AN APPLICATION OF MACHINE LEARNING TO SHIPPING EMISSION INVENTORY

(DOI No: 10.3940/rina.ijme.2018.a4.500)

T Fletcher, **V Garaniya**, and **S Chai**, National Centre for Maritime Engineering and Hydrodynamics, Australian Maritime College, University of Tasmania, Australia, **R Abbassi**, School of Engineering, Faculty of Science and Engineering, Macquarie University, Australia, **H Yu**, **T C Van**, and **R J Brown**, School of Chemistry, Physics, and Mechanical Engineering, Queensland University of Technology, Australia, **F Khan**, Centre for Risk, Integrity and Safety Engineering, Faculty of Engineering & Applied Science, Memorial University of Newfoundland, Canada

SUMMARY

The objective of this study is to develop a shipping emission inventory model incorporating Machine Learning (ML) tools to estimate gaseous emissions. The tools enhance the emission inventories which currently rely on emission factors. The current inventories apply varied methodologies to estimate emissions with mixed accuracy. Comprehensive Bottom-up approach have the potential to provide very accurate results but require quality input. ML models have proven to be an accurate method of predicting responses for a set of data, with emission inventories an area unexplored with ML algorithms. Five ML models were applied to the emission data with the best-fit model judged based on comparing the real mean square errors and the R-values of each model. The primary gases studied are from a vessel measurement campaign in three modes of operation; berthing, manoeuvring, and cruising. The manoeuvring phase was identified as key for model selection for which two models performed best.

1. INTRODUCTION

Trading and the ocean have a relationship that stretches through most of recorded history. International shipping emissions, unlike other trading industries, are increasing due to the growth in demand, slow changing technology, and fairly unrestricted emissions regulations (Acomi & Acomi, 2014; Bencs et al., 2017; Chu et al., 2016; Helfre & Boot, 2013). Marine transportation emission data is scarce considering the size of the industry and number of vessels, leading to one of the largest uncontrolled sources of pollutants (Agrawal et al., 2008;).

The five primary pollutants emitted by international shipping are sulphur oxides (SO_x), Nitrogen oxides (NO_x), carbon dioxide (CO_2), particulate matter (PM), and carbon monoxide (CO). Shipping emission values differ with the different activity modes of the vessel including berthing, manoeuvring, and cruising. The source of vessel emissions is through the main engine combustion (Chu-Van et al., 2018; Fridell et al., 2008; Lack et al., 2009; Lindstad et al., 2013; Mueller et al., 2015).

Essential to the emission discussion for policymakers and scientists studying the effects of climate change is an accurate emission inventory (Endresen et al., 2005). Valuable emissions data includes the quantity of emitted pollutants, the relationship between pollutants and the causes of each, and finally the location of emissions. A thorough breakdown of emissions allows for a comprehensive analysis of the causes, mitigation options and future trends.

Emission inventory to estimate the amount of pollutants released into the atmospheric environment have been widely considered by different researchers (Dadashzadeh et al., 2011; Singer and Harley, 2012). Overall approaches for the estimation of emissions are either Top-down or Bottom-up (Smith et al., 2014). A top down approach uses overall fuel sales to estimate the emissions both internationally and domestically (Smith et al., 2014). The weakness of this approach is the uncertainty associated with the input data (Corbett & Koehler, 2004; Eyring et al., 2005; Olivier & Peters, 1999) and broad application of emission factors (Endresen et al., 2007). The Bottom-up approach uses individual vessel activity data and technical specifications to estimate the emissions by location (Smith et al., 2014). Current approaches applied in emission inventories, apply varied methodologies to estimate emissions for international and domestic shipping with varied accuracy (Miola & Ciuffo, 2011; Moreno-Gutiérrez et al., 2015). Inventories that focus on comprehensive Bottom-up approach have the potential to provide very accurate results due to the larger number of variables, the trade-off being the requirement for more detailed data which is usually difficult and costly to acquire with a high quality (Smit et al., 2010). A limitation to the Bottom-up approach is that the emissions factors applied to the input data are infrequently updated and are sourced from vessel campaigns not necessarily related to the technical details of applicable vessels (Goldsworthy & Galbally, 2011). Bottom-up emission approaches have been developed and implemented, including Emission registration and monitoring shipping (EMS) (Hulskotte & Denier van der Gon, 2010), Tier III (Smith et al., 2014), Methodologies for estimating air pollutant emissions from transport (MEET) (Hickman et al., 1999), Swedish methodology for environmental data (SMED) (Cooper & Gustafsson, 2004), National environmental research institute (NERI) (Olesen et al., 2009), and Monitoring programme on air pollution from sea-going vessels (MOPSEA) (Vangheluwe et al., 2007), Ship traffic emission assessment model (STEAM) (Jalkanen et al., 2009), and emissions from ocean shipping (Corbett & Koehler, 2003).

Alternatively, machine learning (ML) models have proven to be a robust and accurate method of predicting response values to a set of predictor data with applications in industries having previously used classical statistical models for prediction (Ahmed et al., 2010; Bontempi et al., 2013). ML is especially applicable with a dataset containing a complex nonlinear feature space, the advantage being the ability of the algorithms to identify complex patterns in a hyperspace with no prior knowledge or data classification (Li et al., 2016). The main concern in the creation of ML models is the reduction of model training error while maintaining model performance (generalisation capacity) and a low model complexity (Chalimourda et al., 2004).

Shipping emission inventories is an area that remains relatively unexplored with ML algorithms (Mohd Noor et al., 2016). Marine vessels have only been subject to ML research for the purpose of predicting engine speed and power (Chan & Chin, 2016; Mohd Noor et al., 2016). ML has been applied for prediction of diesel engine emissions from on-road sources. The research in this area commonly involves the use of detailed engine parameters measured in laboratory scenarios as predictor variables and this is not possible for a widely implemented vessel emissions inventory (Obodeh & Ajuwa, 2009; Shanmugama et al., 2011). Li et al. (2016) modelled the idling emissions of a light-duty gasoline vehicle using a variety of ML algorithms with the engine parameters and idling duration as input parameters in a process applicable to shipping emissions. In most on-road vehicle studies, an Artificial Neural Network (ANN) is used due to its performance with a large sample size and high quality data which they require to fully describe a condition (Rodriguez-Galiano et al., 2015). The drawback with shipping, and this study, is the limited access to data; therefore, the selected models have a greater flexibility to missing or limited data points.

The algorithms used in ML are classified as either supervised or unsupervised. Supervised learning is used when the features used to classify a finite set of observations are known (Bontempi et al., 2013; Dukart & Hoffmann-La Roche, 2015). Unsupervised learning has no predetermined pattern or relationships for classifying the dataset (Dukart & Hoffmann-La Roche, 2015). Developing the Bottom-up approaches from a classical statistical method to include the use of ML tools can provide more reliable results as models respond to a wide range of dynamic input variables. This research attempts to identify an ML algorithm for training an emission model based on collected vessel data.

2. METHODOLOGY

2.1 MACHINE LEARNING ALGORITHMS

The ML models were selected as they are state-of-the-art and represent the current trend of probabilistic regression model development. The following five ML models were applied to the shipping emission data to identify a best-fit model; Gaussian Process Regression (GPR), Principal Component Regression (PCA), Linear Dynamic System (LDS), Supervised Probabilistic Latent Variable Regression (SSPPLS), Supervised Mixture Probabilistic Latent Factor Regression (SMPPLS). The final prediction outcomes received by these models were compared by normalised Root Mean Square Error (RMSE) and the Pearson product-moment correlation coefficient R. The model with a lower RMSE and higher R-value is the best fit for purpose model. The ML tools considered in this study are discussed briefly in the following sections.

2.1 (a) Gaussian Process Regression (GPR)

GPR is a nonparametric Bayesian modelling technique (Ghassemi et al., 2015; Wang & Chaib-draa, 2017). Bayesian nonparametric models include prior distributions for the assistance in decision making by incorporating uncertainty into the design selection of the regression function Wang and Chaib-draa (2017); Woods et al. (2016). The regression function is treated as a smooth random function which consists of an infinite number of randomly and independently varying variables. The prior distribution for each of the infinite number of random variables is a Gaussian distribution, and the joint prior distribution of all the random variables is described by a prior Gaussian process. The prior Gaussian process can be considered as a random function generator. The variance of such a generator is specified by a kernel function. Such a prior Gaussian process can be updated with the observations of both the predictors and responses. As more data becomes available, the variance of the updated (posterior) Gaussian process reduces until it converges to the true regression function. The Gaussian process is flexible to model a large number of highly varying and smooth functions. However, the computational resources required to update the Gaussian process increases significantly with the size of the training data set. The Gaussian predictive distribution of y^* given a new nonlinear predictor x^* is found by;

$$p\left(y^* \middle| x^*, X, y;\ \hat{\alpha}, \hat{\beta}, \hat{\sigma}\right) \tag{1}$$

with the mean value defined by mean function $\mathrm{E}[f(x)] = m(x)$ and covariance defined by the covariance function $k(x, x') = \left[\left(f(x) - m(x)\right)\left(f(x') - m(x')\right)\right]$, where $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}$ are hyperparameters estimated using an iterative coordinate descent procedure. This model and the kernel elements are further explained by Yu (2017).

