

**UNIVERSITÀ DEGLI STUDI DI NAPOLI
FEDERICO II**



**SCUOLA POLITECNICA E DELLE SCIENZE DI BASE
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE**

Dottorato di Ricerca in Ingegneria Industriale

**STATISTICAL METHODS FOR
INDUSTRIAL PROCESS MONITORING
BASED ON FUNCTIONAL DATA ANALYSIS**

Christian Capezza

TUTOR

Prof. Ing. Biagio Palumbo
Prof. Ing. Antonio Lepore

Coordinatore

Prof. Ing. Michele Grassi

XXXII CICLO

Summary

In this thesis, statistical methods for industrial process monitoring are proposed. The industrial scenario that motivates the research work is the monitoring and prediction of fuel consumption and CO₂ emissions from maritime transportation. The two main objectives are, on the one hand, the prediction of fuel consumption (and/or CO₂ emissions) on the basis of covariates describing the ship operating conditions at each voyage by means of advanced regression methods, and, on the other hand, the statistical process monitoring of the ship operating conditions and the fuel consumption (and/or CO₂ emissions) based on control charts.

The proposed methodologies can be arranged in three groups on the basis of how they treat the data for each voyage of a ship. The first group uses multivariate techniques, which, for each voyage, consider each individual observation of the variables as scalar quantities. Typically, the mean value of each variable over a voyage is considered. The second group considers the data for each voyage as profiles, from which several features are extracted in order to describe them in the best possible way. The third group considers the data for each voyage as functions, i.e. as complex, unique objects that have to be treated using functional data analysis techniques.

The common ground of all the proposed methodologies is the need to provide tools to industrial practitioners that are easily interpretable and give clear indications of anomalies by identifying the related causes, possibly in real-time, and the use of real-data examples to demonstrate their predictive and monitoring abilities.

Contents

Contents	ii
List of Figures	iv
List of Tables	ix
1 Introduction	3
1.1 The industrial scenario	4
1.2 Methods for the prediction of ship fuel consumption and CO ₂ emissions	5
1.3 Methods for statistical process monitoring of ship performance	7
1.4 Outline of the thesis	10
2 Advanced regression techniques for the prediction of ship CO₂ emissions	13
2.1 Introduction	13
2.2 Materials and methods	15
2.3 Comparison framework	19
2.4 Results and discussion	20
2.5 Conclusions	26
3 LS-PLS approach to ship fuel-speed curves	29
3.1 Introduction	29
3.2 The proposed approach	31
3.3 Case study: a Grimaldi Group Ro-Pax cruise ship	38
3.4 Conclusions	45
3.5 Appendix A: predictive ability comparison of PLS and MLR	47
3.6 Appendix B: main steps of the PLS algorithm	47
3.7 Appendix C: simulation study on the coverage of the approximate pointwise prediction intervals	48
4 Ship fuel consumption monitoring and fault detection via control charts based on partial least squares regression	51
4.1 Introduction	51
4.2 Data	53
4.3 The Statistical Approach and Monitoring Tools	59
4.4 Real-case study	62
4.5 Conclusion	67
5 Feature-oriented methods for the prediction of CO₂	69

5.1	Introduction	69
5.2	Materials and Methods	70
5.3	Predictive analytics comparison framework	74
5.4	Results and Discussion	75
5.5	Conclusions	80
6	Analysis of profiles using partial least-squares methods for ship performance monitoring	81
6.1	Introduction	81
6.2	The proposed approach	83
6.3	Real case study	89
6.4	Conclusion	97
7	Monitoring ship operating conditions and CO₂ emissions using control charts based on scalar-on-function regression	99
7.1	Introduction	99
7.2	A Motivating Example	101
7.3	Methodology	102
7.4	A Real-Case Study	108
7.5	Functional real-time control charts	118
7.6	Conclusion	122
8	The R package funcharts	125
8.1	Installation	125
8.2	Usage of funcharts	125
	Conclusion	141
	Bibliography	143

