one scaling factor for each variable. This scaling factor is, in general, the inverse of the total standard deviation of the corresponding variable, i.e., $diag(S_0)_k = 1/std(X_{rec,\,k})$ where $std(X_{rec,\,k})$ is the standard deviation of the k^{th} column of X_{rec}. The scaled matrix is, thus, defined as:

$$X_S^{m\times n} = X_{rec}^{m\times n}.\, S_0^{n\times n} \tag{2}$$

Further, it is mathematically convenient to mean-center the recorded data such that the modelled part ($X_{M,\,A}$) may be viewed as a Tylor's series expansion around a working point X_0. This working point (X_0) is, in general, the mean-center of the cloud formed by the data points in the given high dimensional space. If X_0 is a row matrix containing the mean of each column (or variable) in X_S and I_C is a column matrix of ones, then the mean-centered data matrix is calculated as:

$$X^{m\times n} = X_S^{m\times n} - I_C^{m\times 1}.\, X_0^{1\times n} \tag{3}$$

It should be noted that scaling and mean-centering the data would not affect the final outcome of the model but, in some cases, unscaled data might introduce the rounding-off error during matrix operations [16].

Bilinear Modelling. In PCA, the data-driven mathematical model is regarded as a sum of contributions from different functions of rows and columns. Each of these functions is simply approximated as a linear model. Thus, resulting in a bilinear model as follows:

$$X^{m\times n} = X_{M,\,A}^{m\times n} + E_A^{m\times n} = T_A^{m\times A}.\, P_A'^{\,A\times n} + E_A^{m\times n} \tag{4}$$

Here, matrix T_A contains the so-called scores and matrix P_A contains the so-called loadings with each column corresponding to a Principal Component (PC). P_A' represents the transpose of P_A matrix. The above expression can also be written as summation of A PCs:

$$X^{m\times n} = \sum_{i=1}^{A} t_i^{m\times 1} \times p_i'^{\,1\times n} + E_A^{m\times n} \tag{5}$$

Where t_i and p_i are column matrices or vectors containing scores and loadings of i^{th} PC, respectively. Scores shows the patterns of co-variation among m samples whereas loadings shows the corresponding patterns of co-variation among n variables. It should be noted that the model dimensionality (A) is user-specified but is limited by the maximum number of linearly independent rows or columns in X, commonly known as the rank of the matrix, i.e., $A_{max} = rank(X)$.

Conventionally, in PCA, the score vectors (columns in T_A) are orthogonal[3] to each other and the loading vectors (columns in P_A or rows in P_A') are orthonormal[4]. The scores and loading can be estimated in many different ways. Two of the most popular methods are: Singular Value Decomposition (SVD) [17] and Nonlinear Iterative Partial Least Squares (NIPALS) algorithm [18]. SVD is a direct method which calculates the maximum number of PCs (determined by the rank of the data matrix) whereas NIPALS is an iterative method which calculates 1 PC at a time. NIPALS algorithm can be further modified to accommodate missing values in the dataset using a method given by Martens & Martens (2001) [19].

Singular Value Decomposition (SVD). SVD is a generalization of eigen-decomposition[5] for any $m \times n$ matrix. The data matrix (X), having dimensions $m \times n$ with $m \geq n$, can be decomposed using SVD as follows:

$$X^{m\times n} = U^{m\times n}.\, \Sigma^{n\times n}.\, V'^{\,n\times n} \tag{6}$$

Where U consists of n orthonormalized eigenvectors of ($X.\, X'$), V consists of orthonormalized eigenvectors of ($X'.\, X$)

[3] $T_A'.\, T_A = \lambda$, where λ is a diagonal matrix and $diag(\lambda)_i$ is proportional to the eigenvalue associated with the i^{th} PC.

[4] $P_A'.\, P_A = I$ where I is an identity or unit matrix.

[5] Factorizing a diagonalizable square matrix into eigenvalues and eigenvectors.

and Σ is a diagonal matrix containing non-negative square roots of the scaled eigenvalues of $(X'.\ X)$, also known as singular values. The columns of U and V are also known as left-singular eigenvectors and right-singular eigenvectors of X, respectively.

Comparing Equations (5) and (6), the loading vectors (p_i) corresponds to columns in V and score vectors (t_i) corresponds to columns in $(U.\ \Sigma)$. Moreover, for the model with maximum dimensionality, i.e., $A = A_{max}$, $P = V$ and $T = (U.\ \Sigma)$.

Covariance & Correlation

Variance of a data variable is an absolute measure of variability which quantifies the "spread" of the observations from the expected or mean value of the variable. It can also be interpreted as the mean of the squares of the deviations. Covariance quantifies the relation between the variability of two variables, i.e., it measures the deviation from mean for these two variables with respect to each other. The covariance between a variable and itself is the variance. Covariance is mathematically formulated as:

$$Cov(\alpha,\beta) = E\big((\alpha-\overline{\alpha})(\beta-\overline{\beta})\big) \tag{7}$$

Where α and β are data variables. A high positive or negative value of covariance indicate a strong relationship between variables whereas zero covariance indicate that the variables may be independent of each other.

For more than two variables, the statistical relationship between variables can be quantified as the covariance between two variables at a time and can be presented in the form of a covariance matrix as:

$$Cov(X) = \begin{bmatrix} Cov(X_1,X_1)\ Cov(X_1,X_2)\ Cov(X_1,X_3) \\ Cov(X_2,X_1)\ Cov(X_2,X_2)\ Cov(X_2,X_3) \\ Cov(X_3,X_1)\ Cov(X_3,X_2)\ Cov(X_3,X_3) \end{bmatrix} \tag{8}$$

Where X_1, X_2 and X_3 are data variables. It should be noted that the above matrix is symmetric about the diagonal as $Cov(X_1,X_2) = Cov(X_2,X_1)$ and the diagonal of the matrix contains the variance of the corresponding variable. If X is a mean-centered vector containing n elements or variables $(X_1, X_2,..., X_n)$, the above equation can also be written as:

$$Cov(X) = \frac{1}{n-1} X'.\ X \tag{9}$$

For a better interpretation of covariance matrix, it is sensible to scale the covariance matrix. This is, generally, done by dividing each element of matrix by the product of standard deviation of the corresponding variables. The scaled covariance is also known as the correlation between the variables, calculated as:

$$r(X_1,X_2) = \frac{Cov(X_1,X_2)}{std(X_1).\ std(X_2)} \tag{10}$$

In case of already scaled variables, the covariance matrix is the correlation matrix. Also, the diagonal of the correlation matrix will contain 1 indicating 100% correlation between a variable and itself.