2.1 (b) Principal Component Regression Model (PCR)

PCA is an efficient algorithm for extracting latent features from high dimensional data (Yu et al., 2016). The model can be used to describe a single response variable for several predictor variables. The PCA feature-extraction procedure involves computing the covariance matrix from high dimensional data samples and Eigen decomposition of the covariance matrix for the eigenvectors called principal components (PCs). PCs that explain the majority of the variances are used to form the span of a subspace of a lower dimensionality (Yu et al., 2016). This subspace is referred to as the feature space through subspace projection (Yu et al., 2016). The drawback of a standard PCA model is the sensitivity to the scaling of process data. For samples collected with varying measurement scales, the result can be the extraction of irrelevant PCs (Yu et al., 2016). The PCA process is shown through the following equations; an Eigen decomposition of the original matrix X is performed;

$$V\Sigma_d V^\intercal = eig\left(\frac{X^\intercal X}{n-1}\right) \tag{2}$$

where V is the eigenvector matrix and Σ_d is the eigenvalue matrix. X is projected into the V vector space giving P. The predictor coefficient vector is estimated by$\widehat{W} = V(P^\intercal P)^{-1}P^\intercal y$ where y is a vector of corresponding occurrences. A normalised prediction y is given by applying W to a vector of process variables $x^\intercal$. A further examination of this model is found in (Yu, 2017) where both GPR and PCA are described and applied to a case study.

2.1 (c) Linear Dynamic System (LDS)

The modelling of dynamic and uncertain data sets can be addressed using dynamical Bayesian networks, or more specifically for a time-series state-space model an LDS that incorporates additional quality information is appropriate (Ge & Chen, 2016). An LDS model was introduced for a supervised fault detection case with multiple variables based upon the dual Kalman filtering model (Ge & Chen, 2016). The regression model extension applied in this study considers all response variables in a single model allowing for a complete solution to multiple emission elements where all previous data points are considered for each subsequent prediction. The expectation-maximization (EM) algorithm is used to estimate the maximum likelihood parameter. The EM algorithm is an iterated two-step method. The first step is the expectation step (E-step) where the posterior distribution of the latent variables is calculated according to the current parameter values. The second step is the maximization step (M-step) where the values of all parameters are updated by maximizing the expected log-likelihood function. The model structure of the LDS in a basic form is described by three equations;

$$h_t = Ah_{t-1} + \eta_t^h \tag{3}$$

$$x_t = B_x h_t + \eta_t^x \tag{4}$$

$$y_t = B_y h_t + \eta_t^y \tag{5}$$

where h_t is the latent variable, A is the transition matrix, η_t is the Gaussian noise for the respective variables, B_x and B_y are emission matrices for the respective variables. The conditional probability distributions of the latent and observed measurement are both Gaussian. The joint probability for the sequences of both latent variables h_t and observed variables o_T are both Gaussian distributed and are formulated as;

$$p(o_{1:T}, h_{1:T}) = p(h_1)p(o_1|h_1)\prod_{t=2}^{T} p(h_t|h_{t-1})p(o_t|h_t) \tag{6}$$

This model is further explained in Ge and Chen (2016) from which the model in this study is sourced.

2.1 (d) Supervised Probabilistic Latent Variable Regression Model (SSPPLS)

The Supervised Probabilistic Latent Variable Regression Model is developed as an advancement of the deterministic model such as PCA modelling (Ge, 2016). Like LDS, maximisation algorithms are implemented in both the single probabilistic model and the mixture form. The model is robust in its ability to handle missing data (Ge, 2016). The incorporation of factor analysis (FA) allows for the estimation of noise variance related to each predictor variable to be individually estimated (Ge, 2016). A Gaussian distribution is assumed for both the latent factor and the noise within each model and unity variance is assumed for the latent factor variable (Ge, 2016). The SSPPLS algorithm modelling structure is given as the following;

$$x = A_x t + e_x \tag{7}$$

$$y = A_y t + e_y \tag{8}$$

These rely upon the iterative EM process to maximise the likelihood function $E[L(X, Y, \Theta)]$ during the M-step of the EM algorithm by setting the partial derivatives for each parameter to zero allowing the calculation of the latent factors $\hat{t}_{new}$ and model predictions $\hat{y}_{new}$ for each new input data vector x_{new};

$$\hat{y}_{new} = A_y\hat{t}_{new} = A_y A_x^T(\Sigma_x + A_x A_x^T)^{-1}x_{new} \tag{9}$$

where A_x and A_y are the factor loading matrices, e_x and e_y are measurement noises of X and Y respectively, also Σ_x captures the different variable measurement noise variances.

2.1 (e) Mixture Probabilistic Latent Factor Regression Model (SMPPLS)

The Mixture Probabilistic Latent Factor Regression Model framework is an extension of the Supervised

Probabilistic Latent Variable Regression Model (Ge, 2016). The mixture modelling form, relative to the single latent model, is capable of handling complex nonlinear, multimode regression problems (Ge, 2016). Unlike the single model form, the mixture model does not assume a single Gaussian distribution; this model can be apply many distributions in application to multi-modal distributed data. The number of mixture components is related to the clustered relationship between predictor and response variable data. Several response variables can be predicted using a single model related to one or more time series predictor data sets. This algorithm is given by the following input x_i and output y_i variables;

$$x_i = \sum_{k=1}^{K} p(k) \left(\mu_{x,k} + A_{x,k} t_{i,k} + e_{x,i,k}\right) \quad (10)$$

$$y_i = \sum_{k=1}^{K} p(k) \left(\mu_{y,k} + A_{y,k} t_{i,k} + e_{y,i,k}\right) \quad (11)$$

where $p(k)$ is the mixing proportional value of each model, A_x and A_y are the factor loading matrices, t_k is the Gaussian latent factor vector, and $e_{x,k}$ and $e_{y,k}$ are Gaussian noise vectors. Following the EM algorithm where the log-likelihood function is maximised and the posterior probability $p(k|x_{new})$ for k model is determined;

$$p(k|x_{new}) = \frac{p(x_{new}|k)p(k)}{p(x_{new})} \quad (12)$$

The latent factor variables in each model t_k are used with the factor loading matrix $A_{y,k}$ to make local model predictions $y_{k,new}$;

$$y_{k,new} = A_{y,k} t_{k,new} = A_{y,k} A_{x,k}^T \left(\Sigma_{x,k} + A_{x,k} A_{x,k}^T\right)^{-1} x_{new} \quad (13)$$

The subsequent summation of prediction results from different models k provides the final weighted prediction;

$$y_{new} = \sum_{k=1}^{K} p(k|x_{new}) y_{k,new} \quad (14)$$

The full model description for the single and mixture model forms including the EM algorithm is found in (Ge, 2016). The characteristics of the five models are compared in Table 1 summarising the main elements of each as described in the above sections.

2.2 ACCURACY OF PREDICTED RESPONSES

Validation of a model is completed through cross validation where the original data is separated into one or more training, validation, and test subsets (Chalimourda et al., 2004; Dukart & Hoffmann-La Roche, 2015). Overfitting to a single dataset, and therefore low applicability to new data, occurs when the performance of the model parameters obtained from the training set are not validated with a separate dataset (Chalimourda et al., 2004; Dukart & Hoffmann-La Roche, 2015).

Table 1. Comparison of different ML tools used in this study

GPR	**PCA**	**LDS**	**SSPPLS**	**SMPPLS**
Single response variable	Multiple response variable	Multiple response variables	Multiple response variables	Multiple response variables
Flexible application of Gaussian distributions	Neglects insignificant input variables	Assumes Gaussian distribution	Assumes a single Gaussian distribution	Assumes multiple Gaussian distributions present
Performance decreases with data set size	Computationally inexpensive	Based on dual Kalman filtering model	Robust handling of missing data	Extension of SSPPLS model
Non-parametric model – Bayesian modelling technique	Sensitive to data scaling	Incorporates additional quality information (historical data points)	Factor analysis estimates individual input variable noise	Factor analysis estimates individual input variable noise
Fully specified by mean and covariance functions	Efficiently extracts latent features	Expectation-maximization (EM) algorithm	Expectation-maximization (EM) algorithm	Expectation-maximization (EM) algorithm
No predictor-response assumptions			Advancement of deterministic modelling	Handles complex nonlinear, multimode regression problems

The performance of the parameters obtained through modelling is calculated by the Mean Square Error and compared against other training subsets for a general error (Chalimourda et al., 2004). The Root Mean Square Error (RMSE) is the measure of the separation between the predicted response and the testing data response value. This error indicates alignment between the model and the expected values. A range normalised RMSE is used to compare different ML models and their performance during testing given by;

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{n}} \quad (15)$$

Where y_i and y_i^* are the actual emission and predicted emission at sampling interval i respectively, n is the total number of data points. The fitting level of predicted values to the observed values is measured by the R-value, which is obtained in Equation 16;

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(y_i^* - \overline{y^*})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(y_i^* - \overline{y^*})^2}} \quad (16)$$

Where $\bar{y}$ and $\overline{y^*}$ are the average actual and predicted emission levels. Ideally, the emission model with the lowest RMSE and total highest R-value is the model of best fit.