List of Figures

2.1	Correlation between predictors and CO ₂ emissions.	21
2.2	Analyzing the samples' space: (a) the scores from a PCA model reveal clusters in the data, which can be well identified by the sailed distance over ground (b). Analyzing the samples' space: (a) the scores from a PCA model reveal clusters in the data, which can be well identified by the sailed distance over ground (b).	22
2.3	Distribution of $RMSE_{dcv}$ obtained when predicting CO ₂ emissions for the short route.	23
2.4	Results from LASSO model training for the short route: (a) regression coefficients and (b) the predicted and observed CO ₂ emissions for all 40 iterations of double cross-validation.	24
2.5	Distribution of $RMSE_{dcv}$ obtained when predicting CO ₂ emissions for the medium route.	25
2.6	Results from LASSO model training for the medium route: (a) regression coefficients and (b) the predicted and observed CO ₂ emissions for all 40 iterations of double cross-validation.	25
2.7	Distribution of $RMSE_{dcv}$ obtained when predicting CO ₂ emissions for the long route.	26
2.8	Results from BT model training for the long route: (a) regression coefficients and (b) the predicted and observed CO ₂ emissions for all 40 iterations of double cross-validation.	27
3.1	Graphical scheme of the proposed approach for model building. Bold lines refer to measured variables/factors.	34
3.2	Graphical scheme of the proposed approach for monitoring FCPH. Bold lines refer to measured variables/factors.	36
3.3	The PRESS statistic used to determine the number of components to retain in the PLS model.	39
3.4	(a) FCPH and (b) normalized FCPH reference observations (points) used for the construction of (a) ordinary LS and (b) LS-PLS fuel-speed curves. Vertical scale unit is 1 t/h.	40
3.5	T^2 control chart for monitoring voyages with 99% control limit (dashed line).	41
3.6	SPE_X control chart for monitoring voyages with 99% control limit (dashed line).	42
3.7	Contribution to T^2 for VN 18, 19, 20, 45, and 46.	43
3.8	Contribution to SPE_X for VN 18, 19, 20 e 43.	44
3.9	Monitoring of voyages (triangles) through (a) ordinary LS fuel-speed and (b) LS-PLS fuel-speed curve. Vertical scale unit 1 t/h.	44

3.10	Benchmark comparison of 61 new voyages (triangles) after the EEI via (a) 95% LS prediction intervals and (b) 95% LS-PLS pointwise prediction intervals of the corresponding fuel-speed curve obtained based on 66 reference observations before the considered EEI. Vertical scale unit 1 t/h	45
3.11	Comparison between fuel-speed curves estimated based on reference observations collected before (thin line) and after (bold line) the EEI. Vertical scale unit 1 t/h	46
3.12	Fuel-speed curve (solid line) and 95% pointwise prediction interval of the normalized FCPH (dashes) estimated by (a) MLR and (b) PLS methods for voyages monitored in Phase II (triangles).	47
4.1	Wind components.	55
4.2	Engine room layout.	57
4.3	Ship 1 timeline.	59
4.4	Ship 2 timeline.	59
4.5	Fault detection analysis diagram.	60
4.6	Hotelling T^2 control chart for voyages 2034–2053 of Ship 1.	63
4.7	SPE_x control chart for voyages 2034–2053 of Ship 1.	63
4.8	Prediction error control chart for voyages 2034–2053 of Ship 1.	64
4.9	Contribution of the variables to the Hotelling T^2 statistic, for the voyage 2035 of Ship 1.	64
4.10	Contribution of the variables to the SPE_x statistic, for the voyage 2035 of Ship 1.	64
4.11	Contribution of the variables to the Hotelling T^2 statistic, for the voyage 2036 of Ship 1.	65
4.12	Contribution of the variables to the SPE_x statistic, for the voyage 2041 of Ship 1.	66
4.13	Prediction error control chart for voyages 427–476 of Ship 2, which follow an energy efficiency operation.	67
5.1	The application of a feature-oriented method transforms the input three-way structure \mathbf{X} into a feature matrix \mathbf{F} . AV, PdF and SPA stand for average values, profile-driven features, and statistical pattern analysis, respectively.	72
5.2	Distribution of $RMSE_{dcv}$ for 50 iterations of double cross-validation for the short route.	76
5.3	Distribution of $RMSE_{dcv}$ for 50 iterations of double cross-validation for the medium route.	77
5.4	Elastic net (EN) modeling results with PdF features for the medium route showing (a) the feature importance for different types of features, and (b) predicted vs measured CO_2 emissions.	78
5.5	Distribution of $RMSE_{dcv}$ for 50 iterations of double cross-validation for the long route.	79
5.6	LASSO modeling results with SPA features for the long rout showing (a) the feature importance for different types of features, and (b) the predicted and measured CO_2 emissions.	79
6.1	Hotelling T^2 control chart obtained with MPLS for monitoring voyages, with 95% and 99% upper control limits (dashed and solid line, respectively).	90
6.2	SPE_X control chart obtained with MPLS for monitoring voyages, with 95% and 99% upper control limits (dashed and solid line, respectively)	91