Correlation Loadings. Correlation loadings are defined as the correlation between data variables and a Principal Component (PC). The correlation loadings, thus, can be used to interpret the physical meaning of a PC and it is also useful in visualizing the relationship between individual variables. Two variables which are strongly correlated ($\sim \pm 1$ correlation) with a PC will also be strongly correlated with each other. Additionally, the correlation loadings reflects the contribution of individual variable to a PC and quantifies the amount of variability, contained in that variable, which is accounted for or absorbed by the PC.

The correlation loading matrix can be calculated as the cross-correlation between standardized data variables and standardized PCs. Using Equations (6) and (9), the correlation loading matrix can be calculated as:

$$L = \frac{1}{n-1} X'.\ \sqrt{n-1}\ U = \frac{1}{\sqrt{n-1}} V.\ \Sigma \tag{11}$$

Thus, the correlation loading vectors are, simply, loading vectors scaled by the square root of the respective eigenvalues.

Explained Variance

Each Principal Component (PC) is characterized by three main parameters: scores, loadings and explained variance. Explained variance is a measure of the amount of variability or information, contained in all the variables, taken into account (or absorbed) by a PC. It is often quantified as the percentage of total variance in the data which is accounted for by the current PC.

The variance explained by a PC is the variance contained in the corresponding PC score vector. The total explained variance is obtained as the cumulative sum of the variance explained by consecutive PCs accepted in the model. Thus, the total explained variance represents how well the data fits the model, i.e., it measures the accuracy of the model. The explained variance and total

explained variance are, in general, presented as the percentage of the total variance contained in the original dataset.

DATA

Two main datasets are used for the current work: ship in-service measurement data and weather hindcast data. These two datasets were acquired from different sources. The in-service measurement ship data contains the measurements recorded by various sensors onboard a ship whereas the weather hindcast data represents the external environmental loads that the ship was assumed to be experiencing.

Data Description

The ship data used in the analysis was recorded onboard an approximately 200m long general cargo ship. The ship has a installed propulsion capacity of approximately $10,000KW$ in MCR condition. The vessel is equipped with a comprehensive energy management web application, Marorka Online[6]. Marorka Online is a platform for visualizing fleet data, and it facilitates collaboration between the ship and shore, i.e., the data recorded onboard the ship is transmitted to the shore control center in real-time. The system records parameters which are relevant as performance indicators for the vessel.

Ship Data. The input ship data is about a month-long continuously recorded time series, sampled and stored at every 15 minutes. Figure 1 shows the trajectory of the ship during the voyage. The recorded data contains 26 variables. These variables are classified into different categories with each category representing the nature of the information conveyed by the variables. These categories are: ship identity, navigation, auxiliary power system, propulsion system, and environment. In addition to these categories, time is defined as an independent variable. Table 1 presents the list of categorized ship data variables.

Propulsion system variables are primarily related to the hydrodynamic performance of the ship. But some of these variables, like *State*, *Draft Fore*, and *Draft Aft*, do not directly correlate with ship performance. *State* variable indicates the operational state of the vessel. It has one of the following four values for each time step: 'At Berth', 'Manoeuvring', 'Sea Passage', or 'Anchor/Waiting'. This is used to further discretize the data and only 'Sea Passage' data is used for the analysis. *Draft Fore* and *Draft Aft* are used to introduce two additional variables: mean draft and trim-by-aft. They are more relevant from the hydrodynamics point of view. The cargo weight remained constant during the whole duration of the journey, thus, it cannot be included in the analysis.

[6]www.marorka.com

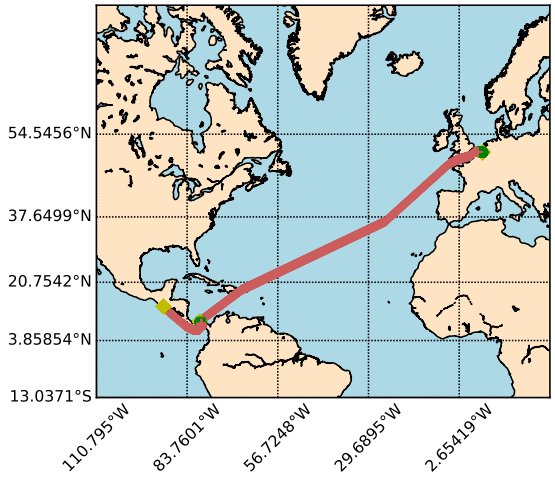

FIGURE 1: SHIP'S TRAJECTORY FOR THE MONTH LONG RECORDED VOYAGE DATA.

Environment variables represent wind loads and sea depth. Incident wind loads strongly influence the hydrodynamic performance of the ship due to air drag. From relative wind speed and direction, longitudinal and transverse incident relative wind speeds are calculated. It is quite obvious that longitudinal and transverse wind speeds would be more correlated to vessel performance. It was observed that the sea depth values were not continuously recorded probably due to the limitation of the depth sensor, so it is not included in the analysis. Navigation variables are used to interpolate hindcast weather data variables, representing environmental loads on the ship.

Ship identity and auxiliary power system variables are not used in the current analysis. As the ship is propelled by a diesel engine, auxiliary power systems hardly influence the hydrodynamic performance of the vessel. In case of an electric propulsion system this might not be the case. Also, in case of a very detailed analysis, say using a numerical model of the ship to determine its hydrodynamic properties, ship identity variables can be used to fetch the building specifications and designs of the ship. This can be very useful to theoretically or empirically predict the calm water hydrodynamic performance of the hull with varying mean draft and trim of the vessel.

Weather Data. The weather data is acquired from two sources: European Centre for Medium-Range Weather Forecast (ECMWF) [20] and Hybrid Coordinate Ocean Model (HYCOM) [21]. The ECMWF data is the ERA-Interim reanalysis data. ERA-Interim is a global atmospheric reanalysis from 1979, which is continuously updated in real time. The spatial resolution

TABLE 1: CATEGORIZED LIST OF VARIABLES RECORDED ONBOARD THE SHIP. 'NAVIGATION', 'PROPULSION SYSTEM' & 'ENVIRONMENT' ARE IMPORTANT CATEGORIES FOR THE CURRENT ANALYSIS.