2.3 TEST PLAN AND DATA COLLECTION

The emissions emitted during the berthing, manoeuvring and cruising phases of the vessel are modelled in this study with the main engine as a consistent source for both vessels. The berthing emissions modelled are from a single vessel (vessel 1), vessel 2 emissions were excluded as the auxiliary engine was in use. The manoeuvring phase of operation has significant variation compared to the other phases. This phase is identified by low-speed manoeuvres conducted normally inside a harbour, exclusive of activity while berthed. The main engine for each of the two vessels is the source of all the data.

The data used in this study is from a previous measurement campaign described in Chu Van et al. (2016) . Two large cargo ships had measurements taken on both main and auxiliary engines in different operating conditions (at berth, manoeuvring, and at sea). The first ship transports from Port of Brisbane to Port of Gladstone, and the second ship from Gladstone to Newcastle. The emission of gaseous pollutants was collected as well as engine parameters; engine power, and shaft speed.

The collected engine test data recorded from vessel instrument panels; shaft speed (rpm) and shaft power (kW) is used as input variables. The emission data collected through installed instruments; NO, CO, CO_2, SO_2, and NO_X all in ppm units are the model outputs. Table 2 explains a list of the input and output variables. For each model, the vessel data for each phase was combined and randomly split into two parts for training and testing. Seventy percent of the data was randomly selected for training the models and thirty percent randomly selected for validation purposes.

3. RESULTS AND DISCUSSION

The emission prediction results from applying each of the five models are presented in Tables 3-7 and Figures 1-6. The same predictor variables are used in each of the models and all six-response variables are modelled as listed in Table 2. The models are compared for their normalised RMSE and R-values to allow for performance comparison between models. The results presented in Tables 3-7 are ranked by the total error for each model. The results of the modelling are illustrated as the individually observed emissions against each model's prediction in time series for each operational phase, and as a function of individual emissions observed and predicted against engine power for each operation phase.

Table 2. Monitored vessels input (X) and output (Y) variables.

Variable number	Variable description	Units
X_1	Shaft speed	rpm
X_2	Engine power	kW
Y_1	Nitrogen monoxide (NO)	ppm
Y_2	Nitrogen dioxides (NO_2)	ppm
Y_3	Nitrogen oxides (NO_X)	ppm
Y_4	Carbon monoxide (CO)	ppm
Y_5	Carbon dioxide (CO_2)	ppm
Y_6	Sulphur dioxide (SO_2)	ppm

Table 3. Normalised relative mean squared error for berthing prediction of the five ML emission models.

Rank	Model	NO	CO	NO_2	CO_2	NO_X	SO_2
1	SMPPLS	10.6%	1.2%	14.0%	8.8%	10.7%	8.2%
2	LDS	11.6%	1.3%	17.3%	9.5%	11.8%	9.1%
3	GPR	13.4%	1.2%	14.2%	9.6%	13.5%	8.8%
4	PCA	12.6%	1.3%	17.8%	9.6%	12.7%	8.9%
5	SSPPLS	13.8%	1.3%	19.9%	10.4%	13.9%	9.7%

Table 4. R-value for prediction of berthing emissions using five ML models.

Rank	Model	NO	CO	NO_2	CO_2	NO_X	SO_2
1	SMPPLS	0.82	0.37	0.81	0.83	0.82	0.83
2	GPR	0.72	0.47	0.80	0.80	0.72	0.80
3	LDS	0.80	0.28	0.72	0.81	0.80	0.80
4	PCA	0.74	0.34	0.67	0.79	0.73	0.79
5	SSPPLS	0.72	0.19	0.60	0.77	0.71	0.77

Table 5. Normalised relative mean squared error for manoeuvring prediction of the five ML emission models

Rank	Model	NO	CO	NO_2	CO_2	NO_X	SO_2
1	SMPPLS	8.8%	6.0%	12.4%	7.1%	8.8%	9.8%
2	GPR	13.3%	6.2%	9.9%	6.5%	11.9%	10.3%
3	PCA	8.9%	6.0%	18.1%	7.8%	9.0%	10.7%
4	SSPPLS	11.8%	6.0%	18.2%	8.0%	11.9%	11.3%
5	LDS	12.4%	6.0%	18.4%	8.4%	12.5%	11.6%

Table 6. R-value for prediction of manoeuvring emissions using five ML models.

Rank	Model	NO	CO	NO_2	CO_2	NO_X	SO_2
1	GPR	0.91	0.28	0.89	0.90	0.92	0.88
2	SMPPLS	0.94	0.23	0.82	0.88	0.93	0.89
3	PCA	0.93	0.08	0.58	0.85	0.93	0.86
4	LDS	0.89	0.16	0.58	0.83	0.89	0.84
5	SSPPLS	0.88	0.01	0.58	0.85	0.88	0.84

Table 7. Normalised relative mean squared error for cruising prediction of the five ML emission models.

Rank	Model	NO	CO	NO_2	CO_2	NO_X	SO_2
1	SMPPLS	2.2%	4.1%	2.0%	2.0%	2.2%	2.5%
2	PCA	2.5%	3.7%	2.4%	2.1%	2.5%	2.9%
3	SSPPLS	2.6%	3.8%	2.4%	2.2%	2.6%	3.0%
4	GPR	3.7%	3.4%	3.0%	3.6%	3.7%	3.6%
5	LDS	3.8%	4.1%	3.9%	3.9%	3.9%	3.8%

The agreement between the predicted and actual emissions for two of the models is more accurate than the others. The berthing phase values are for vessel 1 only as vessel 2 was recorded using its auxiliary engine rather than its main engine. The error values and R-value for this phase are a possible indication of the performance when applied to additional vessels. The value ranges are similar to those in the manoeuvring phase. However, the performance will depend on the actual activity during this operation. During berthing, the engine activity may vary between highly variable when loading, through to a consistent level for low activity times (Figure 1). These activity levels are each similar to manoeuvring and cruising respectively. Therefore, it is expected that the performance will range between these two points and an algorithm can be applied as selected through these two modes.

The manoeuvring phase of operations identified two models as superior to the others; SMPPLS and GPR. The SMPPLS model produced a lower error than the GPR model (Table 5) while both yielded similar R-values (Table 6). All five models performed similarly in predicting four of the six pollutants; NO, CO_2, NO_X, SO_2. The defining area of performance was in predicting CO and NO_2 in which both the SMPPLS and GPR models outperformed the other models allowing them to be disregarded from selection. The much smaller R-value of CO suggests alternative influences outside engine load may be affecting the emissions pattern. The high R-values for four of the pollutants indicate strong ability of the models to explain the vessel emissions with limited input variables, showing that a large portion of the emissions variance is explained by the engine operating levels.

Emission estimation in the cruising phase shows high correlation (Figure 6) and low error values (Table 7 and Figure 5) for all models. This can be attributed to constant engine operation and therefore constant emissions. The high correlation values are due to the lack of variation in the values creating two clusters of input vs output data as seen in Figure 6. These values are easily predicted by the models. All the models perform similarly when applied to input and output variables with lacking variation, therefore, this phase does not eliminate or promote one model over another.

Due to the nature of the results described above, the evaluation of the above five models in the three operational phases is dependent on the performance of each in the manoeuvring phase. The manoeuvring results indicate clearly that the SMPPLS and GPR models performed beyond the other three models. The poor performance of the LDS model indicates there is no temporal correlation between data points. Applying the model to an increasing number of vessels would require further input variables to be incorporated into the models as the pattern of emissions levels varies between ship types.