LIST OF FIGURES

6.3 Hotelling T^2 control chart obtained with tri-PLS for monitoring voyages, with 95% and 99% upper control limits (dashed and solid line, respectively). 91

6.4 SPE_X control chart obtained with tri-PLS for monitoring voyages, with 95% and 99% upper control limits (dashed and solid line, respectively.) 92

6.5 SPE_k control chart for one in control voyage 4 (a) and one out of control voyage 3 (b), with 95% and 99% upper control limits (dashed and solid line, respectively). 92

6.6 Contribution of the variables to the T^2 statistic (a) and to the SPE_X statistic for the voyage 3. 93

6.7 Functional observation of distance from mean route (a), SOG (b), and side wind (c) variables for voyage 3 against reference voyages. Note that in each plot the bold line refers to the voyage 3, while the grey lines are the observations of the corresponding variable for the 52 reference voyages. 94

6.8 SPE_k control chart for voyage 16, with 95% and 99% upper control limits (dashed and solid line, respectively). 95

6.9 Contribution of the variables to the T^2 statistic (a) and to the SPE_X statistic for the voyage 16. 95

6.10 Functional observation of SOG (a), power difference between port and starboard propeller shafts variables (b), and acceleration variables for voyage 16 against reference voyages. Note that in each plot the bold line refers to the voyage 16, while the grey lines are the observations of the corresponding variable for the 52 reference voyages. 96

6.11 Prediction error control chart. 97

7.1 Warping functions mapping at each voyage the cumulative sailing time to the common domain $\mathcal{T} = [0, 1]$, which is the fraction of distance travelled over the voyage. 110

7.2 (a) Fraction of the variance of the functional covariates explained by the multivariate functional principal components (solid line) and threshold (dashed line) equal to 0.01. (b) PRESS statistic calculated by Equation (7.8) as a function of the the first m retained principal components; trivially, when $m = 0$, the PRESS is obtained on the basis of the predictions calculated as the sample mean of the response variable; the points are the PRESS values corresponding to the multivariate functional principal components retained in the model. 111

7.3 (a) Hotelling T^2 , (b) SPE , and (c) response prediction error control charts used for monitoring the Phase II voyages. In each control chart, points joint by a line indicate monitoring statistic values at each voyage, while dashed lines indicate control chart limits as in Equation (7.23), with $\alpha = 0.05$ 112

7.4 Box plots of the contributions of the functional covariates to (a) the Hotelling T^2 statistic and (b) the SPE statistic for the reference voyages. 113

7.5 Contribution of the functional covariates to the SPE statistic for voyage 7. The bars are the contributions of the variables, with the darker ones indicating values exceeding the limit, while the black dashes are the limits calculated on the basis of the reference voyages. 113