Ship Identity	Navigation	Auxiliary Power System	Propulsion System	Environment
Ship Name	Latitude	Aux. Consumed	State	Relative Wind Speed
IMO Number	Longitude	Aux. Electrical Power Output	ME Load Measured	Relative Wind Direction
	Gyro Heading	DG1 Power	Shaft Power	Sea Depth
	COG Heading	DG2 Power	Shaft rpm	
		DG3 Power	Shaft Torque	
			ME Consumed	
			Draft Fore	
			Draft Aft	
			GPS Speed	
			Log Speed	
			Cargo Weight	

for ECMWF data, used here, is 0.75°, i.e., approximately $80km$. It provides wave data variables every 6 hours and wind data variables every 3 hours. The data variables obtained from ECMWF includes northward and eastward wind speed $10m$ above the sea surface, significant wave height, mean wave period and mean wave direction. The data obtained from HYCOM has a spatial resolution of $1/12^\circ$ with a sampling frequency of 1 measurement per day. The data variables obtained from HYCOM includes northward and eastward sea water speed.

The weather data variables are linearly interpolated in space and time to ship's location using the ship's navigation data. The weather data variables obtained from ECMWF and HYCOM are, further, transformed to ship's reference frame, i.e., northward and eastward wind and current speeds are transformed to longitudinal and transverse wind and current speeds using ship's gyro heading. Since the wave data variables cannot be directly transformed to ship's reference frame as in the case of wind and current, only a new variable, relative mean wave direction, is introduced using the mean wave direction and ship's gyro heading.

Data Exploration & Validation

The hydrodynamic performance of a ship is, in general, quantified as the maximum speed achievable for a given propulsive power output. The propulsive power output is, here, measured as the percentage of Maximum Continuous Rating (MCR) load and recorded as the variable *ME Load Measured*. The propulsive power output will be correlated with the measured shaft power. The measured shaft power will differ from the total propulsive power output of the main engine due to transmission losses, which may or may not vary with varying engine power output. The shaft power can be calculated from shaft torque (τ) and rpm (n) as:

$$P = \tau.\omega = \frac{2\pi n}{60}.\tau \tag{12}$$

The acquired data includes shaft power, torque and rpm readings, but the shaft power can also be calculated from measured shaft torque and rpm (Equation (12)). Figure 2 presents the measured and calculated shaft power vs main engine load. A minor difference for a few values is noticeable but otherwise the values are in good agreement.

The main engine power output will be correlated with its fuel oil consumption. *ME Consumed* variable contains the value of fuel consumed by the main engine between the two recorded samples. The correlation between ship speed and propulsive power output will be strongly influenced by environmental loads like wind, waves and current. Figure 3 shows the correlation between ship's speed through water, i.e., log speed and main engine power output for 'Sea Passage' state only. A substantial variation of ship's speed through water for a given main engine load is observed which indicates the influence of environmental loads.

Figure 4 shows the fuel oil consumed by main engine for a measured instantaneous load output. It should be noted that the fuel oil consumption values are recorded as the total fuel consumed between two sampling instants, i.e., fuel consumed in past 15 minutes. Thus, it will include the dynamic effects between these sampling instants as well as the variation of engine performance with environmental loads. Minor variations in fuel oil consumption for the same engine load indicate these effects but the large deviations from the mean trend-line observed between 20% and 50% ME load cannot be explained by this. A deeper analysis into the data shows that the ME fuel consumption readings taken during the initial 24 hours shows abnormally

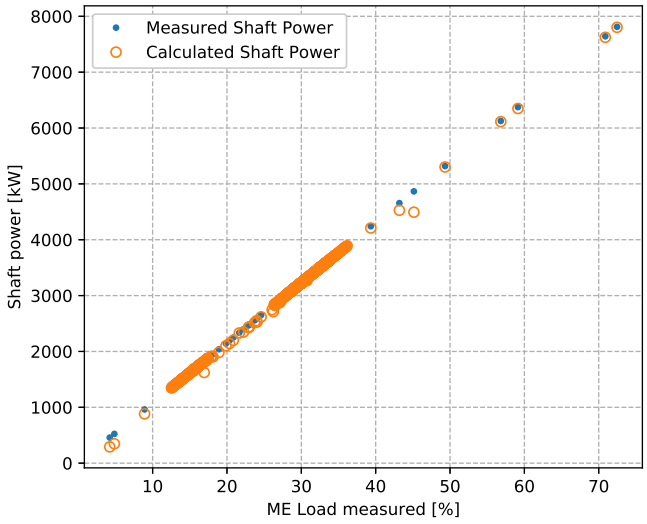

FIGURE 2: COMPARISON OF MEASURED AND CALCULATED SHAFT POWER. CALCULATED SHAFT POWER IS OBTAINED FROM MEASURED SHAFT TORQUE AND RMP.

high values due to unknown reasons. Keeping this in mind, all the data recorded for first two days are removed from the analysis.

Log speed is the measured speed of ship through water whereas GPS speed is the speed of the ship relative to the ground. In the absence of sea current, log speed will coincide with GPS speed. Thus, the difference between log speed and GPS speed is correlated with sea current speed. The difference between these two measured speeds can be considered as an estimate of sea current speed in longitudinal direction of the ship. Thus, it is possible to validate this difference with the sea current speed obtained from HYCOM data. Figure 5 shows the comparison of longitudinal current speed obtained from HYCOM data and the estimated value obtained as the difference between log speed and GPS speed of the ship for 'Sea Passage' state (the gap in the time-series is due to removal of data points when the ship was 'Manoeuvring' across Panama Canal). The two values shows quite good agreement. It should be noted here that high current speeds are well estimated by the difference between log and GPS speeds.

In the current analysis, the wind speed data is obtained from two sources: ship's data and ECMWF. It is, therefore, possible to validate these two sources by comparison. As mentioned above, the relative wind speed and direction obtained from on-board measurements are used to calculate relative longitudinal

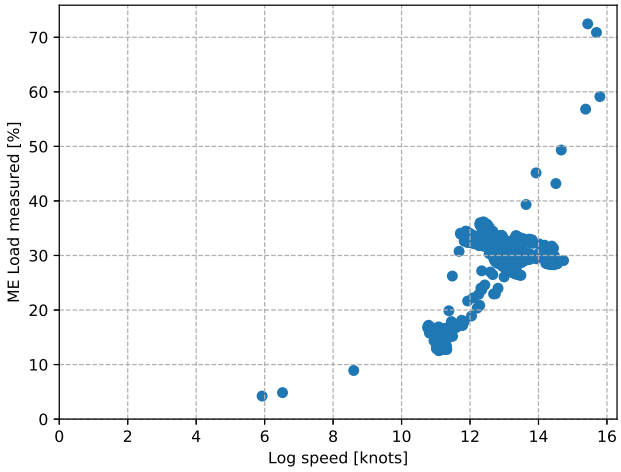

FIGURE 3: MEASURED MAIN ENGINE LOAD (ME LOAD) AS A FUNCTION OF MEASURED LOG SPEED (OR SPEED THROUGH WATER).