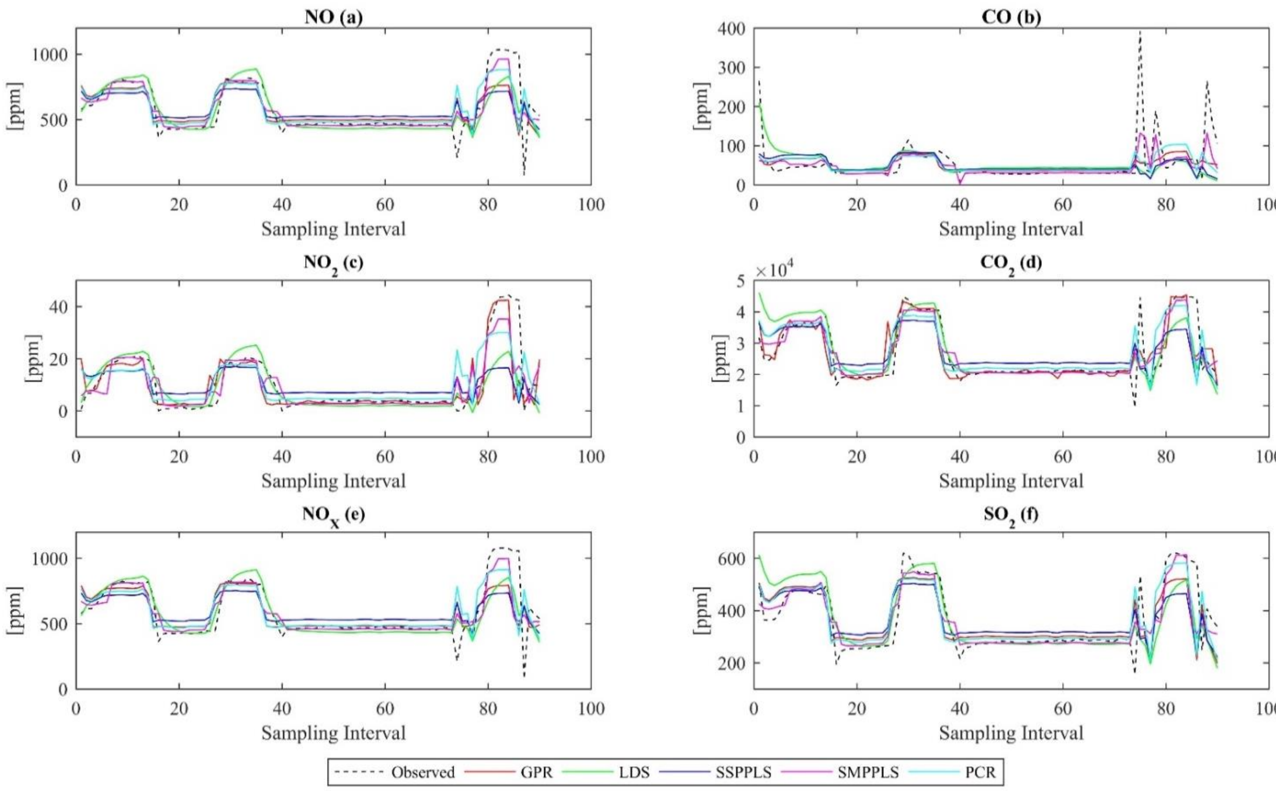

Figure 1: Berthing emissions estimates for all models broken time by pollutant; (a) NO, (b) CO, (c) NO_2, (d) CO_2, (e) NO_x, (f) SO_2

3.1 SMPPLS MODELLING RESULTS

The predicted emissions using the SMPPLS model for each pollutant rank either first or second with the GPR model for each operation phase. The SMPPLS in the manoeuvring phase performs similarly to the GPR model with high R-values in five of the six pollutants measured. The R-values are high for all but the pollutant CO, indicating strong correlations to the actual emissions for the prediction with CO being the only weakly correlated response. The SMPPLS performs well in this case due to the application of multiple Gaussian distributions to the multi-modal data seen in Figure 4 as distinguishable clusters. The advantage of the SMPPLS model, beyond its higher performance, lies in its design where it processes multiple time-series response data simultaneously, in this case, six different pollutant responses, whereas the GPR model requires separate models for each time-series response variable.

3.2 GPR MODELLING RESULTS

The manoeuvring emissions for the GPR model show the R-values for each pollutant, excepting CO, are high with values of between 0.84 and 0.94 indicating a strong correlation between predicted and actual emissions. The value of 0.28 for CO suggests a weak correlation between the predicted and actual emissions. The relative error for the GPR is higher than the SMPPLS model, however this difference is represented for the most part in the prediction of NO_2 and SO_2. In fact, the GPR has a lower or equal error value in four of the pollutant responses (NO, NO_2, CO_2, NO_X). The weakness of the GPR model is its limitation to only predict a single output for an input time series. The requirement to repeat modelling for each output, combined with decreasing performance with the size of a data set, indicates an increasingly computationally expensive model, unsuitable for wider implementation.

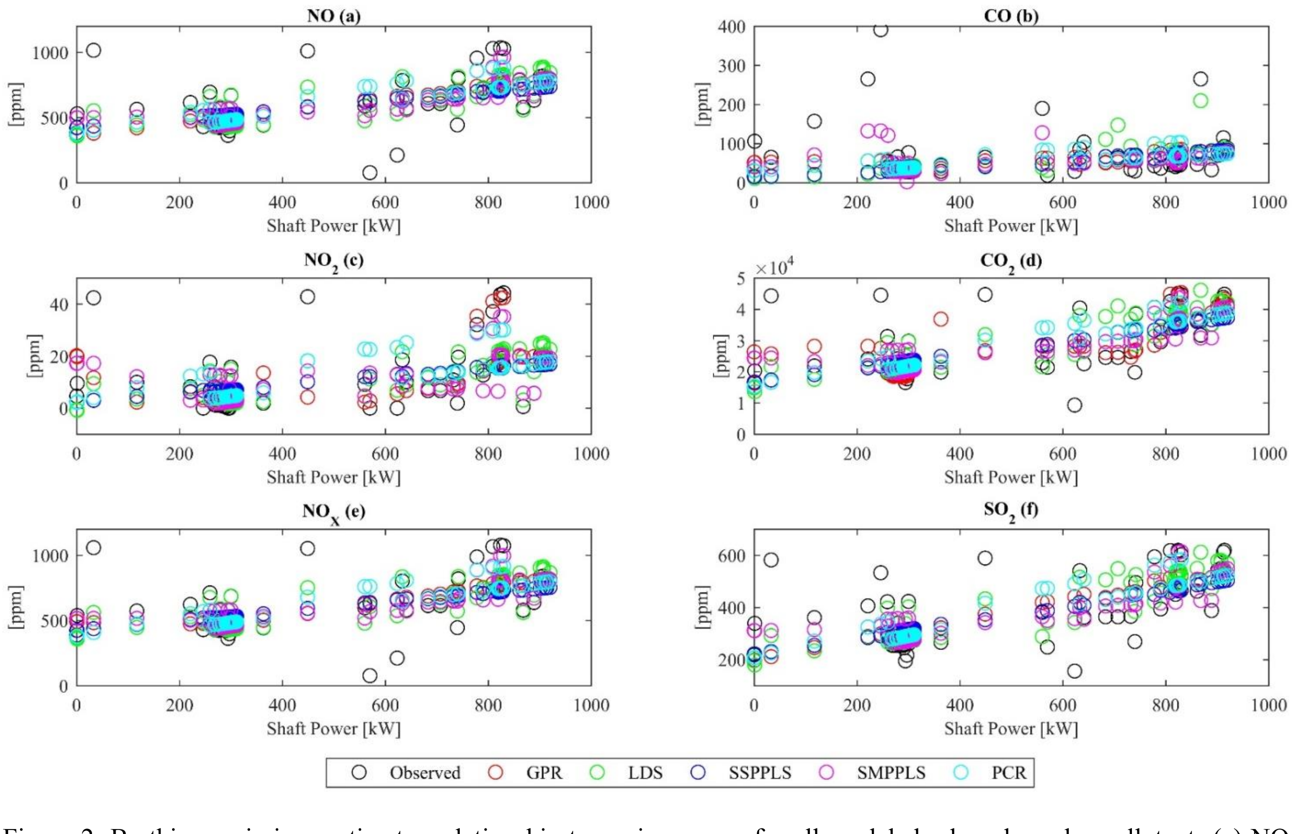

Figure 2: Berthing emissions estimates relationship to engine power for all models broken down by pollutant; (a) NO, (b) CO, (c) NO_2, (d) CO_2, (e) NO_x, (f) SO_2