7.6 Observations of the critical (i.e., indicated as responsible of OC by contributions plots in Figure 7.5) functional covariates for Phase II monitoring of the voyage 7, viz. (a) SOG (V), (b) cumulative sailing time (H), (c) acceleration (A), (d) power difference between port and starboard propeller shafts (ΔP), and (e) longitudinal wind component (W_L). In each plot, the black line indicates the current observation, while the reference functional observations are plotted in grey. 114

7.7 Contribution of the functional covariates to the Hotelling T^2 statistic (a) and SPE statistic (b) for voyage 18. The bars are the contributions of the variables, with the darker ones indicating values exceeding the limit, while the black dashes are the limits calculated on the basis of the reference voyages. 115

7.8 Observations of the critical (i.e., indicated as responsible of OC by contributions plots in Figure 7.7) functional covariates for Phase II monitoring of the voyage 18, viz. (a) SOG (V), (b) cumulative sailing time (H), (c) acceleration (A), (d) power difference between port and starboard propeller shafts (ΔP), and (e) longitudinal wind component (W_L). In each plot, the black line indicates the current observation, while the reference functional observations are plotted in grey. 116

7.9 Contribution of the functional covariates to the Hotelling T^2 statistic (a) and SPE statistic (b) for voyage 19. The bars are the contributions of the variables, with the darker ones indicating values exceeding the limit, while the black dashes are the limits calculated on the basis of the reference voyages. 117

7.10 Observations of the critical (i.e., indicated as responsible of OC by contributions plots in Figure 7.9) functional covariates for Phase II monitoring of the voyage 19, viz. (a) SOG, (b) cumulative sailing time (H), (c) acceleration, (d) power difference between port and starboard propeller shafts (ΔP), (e) longitudinal wind (W_L), and (f) transverse wind (W_T). In each plot, the black line indicates the current observation, while the reference functional observations are plotted in grey. 118

7.11 Contribution of the functional covariates to the Hotelling T^2 statistic (a) and SPE statistic (b) for voyage 24. The bars are the contributions of the variables, with the darker ones indicating values exceeding the limit, while the black dashes are the limits calculated on the basis of the reference voyages. 119

7.12 Observations of the critical (i.e., indicated as responsible of OC by contributions plots in Figure 7.11) functional covariates for Phase II monitoring of the voyage 24, viz. (a) SOG (V), (b) power difference between port and starboard propeller shafts (ΔP), (c) longitudinal wind component (W_L), and (d) acceleration (A). In each plot, the black line indicates the current observation, while the reference functional observations are plotted in grey. 120

7.13 (a) Scalar control chart for the covariates based on the Hotelling statistic (7.30) and (b) scalar control chart for the response prediction error used for monitoring the Phase II voyages. In each control chart, points joint by a line indicate monitoring statistic values at each voyage, while solid lines indicate control chart limits. 121

7.14 Graphical example showing how $k^* = d^*/d$ is determined for a new voyage. The dashed curve represents the route travelled by the ship up to the current GPS position, which is the point labeled as P^* . The solid curve represents the nominal route and its point nearest to P^* is labeled as \bar{P}^* . d^* is the length of the portion of the solid curve from the departure port to \bar{P}^* . d is the total length of the nominal route. 122

LIST OF FIGURES

7.15 Real-time monitoring of voyages 7 (a), 18 (b), 19 (c), 23 (d) and 24 (e). For each voyage, the T^2 , SPE , and response prediction error control charts are plotted. In each plot, solid lines indicate the profiles of the monitoring statistics during the voyage, and dashed lines indicate the corresponding control limits as in Equation (7.23), with $\alpha = 0.05$ 123