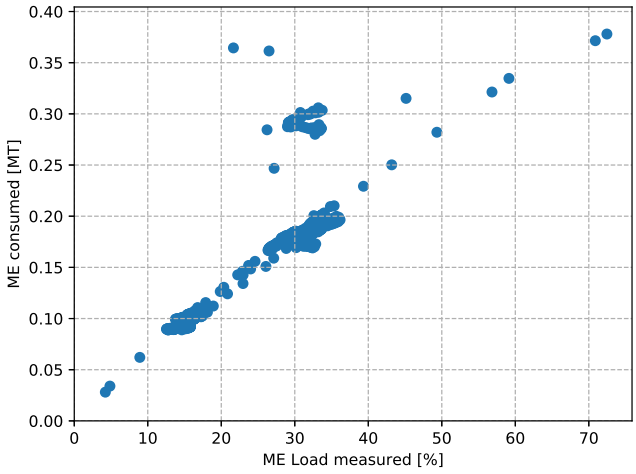

FIGURE 4: MEASURED ME FUEL CONSUMED (BETWEEN 2 SAMPLING INTERVALS) AS A FUNCTION OF MEASURED ME LOAD.

and transverse wind speeds. Also, the northward and eastward wind speeds obtained from ECMWF are transformed to longitudinal and transverse wind speeds. Assuming negligible speed of ship in transverse direction (i.e., no sway motion), the relative transverse wind speed should match with transverse wind speed.

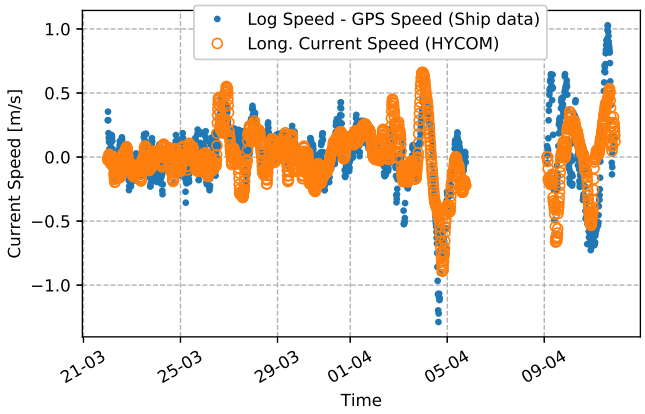

FIGURE 5: COMPARISON OF LONGITUDINAL SEA CURRENT SPEED ESTIMATED FROM SHIP DATA AND HINDCAST (HYCOM) MODEL FOR 'SEA PASSAGE' STATE.

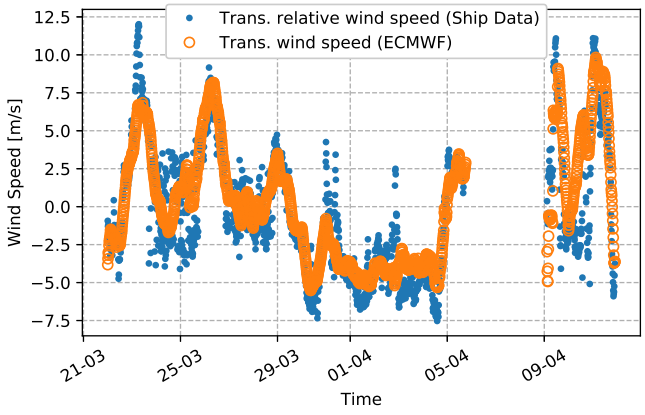

FIGURE 6: COMPARISON OF TRANSVERSE WIND SPEED OBTAINED FROM SHIP DATA AND HINDCAST (ECMWF) MODEL FOR 'SEA PASSAGE' STATE.

Similarly, it is possible to compare the longitudinal wind speed (from ECMWF) with the difference between relative longitudinal wind speed and GPS speed (from ship's data).

Figure 6 and 7 shows the comparison for transverse and longitudinal wind speeds from the two data sources. The two sources of data are seen to be in quite good agreement but the values obtained from ship's data seems to be unreliable, specially in case of longitudinal wind speed as the wind speed is changing sign or direction without any probable cause. Thus, the

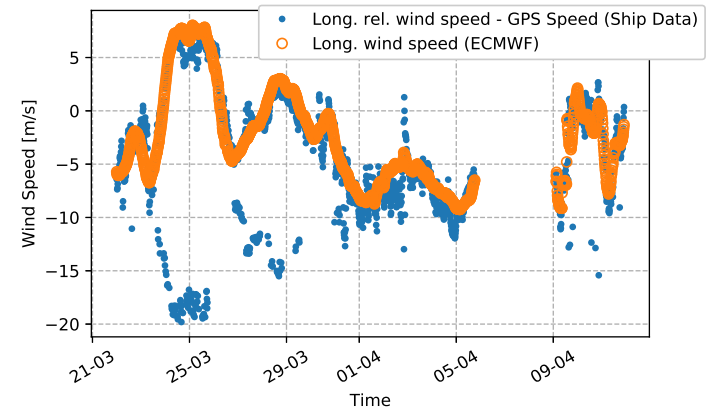

FIGURE 7: COMPARISON OF LONGITUDINAL WIND SPEED OBTAINED FROM SHIP DATA AND HINDCAST (ECMWF) MODEL FOR 'SEA PASSAGE' STATE.