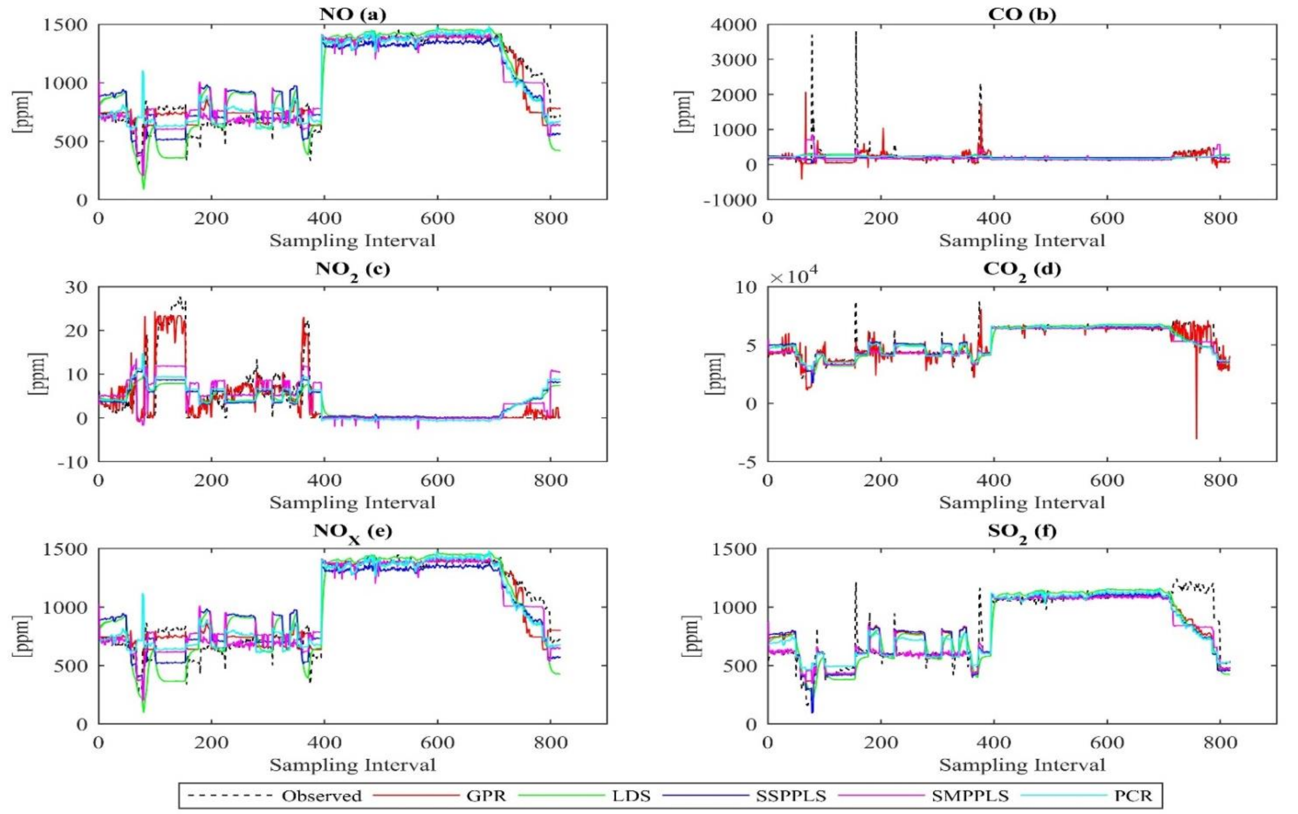

Figure 3: Manoeuvring emissions estimates for all models broken time by pollutant; (a) NO, (b) CO, (c) NO_2, (d) CO_2, (e) NO_x, (f) SO_2

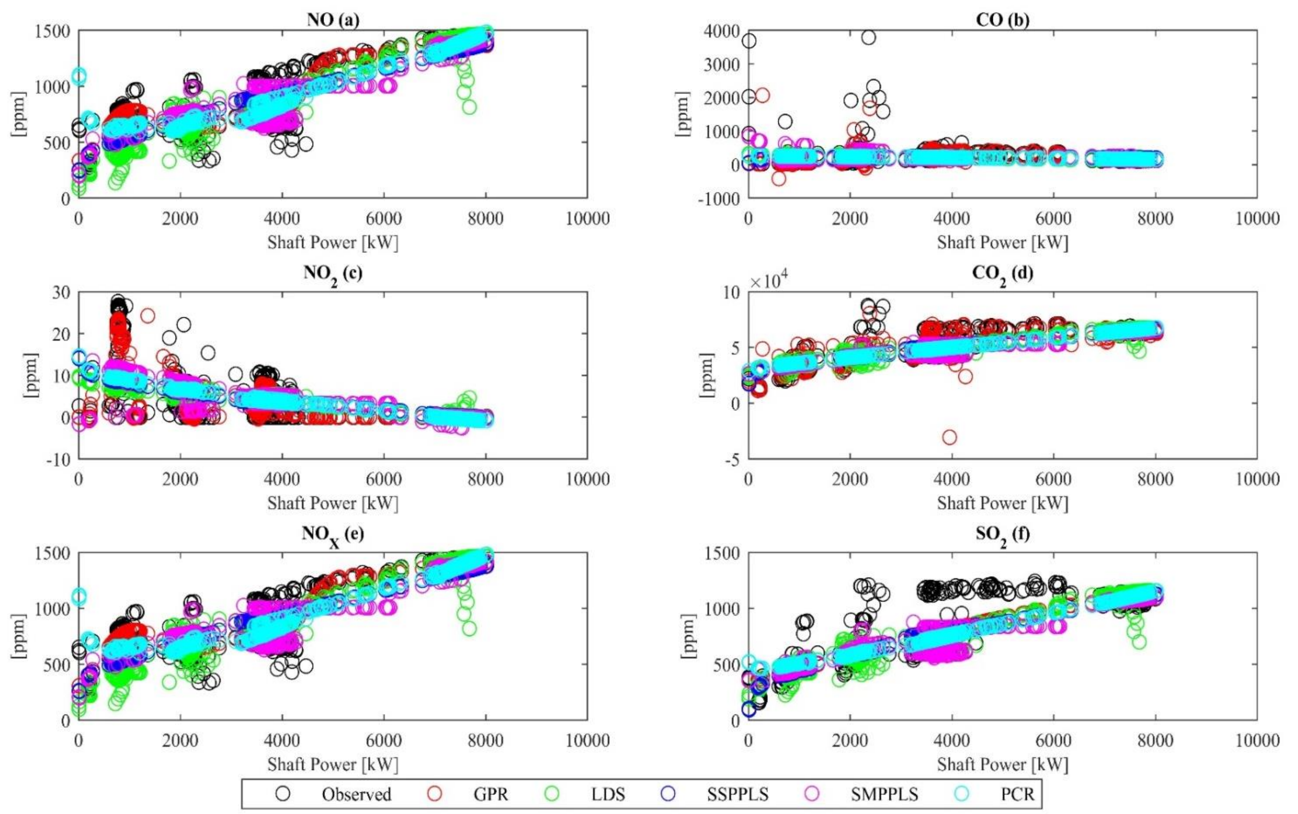

Figure 4: Manoeuvring emissions estimates relationship to engine power for all models broken down by pollutant; (a) NO, (b) CO, (c) NO_2, (d) CO_2, (e) NO_x, (f) SO_2

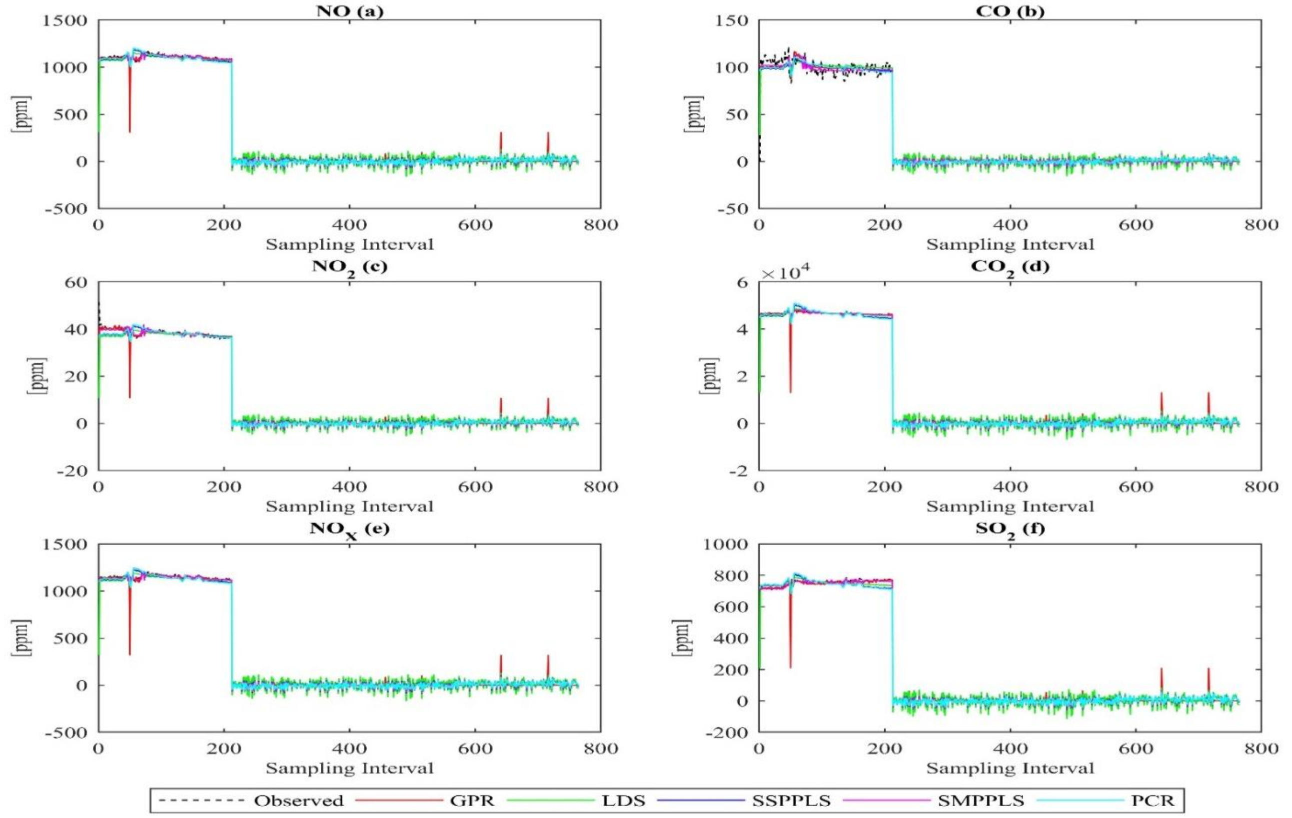

Figure 5: Cruising emissions estimates for all models broken time by pollutant; (a) NO, (b) CO, (c) NO_2, (d) CO_2, (e) NO_x, (f) SO_2