List of Tables

2.1	Operational variables measured for each voyage.	15
2.2	Hyper-parameters tested during model training. The suitable value for each method is selected by 10-fold cross-validation.	20
3.1	ANOVA table for hypothesis testing of the effectiveness of an EEI.	38
3.2	Operational variables measured for each voyage.	39
3.3	ANOVA table for hypothesis testing of the effectiveness of an EEI.	45
3.4	Mean, median, and 95% confidence interval of the approximate prediction interval coverage.	49
4.1	Physical variables acquired at each voyage considered in the proposed approach.	54
5.1	Variables used as covariates in the feature-oriented approach.	72
5.2	Object-profiles and their respective features considered in the PdF dictionary.	73
7.1	Functional covariates used in the scalar-on-function regression model.	109

Acknowledgements

I am grateful to my PhD supervisors Biagio Palumbo and Antonio Lepore for giving me the opportunity to invest my last three years on an intense and productive activity, which made me grow as a researcher, and for their continuous support. Thanks to Biagio for being able to manage the entire activity of the research group and to meet different needs of many people, and for all his suggestions. Thanks to Antonio for pushing me and all our research group to aim always at the excellence. I am also grateful to my colleague Fabio Centofanti, for sharing so many ideas, for helping me to focus on the important things first, and because we often motivated each other to do our best. Thanks also to Chiara, Francesco, and Maria Sole for being great office mates and creating a better working atmosphere.

Thanks to the professors of the Department of Statistical Sciences of the University of Padova for the incredible value they added to my education in Statistics thanks to the PhD courses I followed during the academic year 2017/2018. I am also grateful to all the PhD students of the XXXIII cycle, which were my colleagues during the courses. I am especially grateful to Michele Lambardi di San Miniato for sharing with me his great knowledge about Mathematics, Statistics, and many other things.

Thanks to Matteo Fasiolo and Simon N. Wood for giving me the opportunity to spend seven months as visiting postgraduate research student at the School of Mathematics of the University of Bristol, where we have collaborated on a research project. I am grateful to Simon for his suggestions and ideas which helped to improve our research work. I am especially grateful to Matteo for all the time we spent together in and outside the university, for all the sport activity we did together, for introducing me to squash, for all the things I learned about statistics from both computational and methodological point of views, about mgcv and generalized additive models.

Thanks to my girlfriend Mara for supporting me all the time, especially when we were far away from each other, while I was in Padova and Bristol. *Infine, grazie nuovamente a Mara, ma soprattutto grazie a mamma, papà, Salvio, alla mia famiglia e ai miei amici per essere sempre disponibili ad aiutarmi a ritrovare fiducia in me stesso qualora ne avessi bisogno, e senza i quali non sarei mai arrivato fin qui oggi.*

Chapter 1

Introduction

The aim of this thesis is to provide statistical methods in the industrial context of prediction and monitoring of fuel consumption and CO₂ emissions from maritime transportation. The two main objectives are the prediction of fuel consumption (and/or CO₂ emissions) on the basis of covariates describing the ship operating conditions at each voyage by means of advanced regression methods, and the statistical process monitoring of ship operating conditions and fuel consumption (and/or CO₂ emissions) based on control charts.

The relevance of the industrial scenario is highlighted by the new international regulations, such as the EU regulation 2015/757, which urge shipping operators to set up systems for the monitoring, reporting and verification of CO₂ emissions. These regulations in the shipping sector aim to give greater transparency to operations and public access to CO₂ emissions data. On the other hand, the continuous acquisition of operational data, which is performed on most of the modern ships, urgently calls for the application of new and opportune statistical methods able to deal with high-dimensional data. Modern multi-sensor systems are able to stream massive amounts of high-frequency observational data, which can be considered to be varying over a continuous domain, therefore ship operating conditions can be described by sensor signals collected throughout each voyage and stored as profiles. However, in today's market, there is no standard solution or method available that can be robustly adopted in real environments for the shipping industry.

The methodologies proposed in this thesis are shown in the next chapters and can be arranged in three groups on the basis of how they treat the data for each voyage of a ship. The first group of methodologies uses multivariate techniques, that is, for each voyage, observations of the variables are scalar quantities. The second group of methodologies considers the data for each voyage as profiles, from which several features are extracted in order to describe them in the best possible way. The third group of methodologies considers the data for each voyage as functions, i.e. as complex, unique, objects, which are to be treated using functional data analysis techniques.