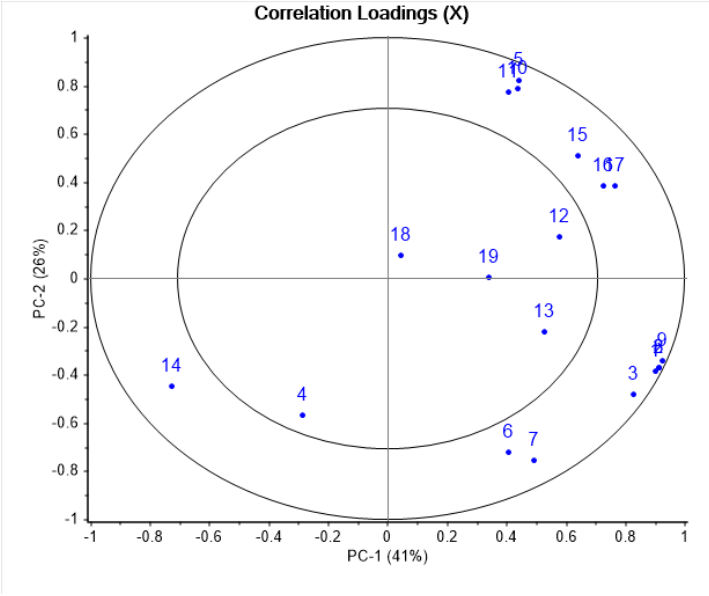

FIGURE 8: PRELIMINARY PCA MODEL: GRAPHICAL REPRESENTATION OF CORRELATION LOADINGS FOR INPUT VARIABLES (TABLE 2) IN PC-1 VS. PC-2 SPACE.

wind speed data obtained from ship's data is not included in the analysis any further.

TABLE 2: PRELIMINARY PCA MODEL: CORRELATION LOADINGS. SHOWING THE CORRELATION BETWEEN PRINCIPAL COMPONENTS AND INPUT VARIABLES. RED COLOR INDICATES STRONG CORRELATION WHILE YELLOW INDICATES NIL CORRELATION.

Sl. No.	Variables	PC-1	PC-2
1	ME consumed	0.9017	-0.3826
2	Shaft power	0.9121	-0.3727
3	Shaft rpm	0.8280	-0.4792
4	Draft fore	-0.2879	-0.5690
5	Draft aft	0.4425	0.8232
6	GPS speed	0.4075	-0.7195
7	Log speed	0.4939	-0.7568
8	ME Load measured	0.9121	-0.3727
9	Shaft Torque	0.9239	-0.3403
10	Mean draft	0.4380	0.7883
11	Trim-by-aft	0.4087	0.7724
12	Long. wind speed	0.5769	0.1749
13	Trans. wind speed	0.5277	-0.2213
14	Relative mean wave direction	-0.7242	-0.4460
15	Significant wave height	0.6416	0.5118
16	Mean wave direction	0.7243	0.3832
17	Mean wave period	0.7653	0.3862
18	Long. current speed	0.0437	0.0957
19	Trans. current speed	0.3393	0.0034

RESULTS

Based on the available dataset and observations made during data exploration, a preliminary Principal Component Analysis (PCA) model was created including 19 variables and 1688 samples from 'Sea Passage' state only. Table 2 presents the list of variables included in the preliminary PCA model and the obtained correlation loadings for each variable with PC-1 and PC-2. Figure 8 shows the correlation loadings in graphical format. The purpose of the preliminary PCA model is to do the following: (a) Check the correlation between variables and perform variable selection for final model; (b) Detect and investigate potential outliers.

Variable Selection

From the correlation loadings (Figure 8), it can be observed that longitudinal and transverse current speeds are very loosely correlated with the PCs as well as with other variables. In other words, longitudinal and transverse current speeds do not contribute much to the model. Thus, it is better to remove these variables from the analysis to get a better fitting and compact model.

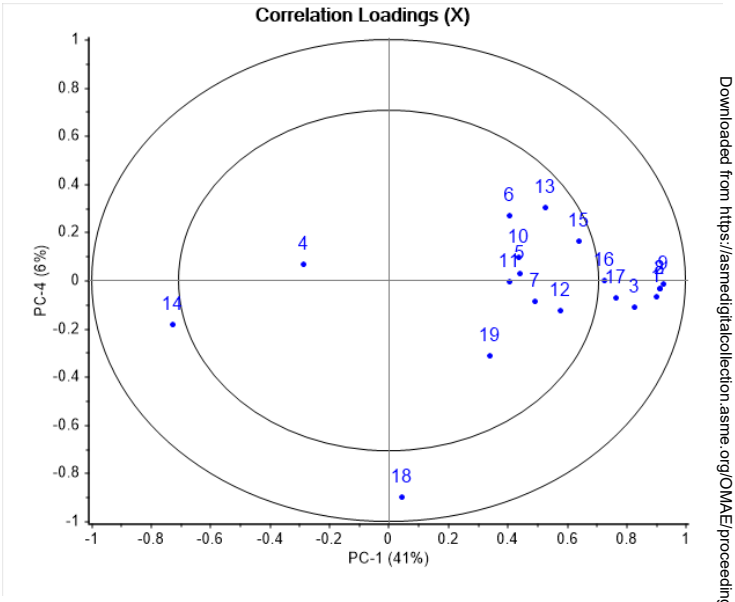

FIGURE 9: PRELIMINARY PCA MODEL: GRAPHICAL REPRESENTATION OF CORRELATION LOADINGS FOR INPUT VARIABLES (TABLE 2) IN PC-1 VS. PC-4 SPACE.

A consequence of not removing an uncorrelated variable can be understood as follows. The primary aim of a PCA model is to explain the variance in the complete dataset via minimum number of Principal Components (PCs). Therefore, the variance in any uncorrelated variable must also be explained. So as to achieve this, the model will create an extra PC just to explain the variance in this uncorrelated variable. Figure 9 shows that PC-4 is the undesirable extra PC created by the model to explain the variance in longitudinal current speed.

Outliers

Figure 10 shows the influence plot for PC-1 with 5% confidence limits. The samples marked by circles are potential outliers as they have high residuals and high influence on the model. The residuals are calculated as Q-residuals and the influence is calculated as Hotelling's T2 values. Further investigation revealed that these potential outlier are different from the rest of the sample set. They have either too low or too high shaft rpm than the rest of the samples, as shown in Figure 11. Thus, these samples are not erroneous values, rather they are very rare samples which are entirely different from the rest of the sample set.