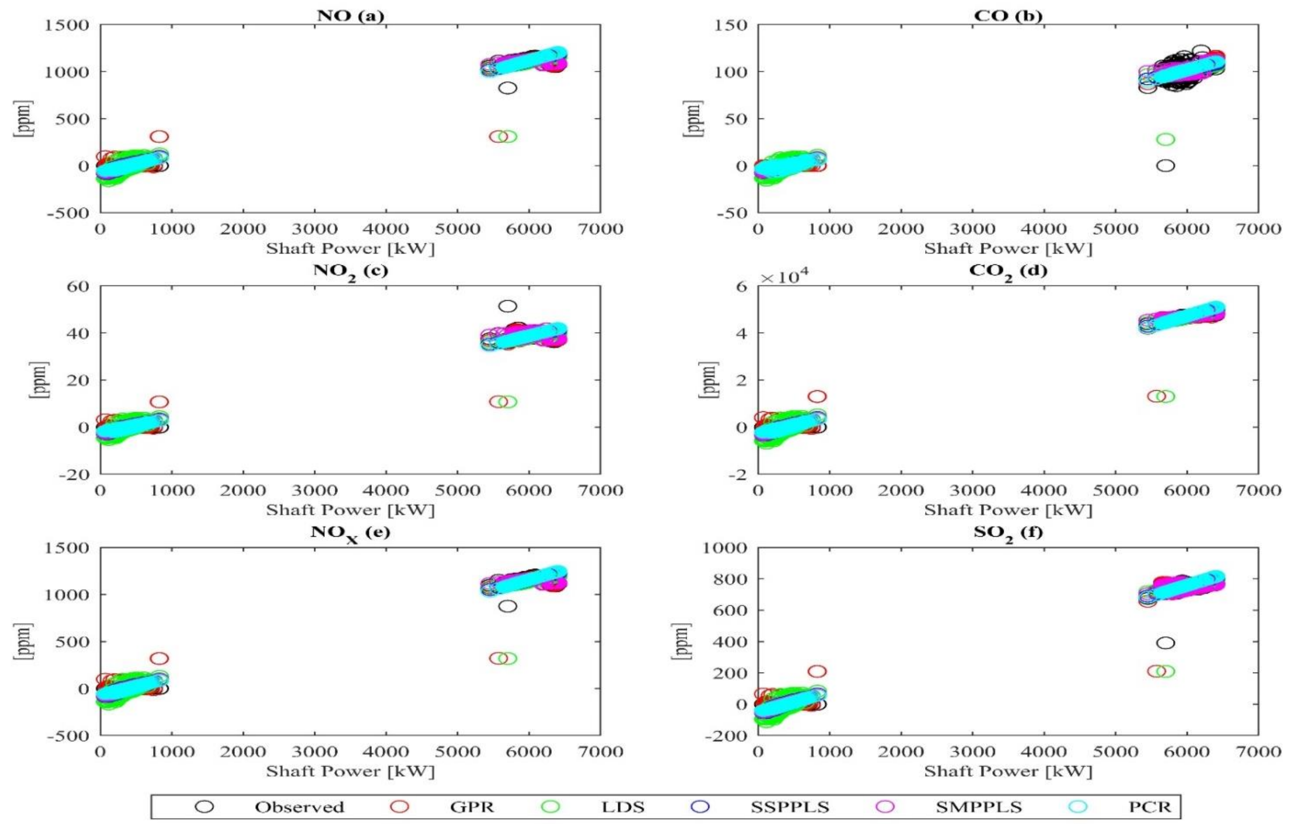

Figure 6: Cruising emissions estimates relationship to engine power for all models broken down by pollutant; (a) NO, (b) CO, (c) NO_2, (d) CO_2, (e) NO_x, (f) SO_2

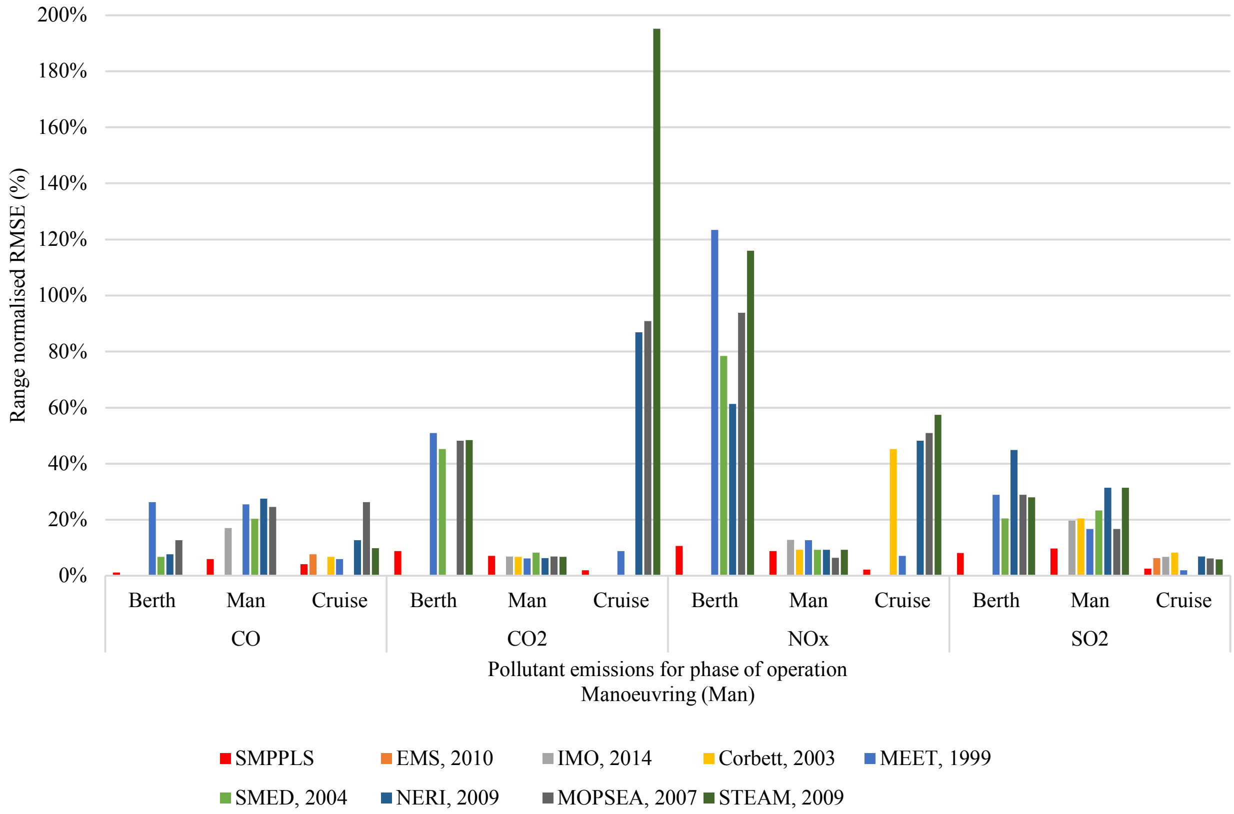

Figure 7: Range normalised RMSE for all phase predictions of SMPPLS ML model and existing shipping emission models

3.3 COMPARISON WITH EXISTING MODELS

The emission prediction results for the SMPPLS ML model are presented in Figure 7 alongside the pollutant emissions estimated by existing Bottom-up approaches (Cooper & Gustafsson, 2004; Corbett & Koehler, 2003; Hickman et al., 1999; Hulskotte & Denier van der Gon, 2010; Jalkanen et al., 2009; Olesen et al., 2009; Smith et al., 2014; Vangheluwe et al., 2007). The SMPPLS model was chosen for comparison due to its superior performance compared to the other ML models studied. The emissions estimation from each model are created from engine data from which all estimation in this study has been based. The range normalised RMSE is plotted for each model where emission factors are provided in each mode of operation for four pollutants (CO, CO_2, NO_x, SO_2).

The comparison between the models in Figure 7 illustrates that there is variation in the performance of each, depending on the mode of operation and pollutant studied. However, for each broken-down section, the ML model performed with lower error than the traditional Bottom-up approaches utilising an emission factor for estimation. The trends in performance when ML models are applied to fleet wide data are uncertain due to the required development required in model development when the variation in predictor parameters increases. The results shown are an indication of how the use of machine learning in estimating pollutant emissions can be an improvement to the performance of an emission inventory over the use of emission factors.