In order to give an overview of the entire research work, we briefly introduce the industrial scenario motivating the research work in Section 1.1, then we introduce methods for the prediction of ship fuel consumption and CO₂ emissions in Section 1.2 and methods for statistical process monitoring of ship performance in Section 1.3.

1.1 The industrial scenario

In the last years, the problem of monitoring CO₂ emissions in the maritime transportation field has become of paramount importance in view of the climate change and global warming issues. The shipping industry is nowadays facing a new regulatory regime that aims to give public access to emissions data. At the European level, the application of the EU regulation 2015/757 (European Commission 2015), which is mandatory from January 2018, urges shipping companies to set up a system for daily monitoring, reporting, and verification (MRV) of emissions for each ship. At the world level, environmental directives imposed by the Kyoto Protocol and the International Maritime Organization (IMO) are coming into force to strictly control greenhouse gases emissions, especially into the so-called emission controlled areas (Buhaug et al. 2009, IMO 2012*a,b,c,d*, European Commission 2013). Then, the extensive and increasingly demanding air pollution programs make predicting ships' CO₂ emissions not only a strategic, but also a mandatory task for shipping companies.

Concurrently, automatic multi-sensor acquisition systems installed on-board of modern ships allow data to be uploaded and stored without manual intervention and in very short periods of time. Thus, the initial problem of acquiring and transferring data to a remote server (Chen 1989) now turns into a problem of correctly handling and processing the information hidden in the massive amount of operational data (Bertram 2011, Løvoll & Kadal 2014) available (e.g. vessel's resistance, power needed for propulsion, fuel consumption, speed, different routes, weather conditions, sea waves, displacement, draughts, trim, engine operation mode, etc.) and converting it into value. The profile of these variables is typically complex, unstructured, intrinsically collinear and with non-stationary behaviour.

The critical issue of predicting ship fuel consumption and CO₂ emissions in the shipping industry is a starting point to devise better operational strategies to mitigate this environmental aspect and decrease the ecological footprint of the shipping activity. The most common method used in the naval literature to estimate the fuel consumption concerns the use of the so-called power-speed curves, because a direct proportional relationship also holds between the engine power and the fuel consumption through the power-based specific fuel oil consumption coefficient (see e.g. Corbett & Koehler (2003)). Methods for estimating the relationship between operational data and some response (such as fuel consumption or CO₂ emissions) are usually classified as *white-box*, when based only on principles of physics and *black-box*, when the relationship between response and predictor variables is purely data-driven. However, these two extreme categories are not so well distinct and *grey-box* is the most common label for methods that start from physics principles but also learn from the observed data. The marine engineering literature mainly relies on the use of white-box models. The most common example to estimate the fuel consumption and CO₂ emissions is the use of the so-called power-speed curves (Van Manen et al. 1988, Schrady et al. 1996), which describe an ideal univariate relationship between vessel speed and the engine power. In fact, a direct proportional relationship also holds between the engine power and the fuel consumption through the power-based specific fuel oil consumption coefficient (see e.g. Corbett & Koehler (2003)). However, these curves are usually calibrated through dedicated tests and overlook other factors affecting the vessel during navigation, then this leads to poor predictions of fuel consumption and harmful emissions. In this context and in the recent years, a few black- and grey-box approaches have been proposed to circumvent these limitations. Perera & Mo (2016) drew empirical relationships between ship resistance and speed through data visualization methods. Petersen et al. (2012) investigated artificial neural networks and Gaussian Process approaches for statistical modeling of fuel efficiency. Lu et al.

(2015) developed a semi-empirical ship operational performance predictive model to estimate the ship's added resistance considering specific additional variables. Bocchetti et al. (2015) proposed a statistical approach founded on multiple linear regression which allows for both point-wise and interval predictions of the fuel consumption at given operating conditions.