The best way to deal with this type of problem is to gather additional samples which are similar to such rare samples. Since this is not possible here, it is better to remove these samples so

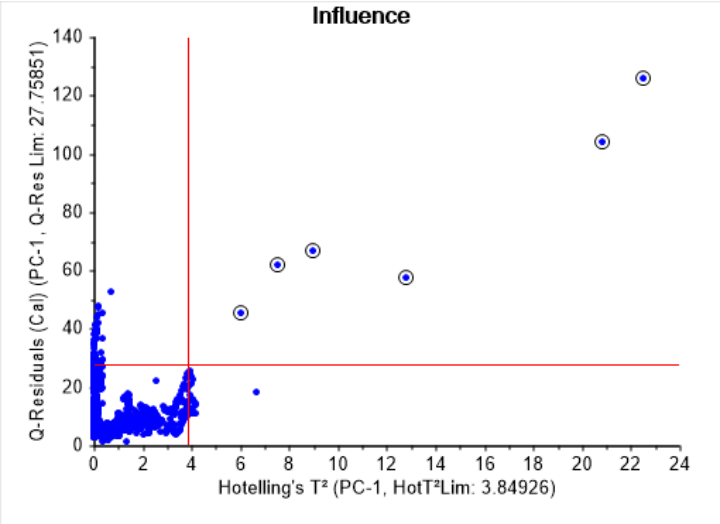

FIGURE 10: PRELIMINARY PCA MODEL: PC-1 INFLUENCE PLOT WITH 5% CONFIDENCE LIMITS SHOWING POTENTIAL OUTLIERS (MARKED BY CIRCLES TO THE TOP-RIGHT OF RED LINES).

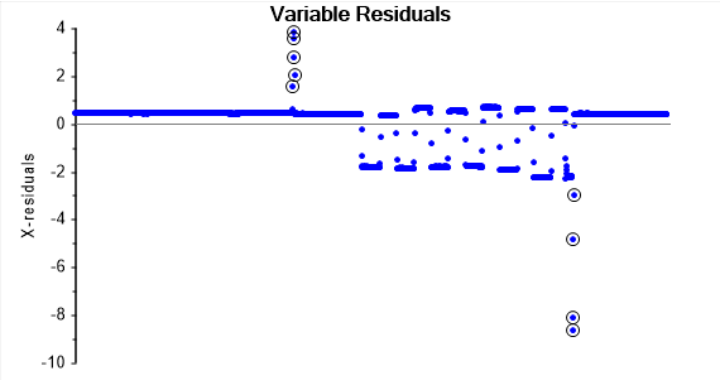

FIGURE 11: SCALED AND MEAN-CENTERED SHAFT RMP. SAMPLES MARKED BY CIRCLE ARE RARE AS THEY LIE FAR AWAY FROM THE MEAN (ZERO) LINE AS COMPARED TO REST OF THE SAMPLE SET.

that the model fits better to the remaining sample set. In view of this, all such rare samples were removed from the final PCA model.

PCA Model

Scores & Loadings. Based on the results from the preliminary PCA model, only 17 variables are included in the final PCA model. Table 3 shows the correlation loadings for first 7 Principal Components (PCs) calculated by the model. Figure 12

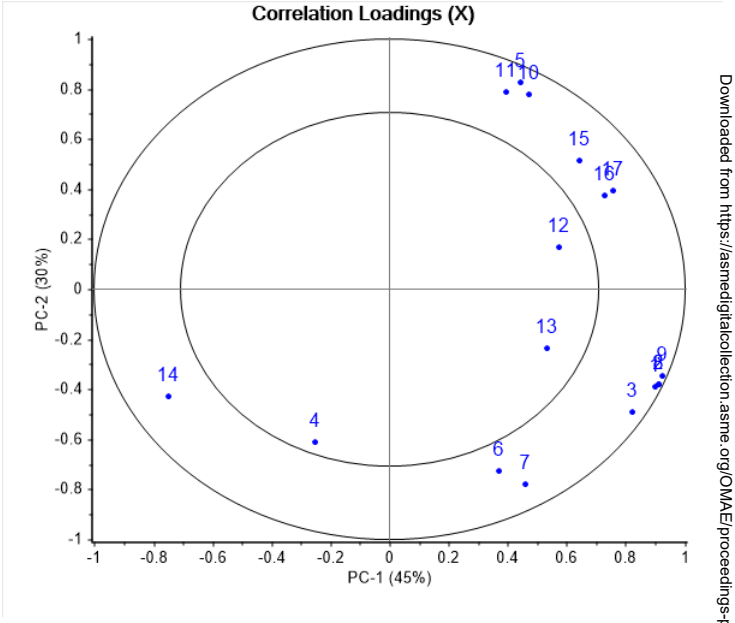

FIGURE 12: FINAL PCA MODEL: GRAPHICAL REPRESENTATION OF CORRELATION LOADINGS FOR INPUT VARIABLES (TABLE 3) IN PC-1 VS. PC-2 SPACE.

presents the correlation loadings for PC-1 and PC-2 in graphical format. It is observed that the 17 input variables are separated into 4 main groups of strongly correlated variable: Power parameters (1-3, 8, 9), wave parameters (14-17), draft parameters (4, 5, 10, 11) and speed parameters (6, 7). Each of these groups are oriented along a different direction in PC-1 vs. PC-2 space but none of these groups are completely aligned with either PC-1 or PC-2.

Similar observations can drawn in case of PC-3. This makes it difficult to interpret the physical meaning of these PCs. Additionally, it should be observed that PC-5 to PC-7 do not show good correlation with any of the variables, indicating that they are mostly representing noise in the given dataset. Thus, it is possible to retain only first 4 Principal Components and discard the remaining.

From Figure 12, it is surprising to observe that transverse wind speed (13) is correlated with PC-1 and PC-2 almost as much as longitudinal wind speed (12). It even shows higher correlation with PC-3 (refer Table 3), indicating that it is an important parameter for this model.

Explained Variance & Validation. Figure 13 presents the explained variance for the PCA model. The model can explain about 90% variance with just 4 PCs and further PCs do not really contribute to the model. Thus, the model with only first 4 PCs is a very good fit for the given dataset. Figure 13 also shows

TABLE 3: FINAL PCA MODEL: CORRELATION LOADINGS. SHOWING THE CORRELATION BETWEEN PRINCIPAL COMPONENTS AND INPUT VARIABLES. RED COLOR INDICATES STRONG CORRELATION WHILE YELLOW INDICATES NIL CORRELATION.