4. EMISSIONS INVENTORY

This section proposes a methodology for addressing the uncertainty that exists in current shipping emission inventories in the manoeuvring phase of operation. The emission of pollutants in an in-port environment differs from that of an at-sea environment, with engine loads varying between high and low power for short time periods. This variability has not been captured in traditional approaches with existing inventories typically estimating an average speed and duration for specific ports and applying a small sample size of emission measurements to the wider fleet.

A new approach that incorporates the use of ML algorithms is suggested which uses the engine load information of specific ship types and the corresponding emissions measurements from a sample of vessels to predict future emissions based on engine load data. This data is easily accessible from individual vessels although the measurement of emissions from individual vessels is an intensive process. The incorporation of an ML model trained on similar ship types allows for a non-intensive estimation of in-port shipping emissions in any geographic location.

The division of an emission inventory into in-port and at-sea inventories allows for the varied operating conditions to be targeted specifically. The steady state operation of vessels can be estimated using an extensive Automatic Identification System (AIS) based inventory such as that of Jalkanen et al. (2009). The increasing applicability of an inventory such as this will rely upon the distribution and accessibility of the AIS system. However, with an ML based inventory estimating in-port emissions, the data that local regulators and port authorities require to inform decisions and planning can be accurately available with a low cost beyond the collection of engine operation records for the period or location of interest. The engine data required to be modelled is measured and displayed digitally, thus storing this information in a format capable of distribution, if that is not already the case, is a simple matter.

A possible development of the at-sea emission inventory model of Jalkanen et al. (2009), where the comprehensive data modelling conducted for engine power estimation, would be an ideal input to an ML model for emissions prediction. This combination would allow the prediction of emissions without the use of emission factors. The difficulty of this suggested model is the requirement for measurement campaign data. However, the nature of the required measurement campaigns would be in line with those for emission factors, therefore existing data would be applicable for model development and any new effort for emission factor development would be an opportunity for extending emission inventories in two paths for more extensive validation studies.

5. CONCLUSIONS

With the aim of identifying a model to predict shipping emissions based on engine parameters, GPR, PCA, SMPPLS, SSPPLS, and LDS models are proposed. The emissions trained and tested in each model are NO, NO_X, SO, SO_2, CO, CO_2. Performance of the models has been evaluated by calculating the RMSE and R-value for each model. The results from the cruising phase illustrated the low variation in emission with constant engine activity and therefore had similar performance across the five models. The berthing emission results showed similar activity to the manoeuvring phase. However, it is acknowledged that the results are likely to be similar to either a cruising phase or manoeuvring depending on the current activity. The manoeuvring phase was therefore identified as the phase to dictate model performance. The predictions obtained from this phase were very good for all types of emissions, excepting CO, with two models; the SMPPLS and GPR exceeding the performance of the others. The SMPPLS model performs better by applying multiple Gaussian distributions to the multi-modal data and has an advantage over the GPR model in that it estimates multiple time-series variables simultaneously while the GPR model estimates each individually. The comparison of the SMPPLS model to existing Bottom-up approaches is an indication of the improved estimation ability of a model incorporating ML tools rather than emission factors. Applying the developed model as a specific in-port emissions inventory would allow for complex engine conditions to be captured and the accuracy of vessel emissions to be increased. The described methodology serving in addition to ‘at-sea’ inventories allows for a full measurement of shipping emissions.

6. FURTHER RESEARCH

The methodology and modelling described in this study are focussed on the manoeuvring emissions of two vessels. There are however many opportunities for expanding this program of study through ship types. The requirements for further study are for additional or existing emission measurement campaign data upon a greater and appropriate sample size of ship types with a variety of designs. Expanding study into an increasing number of ship types will require the capture of input variables such as fuel and engine data because as the variety of vessels increases, the variance in emissions requires more input variables to explain the variation. Areas of study appropriate for expansion lie in the pollutants monitored; particulate matter is a pollutant from shipping that is important for measurement and estimation due to its environmental and health effects. Trialling particulate matter emissions with ML algorithms would increase the quality and relevance of estimated emissions.

7. ACKNOWLEDGEMENTS

The author acknowledges the support of the International Association of Maritime Universities (IAMU) and the Nippon Foundation in Japan.

8. REFERENCES

1. ACOMI, N., & ACOMI, O. C. (2014). *The Influence of Different Types of Marine Fuel over the Energy Efficiency Operational Index*. Energy Procedia, 59(Supplement C), 243-248. doi:10.1016/j.egypro.2014.10.373
2. AGRAWAL, H., MALLOY, Q. G. J., WELCH, W. A., WAYNE MILLER, J., & COCKER, D. R. (2008). *In-use gaseous and particulate matter emissions from a modern ocean going container vessel*. Atmospheric Environment, *42*(21), 5504-5510. doi:10.1016/j.atmosenv.2008.02.053

3. AHMED, N. K., ATIYA, A. F., GAYAR, N. E., & EL-SHISHINY, H. (2010). *An Empirical Comparison of Machine Learning Models for Time Series Forecasting*. Econometric Reviews, 29(5-6), 594-621. doi:10.1080/07474938.2010.481556
4. BENCS, L., HOREMANS, B., JOLANTA BUCZYŃSKA, A., & VAN GRIEKEN, R. (2017). *Uneven distribution of inorganic pollutants in marine air originating from ocean-going ships*. Environmental Pollution, 222, 226-233. doi:10.1016/j.envpol.2016.12.052
5. BONTEMPI, G., BEN TAIEB, S., & LE BORGNE, Y.-A. (2013). *Machine Learning Strategies for Time Series Forecasting*. In M.-A. Aufaure & E. Zimányi (Eds.), Business Intelligence: Second European Summer School, eBISS 2012, Brussels, Belgium, July 15-21, 2012, Tutorial Lectures (pp. 62-77). Berlin, Heidelberg: Springer Berlin Heidelberg.
6. CHALIMOURDA, A., SCHÖLKOPF, B., & SMOLA, A. J. (2004). *Experimentally optimal ν in support vector regression for different noise models and parameter settings*. Neural Networks, 17(1), 127-141. doi:10.1016/S0893-6080(03)00209-0
7. CHAN, T. K., & CHIN, C. S. (2016, 2016/11/22-25). *Data analysis to predictive modeling of marine engine performance using machine learning*. Paper presented at the 2016 IEEE Region 10 Conference (TENCON).
8. CHU, T. V., RAINEY, T., RISTOVSKI, Z., POURKHESALIAN, M., GARANIYA, V., ABBASSI, R., YANG, L., & BROWN, R. J. (2016). *Emissions from a marine auxiliary diesel engine at berth using heavy fuel oil*. 10th Australasian Heat and Mass Transfer Conference, Brisbane, Qld, Australia.
9. CHU-VAN, T., RISTOVSKI, Z., POURKHESALIAN, A. M., RAINEY, T., GARANIYA, V., ABBASSI, R., JAHANGIRI, S., ENSHAEI, H., KAM, U. S., KIMBALL, R., YANG, L., ZARE, A., BARTLETT, H., & BROWN, R.J. (2018). *On-board measurements of particle and gaseous emissions from a large cargo vessel at different operating conditions*. Environmental Pollution, 237*:* 832-841.
10. COOPER, D., & GUSTAFSSON, T. (2004). *Methodology for calculating emissions from ships: 1*. Update of emission factors. In: IVL Swedish Environmental Research Institute.
11. CORBETT, J. J., & KOEHLER, H. W. (2003). *Updated emissions from ocean shipping*. Journal of Geophysical Research: Atmospheres, 108(D20). doi:10.1029/2003JD003751
12. CORBETT, J. J., & KOEHLER, H. W. (2004). *Considering alternative input parameters in an activity-based ship fuel consumption and emissions model*: Reply to comment by Øyvind Endresen et al. on "Updated emissions from ocean shipping". Journal of Geophysical Research: Atmospheres, *109*(D23), 1-7. doi:10.1029/2004JD005030
13. DADASHZADEH, M., KHAN, F., HAWBOLDT, K., & ABBASSI, R. (2011). *Emission factor estimation for oil and gas facilities*. Process Safety and Environmental Protection, 89: 295-299.
14. DUKART, J., & HOFFMANN-LA ROCHE, F. (2015). *Basic Concepts of Image Classification Algorithms Applied to Study Neurodegenerative Diseases*. In Brain Mapping: An Encyclopedic Reference (pp. 641-646). Waltham: Academic Press.
15. ENDRESEN, Ø., BAKKE, J., SØRGÅRD, E., BERGLEN, T., & HOLMVANG, P. (2005). *Improved modelling of ship SO2 emissions - A fuel-based approach*. Atmospheric Environment, 39(20), 3621-3628. doi:10.1016/j.atmosenv.2005.02.041
16. ENDRESEN, Ø., SØRGÅRD, E., BEHRENS, H. L., BRETT, P. O., & ISAKSEN, I. S. A. (2007). *A historical reconstruction of ships' fuel consumption and emissions*. Journal of Geophysical Research: Atmospheres, 112(D12). doi:10.1029/2006JD007630
17. EYRING, V., KÖHLER, H. W., VAN AARDENNE, J., & LAUER, A. (2005). *Emissions from international shipping: 1. The last 50 years*. Journal of Geophysical Research: Atmospheres, *110*(D17). doi:10.1029/2004JD005619
18. FRIDELL, E., STEEN, E., & PETERSON, K. (2008). *Primary particles in ship emissions*. Atmospheric Environment, 42(6), 1160-1168. doi:10.1016/j.atmosenv.2007.10.042
19. GE, Z. (2016). *Supervised Latent Factor Analysis for Process Data Regression Modeling and Soft Sensor Application*. IEEE Transactions on Control Systems Technology, 24(3), 1004-1011. doi:10.1109/TCST.2015.2473817
20. GE, Z., & CHEN, X. (2016). *Supervised linear dynamic system model for quality related fault detection in dynamic processes*. Journal of Process Control, 44(Supplement C), 224-235. doi:10.1016/j.jprocont.2016.06.003
21. GHASSEMI, M., PIMENTEL, M. A. F., NAUMANN, T., BRENNAN, T., CLIFTON, D. A., SZOLOVITS, P., & FENG, M. (2015, 2015/1/25). *A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data*. Paper presented at the AAAI Conference on Artificial Intelligence.
22. GOLDSWORTHY, L., & GALBALLY, I. E. (2011). *Ship engine exhaust emissions in waters around Australia - an overview*. Air Quality and Climate Change, 45(4), 24-32.
23. HELFRE, J.-F., & BOOT, P. A. C. (2013). *Emission Reduction in the Shipping Industry:*