Classical approaches used in the naval literature may fail or are at least suboptimal, since they are often limited to the analysis of averages per voyage. However, in spite of the easier interpretability, compressing a variable profile into a single average value may lead to significant information loss and to discarding most of the relevant dynamic patterns. For these reasons, in this thesis we propose the methodologies briefly summarized in the following sections.

1.2 Methods for the prediction of ship fuel consumption and CO₂ emissions

Nowadays, the statistical and data science domains offer potentially interesting tools. Statisticians, engineers and data scientists are naturally called upon in order to propose and test alternative predictive techniques that might answer positively to the issues referred above, especially when applied to current complex data, which is typically less structured, larger in size, and requires merging several datasets from various sources.

We want to clarify that in this thesis, the term “prediction” is not used in the sense of “forecasting”, i.e. being able to predict the amount of CO₂ emissions before the end of a voyage, or of voyages following the current one. The main aim with prediction is to “model the variability” of the scalar response variable conditional on all factors that can influence it. This is useful because, at the end of a new voyage to be monitored, for example, the practitioner is able to understand that the amount of CO₂ emissions was higher than average because of adverse weather conditions and then its variability was not unusual given the model, while, on the other hand, in some other situation the amount of CO₂ emissions was close to the average, but its prediction error was too large, then this situation should be signaled as potentially anomalous. It is also important to note that there are both controllable and uncontrollable covariates, for example the speed of a ship is a controllable factor, while the wind speed is uncontrollable. Using both controllable and uncontrollable factors is crucial, because they allow to model the variability of the scalar response better.

The next subsections illustrate the methodologies proposed in this thesis for prediction of ship fuel consumption and CO₂ emissions, by following the increasing complexity of the data, i.e. we consider multivariate methods, feature-oriented methods, and scalar-on-function regression.

Multivariate methods

In Chapter 2 multivariate methods for the prediction of ship fuel consumption and CO₂ emissions are proposed. A first aim of this thesis is in fact to compare and discuss modern regression techniques and their ability for predicting ships' CO₂ emissions based on the navigation information available in modern ships and to develop a robust and accurate model for this critical aspect. A large number of regression methods are available in the literature and can be used in the present context of predicting CO₂ emissions. Chapter 2 proposes four different classes of regression methods representing a rich variety of modeling approaches usually considered for handling large datasets possibly presenting collinearity and sparsity issues: variable selection methods, penalized regression methods, latent variable methods

and tree-based ensemble methods. Each class contains methods that share similarities in terms of prior assumptions regarding the predictor variables, the response variable (ship fuel consumption or CO₂ emissions), and the relationship between them.

Variable selection methods. The variable selection class assumes that only some predictors are relevant for the prediction model, i.e. some predictor variables carry relevant information regarding the response while others are irrelevant and should be discarded in order to obtain more simple, robust and parsimonious models.

Penalized regression methods. The class of penalized regression methods imposes a penalty on the magnitude of the regression coefficients, thus constraining their values to be small. The penalty increases the bias of the regression model, but stabilizes the estimator variance so that better estimates of the regression coefficients can be obtained. Note that the variance decrease often compensates the bias increase so that the overall prediction performance can be improved. This includes four regression methods: ridge regression (Draper & Smith 2014), least absolute shrinkage and selection operator (Tibshirani 1996, Rasmussen & Bro 2012) (LASSO), elastic net and support vector regression (Smola & Schölkopf 2004, Ahmed et al. 2010, Canu et al. 2005).

Latent variable methods. The third class of methods is based on the latent variable framework (Burnham et al. 1999) where a set of unmeasured quantities are considered to be responsible for the observed variability in both predictors and response variable. Linear combinations of the measured variables are used to estimate the latent variables and the model obtained can be used for predicting the response based on the available predictors. This class includes three regression methods: principal component regression (Jolliffe 2002, Jackson 2005, Wold et al. 1987), principal component regression with the scores added in a forward stepwise fashion (PCR_{FS}) and partial least squares regression (Wold et al. 2001, Geladi & Kowalski 1986, Wold et al. 1984).