Sl. No.	Variables	PC-1	PC-2	PC-3	PC-4	PC-5	PC-6	PC-7
1	ME consumed	0.9016	-0.3893	0.1181	-0.0774	0.0127	-0.0149	0.0784
2	Shaft power	0.9126	-0.3792	0.0633	-0.0636	0.0102	-0.0245	0.0760
3	Shaft rpm	0.8247	-0.4927	0.2330	-0.0609	0.0462	-0.0039	0.0816
4	Draft fore	-0.2544	-0.6107	-0.6950	-0.0257	0.0512	-0.2620	0.0098
5	Draft aft	0.4427	0.8281	0.2983	-0.0109	-0.1109	-0.0999	0.0270
6	GPS speed	0.3706	-0.7254	0.2980	0.1898	-0.1275	-0.0730	-0.4044
7	Log speed	0.4608	-0.7770	0.3185	0.1361	0.0353	-0.0497	0.0418
8	ME Load measured	0.9126	-0.3792	0.0633	-0.0636	0.0102	-0.0245	0.0760
9	Shaft Torque	0.9226	-0.3468	0.0182	-0.0675	0.0028	-0.0304	0.0761
10	Mean draft	0.4721	0.7781	-0.0890	-0.0364	-0.1279	-0.3548	0.0484
11	Trim-by-aft	0.3965	0.7897	0.4532	0.0016	-0.0949	0.0240	0.0152
12	Long. wind speed	0.5750	0.1699	-0.2688	-0.7265	-0.0422	0.0865	-0.1020
13	Trans. wind speed	0.5333	-0.2347	-0.5990	0.2794	-0.3868	0.1609	0.0988
14	Relative mean wave direction	-0.7482	-0.4300	0.2344	0.0012	0.2447	-0.0698	0.1753
15	Significant wave height	0.6418	0.5128	-0.1387	0.4909	0.0795	-0.0085	0.0217
16	Mean wave direction	0.7295	0.3732	-0.3334	0.0717	0.3372	-0.0637	-0.1637
17	Mean wave period	0.7556	0.3947	-0.2025	0.1089	0.3138	0.1840	-0.0158

the explained variance for validation dataset. Model validation is done using cross-validation technique with 20 segments of randomly picked nonconsecutive samples, each segment containing about 145 samples. The validation dataset presents similar results as the calibration dataset.

Interpreting PCs. The simplest way to interpret Principal Components (PCs) is by looking at the correlation loadings or observing trends in sample score space, say, by means of sample grouping the scores. Sample grouping could not be used in the given case due to the complexity of PCs but by looking at correlation loadings (Table 3), it can said that PC-4 is mainly a combination of longitudinal wind speed (12) and significant wave height (15), i.e., it signifies the severity of environmental loads. Thus, a sample with high score for PC-4 would represent high environmental loads. PC-1, PC-2 and PC-3 are a combination of many variables as clearly observed from Table 3 and Figure 12.

It is also possible to understand the physical meaning of PCs from variable contributions point of view, i.e., by looking at variable residuals for all the PCs. Figure 14 shows that shaft power contributes to PC-1 and PC-2 only whereas Figure 15 shows that trim-by-aft contributes to PC-1, PC-2 and PC-3.

CONCLUSION

A data-driven mathematical approach was used to process the high dimensional sensor data recorded onboard a ship during a sea voyage. Principal Component Analysis (PCA) was used to perform variable selection and detect potential outliers. The high dimensional dataset obtained from the sensors onboard the ship and weather hindcast, representing the hydrodynamic performance of the ship, was greatly reduced in dimensions by PCA. The PCA model achieved upto 90% explained variance with only 4 Principal Components (PCs).

Wind and sea current hindcast data, obtained from ECMWF and HYCOM respectively, was found to be in good agreement and, in some cases, more reliable than the in-service measurements recorded onboard the ship. The sea current speed variables were eliminated during the variable selection process as they did not contribute to the PCA model for the given dataset. Transverse wind speed was observed to be an important parameter for the given dataset. Investigation needs to be done on a larger dataset in order to draw any further conclusions.

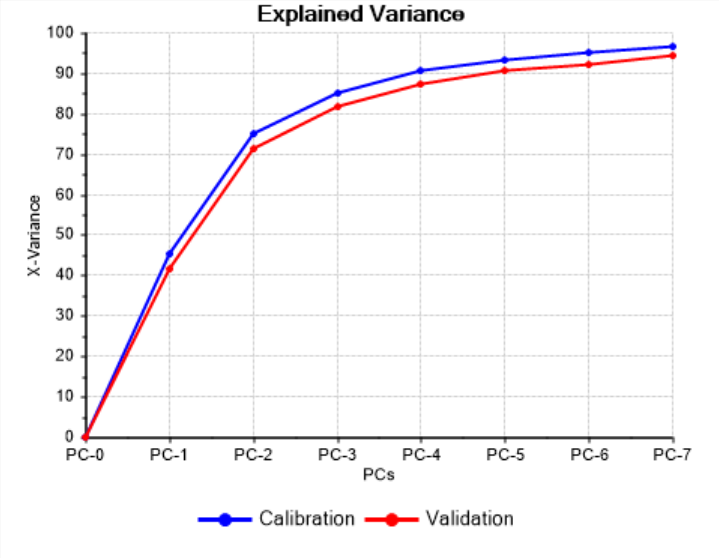

FIGURE 13: FINAL PCA MODEL: EXPLAINED VARIANCE. SHOWING THE VARIANCE IN DATASET ABSORBED BY CONSECUTIVE PCS. FIRST 4 PCS EXPLAINING ABOUT 90% VARIANCE.

FUTURE WORK

The PCA model presented in the current work is developed using the data recorded onboard a ship for about a month long period. A similar model can be easily developed on a larger set of data and for a different vessel. It would be interesting to corroborate the current findings for new and variant sets of such data.

Based on the current model, it is possible to quantify the performance of the vessel using the location of a sample in Principal Component Analysis (PCA) score space. But, in order to do so, a benchmark or standard basis needs to be established in PCA score space to mark, say, 100% performance. Alternatively, it is possible to develop a regression model based on the current analysis model, as demonstrated by Massy (1965) [22]. The regression model would be able to predict the speed or fuel consumption for a given state of the ship. Thus, it would be possible to quantify the performance of the ship in terms of speed loss or excess fuel consumption.

ACKNOWLEDGMENT

The authors would like to acknowledge using The Unscrambler X, a very useful commercial application developed by Prof. Harald Martens and CAMO Software.

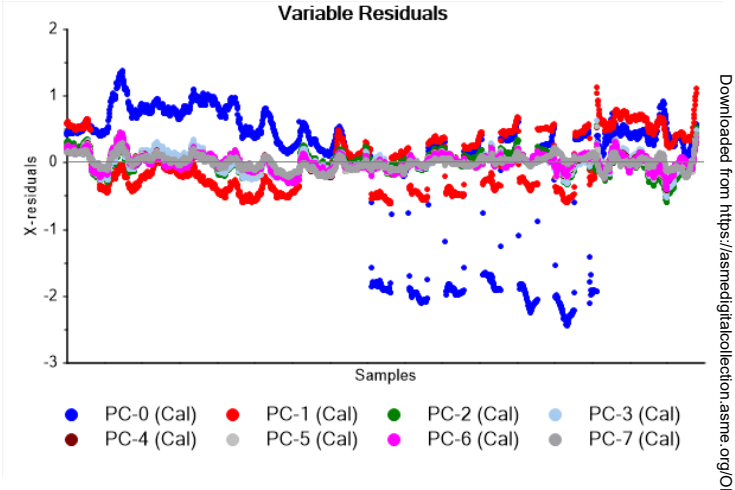

FIGURE 14: FINAL PCA MODEL: SHAFT POWER RESIDUALS. SHOWING THE VARIANCE IN SHAFT POWER ABSORBED BY CONSECUTIVE PCS. FIRST 2 PCS EXPLAINING MOST OF THE VARIANCE IN SHAFT POWER.