Regulations, Exposure and Solutions. Retrieved from http://www.sustainalytics.com/sites/default/files/shippingemissions_july2013.pdf
24. HICKMAN, J., HASSEL, D., JOUMARD, R., SAMARAS, Z., & SORENSON, S. (1999). *Methodology for calculating transport emissions and energy consumption*. Retrieved from
25. HULSKOTTE, J. H. J., & DENIER VAN DER GON, H. A. C. (2010). *Methodologies for estimating shipping emissions in Netherlands* (EMS).
26. JALKANEN, J. P., BRINK, A., KALLI, J., PETTERSSON, H., KUKKONEN, J., & STIPA, T. (2009). *A modelling system for the exhaust emissions of marine traffic and its application in the Baltic Sea area*. Atmospheric Chemistry and Physics, 9(23), 9209-9223. doi:10.5194/acp-9-9209-2009
27. LACK, D. A., CORBETT, J. J., ONASCH, T., LERNER, B., MASSOLI, P., QUINN, P. K., . WILLIAMS, E. (2009). *Particulate emissions from commercial shipping: Chemical, physical, and optical properties*. Journal of Geophysical Research: Atmospheres, 114(D7). doi:10.1029/2008JD011300
28. LI, Q., QIAO, F., & YU, L. (2016). *A Machine Learning Approach for Light-Duty Vehicle Idling Emission Estimation Based on Real Driving and Environmental Information*. Environment Pollution and Climate Change, 1(1). doi:10.4172/2573-458X.1000106
29. LINDSTAD, H., JULLUMSTRØ, E., & SANDAAS, I. (2013). *Reductions in cost and greenhouse gas emissions with new bulk ship designs enabled by the Panama Canal expansion*. Energy Policy, 59, 341-349. doi:10.1016/j.enpol.2013.03.046
30. MIOLA, A., & CIUFFO, B. (2011). *Estimating air emissions from ships: Meta-analysis of modelling approaches and available data sources*. Atmospheric Environment, 45(13), 2242-2251. doi:10.1016/j.atmosenv.2011.01.046
31. MOHD NOOR, C. W., MAMAT, R., NAJAFI, G., MAT YASIN, M. H., IHSAN, C. K., & NOOR, M. M. (2016). *Prediction of marine diesel engine performance by using artificial neural network model*. Journal of Mechanical Engineering and Sciences (JMES), 10(1), 1917-1930. doi:10.15282/jmes.10.1.2016.15.0183
32. MORENO-GUTIÉRREZ, J., CALDERAY, F., SABORIDO, N., BOILE, M., RODRÍGUEZ VALERO, R., & DURÁN-GRADOS, V. (2015). *Methodologies for estimating shipping emissions and energy consumption: A comparative analysis of current methods*. Energy, 86(Supplement C), 603-616. doi:10.1016/j.energy.2015.04.083
33. MUELLER, L., JAKOBI, G., CZECH, H., STENGEL, B., ORASCHE, J., ARTEAGA-SALAS, J. M., . ZIMMERMANN, R. (2015). *Characteristics and temporal evolution of particulate emissions from a ship diesel engine*. Applied Energy, 155(Supplement C), 204-217. doi:10.1016/j.apenergy.2015.05.115
34. OBODEH, O., & AJUWA, C. I. (2009). *Evaluation of artificial neural network performance in predicting diesel engine NOx emissions*. European Journal of Scientific Research, 33(4), 642-653.
35. OLESEN, H. R., WINTHER, M., ELLERMANN, T., PLEJDRUP, M., & CHRISTENSEN, J. (2009). *Ship emissions and air pollution in Denmark: Present situation and future scenarios*. Retrieved from
36. OLIVIER, J. G. J., & PETERS, J. A. H. W. (1999). *International marine and aviation bunker fuel: trends, ranking of countries and comparison with national CO2 emission*. RIVM Rapport 773301002.
37. PETERSEN, JÓAN, P., JACOBSEN, DANIEL, J., & WINTHER, O. (2011). *Statistical modelling for ship propulsion efficiency*. Journal of Marine Science and Technology, 17(1), 30-39. doi:10.1007/s00773-011-0151-0
38. RASSMUSSEN, C. E., & WILLIAMS, C. K. I. (2004). *Gaussian processes in machine learning*. Lecture notes in computer science, 3176, 63-71.
39. RODRIGUEZ-GALIANO, V., SANCHEZ-CASTILLO, M., CHICA-OLMO, M., & CHICA-RIVAS, M. (2015). *Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines*. Ore Geology Reviews, 71(Supplement C), 804-818. doi:10.1016/j.oregeorev.2015.01.001
40. SHANMUGAMA, P., SIVAKUMARB, V., MURUGESANA, A., & ILANGKUMARANA, M. (2011). *Performance and Exhaust Emissions of a Diesel Engine Using Hybrid Fuel with an Artificial Neural Network*. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 33(15), 1440-1450. doi:10.1080/15567036.2010.539085
41. SINGER, B. C., & HARLEY, R. A. (2012). *A fuel-based motor vehicle emission inventory*. Journal of the Air & Waste Management Association, 46(6): 581-593.
42. SMIT, R., NTZIACHRISTOS, L., & BOULTER, P. (2010). *Validation of road vehicle and traffic emission models - A review and meta-analysis*. Atmospheric Environment, 44(25), 2943-2953. doi:10.1016/j.atmosenv.2010.05.022
43. SMITH, T. W. P., JALKANEN, J. P., ANDERSON, B. A., CORBETT, J. J., FABER,

J., HANAYAMA, S., ETTING. (2014). *Third IMO GHG study*. In: International Maritime Organization.

44. CHU VAN, T., & BROWN, R. J. (2016). *Particle emissions from ships at berth using heavy fuel oil*. Paper presented at the Proceedings of the 17th International Association of Maritime Universities (IAMU) Annual General Assembly (AGA).
45. VANGHELUWE, M., MEES, J., & JANSSEN, C. (2007). *Monitoring programme on air pollution from sea-going vessels (MOPSEA)*. Retrieved from www.belspo.be
46. WANG, Y., & CHAIB-DRAA, B. (2017). *An online Bayesian filtering framework for Gaussian process regression: Application to global surface temperature analysis*. Expert Systems with Applications: An International Journal, 67(C), 285-295. doi:10.1016/j.eswa.2016.09.018
47. WOODS, D. C., OVERSTALL, A. M., ADAMOU, M., & WAITE, T. W. (2016). *Bayesian design of experiments for generalised linear models and dimensional analysis with industrial and scientific application*. Quality Engineering, 0-0. doi:10.1080/08982112.2016.1246045
48. YU, H. (2017). *Dynamic risk assessment of complex process operations based on a novel synthesis of soft-sensing and loss function*. Process Safety and Environmental Protection, 105(Supplement C), 1-11. doi:10.1016/j.psep.2016.10.006
49. YU, H., KHAN, F., & GARANIYA, V. (2016). *An alternative formulation of PCA for process monitoring using distance correlation*. Industrial & Engineering Chemistry Research, 55(3), 656-669. doi:10.1021/acs.iecr.5b03397