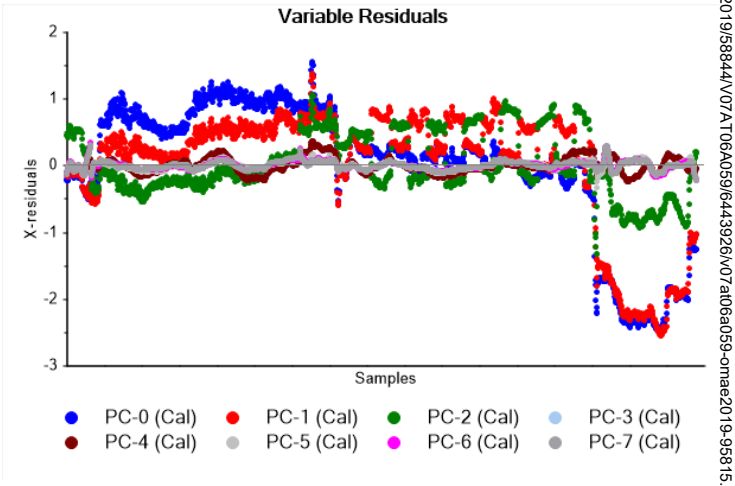

FIGURE 15: FINAL PCA MODEL: TRIM-BY-AFT RESIDUALS. SHOWING THE VARIANCE IN TRIM-BY-AFT ABSORBED BY CONSECUTIVE PCS. FIRST 3 PCS EXPLAINING MOST OF THE VARIANCE IN TRIM-BY-AFT.

REFERENCES

[1] van den Boom, H., and van der Hout, I., 2008. "Speed-power performance of ships during trials and in service".

[2] Prpić-Oršić, J., and Faltinsen, O., 2012. "Estimation of ship speed loss and associated co2 emissions in a seaway". *Ocean Engineering,* ***44***, 04, pp. 1–10.

[3] Feng, P., Ma, N., and Gu, X., 2010. "Long-term prediction

of speed reduction due to waves and fuel consumption of a ship at actual seas". Vol. 4, pp. 199–208. cited By 4.

[4] Lu, L., Mao, Y.-T., and Hsin, C.-Y., 2018. "Computation of the speed loss in seaway by different approaches". Vol. 2018-June, pp. 128–134. cited By 0.

[5] Seo, M.-G., Park, D.-M., Yang, K.-K., and Kim, Y., 2013. "Comparative study on computation of ship added resistance in waves". *Ocean Engineering,* ***73***, pp. 1–15. cited By 33.

[6] Kim, S.-W., Kim, J.-H., Seo, M.-G., Choi, J.-W., Lee, Y.-B., and Han, S.-K., 2017. "Assessment of ship operating performance by using full scale measurement". pp. 1001–1006. cited By 0.

[7] Mao, W., Rychlik, I., Wallin, J., and Storhaug, G., 2016. "Statistical models for the speed prediction of a container ship". *Ocean Engineering,* ***126***, pp. 152–162. cited By 4.

[8] Pedersen, B. P., 2014. "Data-driven vessel performance monitoring". PhD thesis.

[9] Gjølme, J. C., 2017. "Estimation of Speed Loss due to Current, Wind and Waves". MS Thesis, Norwegian University of Science and Technology (NTNU), Trondheim, NO, June. "See also URL `http://hdl.handle.net/11250/2453420`".

[10] Bal Beşikçi, E., Arslan, O., Turan, O., and Ölçer, A., 2016. "An artificial neural network based decision support system for energy efficient ship operations". *Computers and Operations Research,* ***66***, pp. 393–401. cited By 23.

[11] Perera, L., 2017. "Handling big data in ship performance and navigation monitoring".

[12] Sagiroglu, S., and Sinanc, D., 2013. "Big data: A review". pp. 42–47. cited By 301.

[13] Holzinger, A., 2018. "From machine learning to explainable ai". pp. 55–66.

[14] Brinton, C., 2017. "A framework for explanation of machine learning decisions". In IJCAI 2017 Workshop on Explainable Artificial Intelligence.

[15] Jolliffe, I., 2002. *Principal Component Analysis*. Springer Series in Statistics. Springer.

[16] Turing, A., 1948. "Rounding-off errors in matrix processes". *Quarterly Journal of Mechanics and Applied Mathematics,* ***1***(1), pp. 287–308. cited By 157.

[17] Golub, G., and Reinsch, C., 1970. "Singular value decomposition and least squares solutions". *Numerische Mathematik,* ***14***(5), pp. 403–420. cited By 1464.

[18] Vandeginste, B., Sielhorst, C., and Gerritsen, M., 1988. "Nipals algorithm for the calculation of the principal components of a matrix". *TrAC - Trends in Analytical Chemistry,* ***7***(8), pp. 286–287. cited By 17.

[19] Martens, H., and Martens, M., 2001. *Multivariate Analysis of Quality: An Introduction*. Wiley.

[20] Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hersbach, H., Hólm, E., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., Mcnally, A., Monge-Sanz, B., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F., 2011. "The era-interim reanalysis: Configuration and performance of the data assimilation system". *Quarterly Journal of the Royal Meteorological Society,* ***137***(656), pp. 553–597. cited By 9796.

[21] Chassignet, E., Hurlburt, H., Smedstad, O., Halliwell, G., Hogan, P., Wallcraft, A., Baraille, R., and Bleck, R., 2007. "The hycom (hybrid coordinate ocean model) data assimilative system". *Journal of Marine Systems,* ***65***(1-4 SPEC. ISS.), pp. 60–83. cited By 362.

[22] Massy, W., 1965. "Principal components regression in exploratory statistical research". *Journal of the American Statistical Association,* ***60***(309), pp. 234–256. cited By 398.