Tree-based ensemble methods. The last class of regression methods is based on ensembles of regression trees (Hastie et al. 2009, Dietterich 2000, Strobl et al. 2009, Breiman et al. 1984). A regression tree is a model that splits the predictors' space into regions where the response variability is small. Then, the mean response of the samples within the leaf is the predicted value. Ensemble methods are adopted to decrease the variance of the trees, which is indeed very high when trees are deep and therefore could overfit the training sets. This class includes three regression methods: bagging of regression trees (BRT), random forests (RF) and boosting of regression trees (BT). BRT applies bootstrap to the original dataset in order to generate many datasets. Each one of these datasets is used to build a regression tree and the average prediction from all trees in the ensemble constitute the predicted response. Similarly, RF uses bootstrap to generate datasets, which however only contain a smaller number (e.g. \sqrt{p}) of randomly selected predictors. Lastly, BT (Elith et al. 2008, Cao et al. 2010, Freund et al. 1996) iteratively fits regression trees to the residuals from previous trees so that each tree focuses on samples that were poorly modelled.

Feature-oriented methods

In Chapter 5, feature-oriented methods for the prediction of ship fuel consumption and CO₂ emissions are proposed. Automatic multi-sensor acquisition systems installed on-board of modern ships facilitate collecting operational data (i.e., ship operating conditions) from a massive number of variables. The profile of these variables is typically complex, unstructured, intrinsically collinear and with non-stationary behavior. In this scenario, classical approaches used in the naval literature may fail or are at least suboptimal, since they are limited to the

analysis of averages per voyage. However, in spite of the easier interpretability, compressing a variable profile into a single average value may lead to significant information loss and to discarding most of the relevant dynamic patterns.

In the opposite spectrum of complexity, multivariate statistical methods commonly used for monitoring batch processes usually require the implementation of data pre-processing techniques that constitute an additional challenge for practitioners and may hamper their practical usability. For example, data needs to be correctly unfolded to handle its three-way structure, resulting in a very large number of pseudo-variables and model parameters. Furthermore, complex synchronization methods are required in order to ensure that the voyages' major landmarks are aligned and that all voyages have the same number of observations.

The batch process monitoring literature is vast and another class of approaches that is growing in importance is the class of feature-oriented methods. These methods are simpler to apply because they do not require synchronization and tend to be more parsimonious since the number of model parameters is smaller. Examples of feature-oriented methods include profile-driven features, recently proposed by Rendall et al. (2017), and statistical pattern analysis, proposed by He & Wang (2011). These techniques compress each variable into a small number of features that can be utilized for data-driven model building.

Scalar-on-function regression method

In Chapter 7, functional data analysis methods for the prediction of ship fuel consumption and CO₂ emissions are proposed. The most complex approach to the prediction problem avoids the simplification of feature-oriented methods that extract relevant features from the profiles, but it treats the data describing the operating conditions of a ship over each voyage as unique, complex mathematical objects, such as functions or vectors of functions. In this context, functional data analysis techniques, in particular scalar-on-function regression, can be used to predict the fuel consumption or CO₂ emissions at each voyage.

Denote by $\mathbf{X} = (X_1, \dots, X_P)$ a vector of random elements that take value in the space of square integrable functions, and with y the scalar response variable representing the fuel consumption or the CO₂ emissions at each voyage, let $\{(\mathbf{X}_i, y_i)\}_{i=1, \dots, n}$ be a random sample from (\mathbf{X}, y) , with $\mathbf{X}_i = \mathbf{X}_i(t) = (X_{i1}(t), \dots, X_{iP}(t))$ being a vector of functional covariates. The conditional distribution of y_i given the corresponding observation of the functional covariates \mathbf{X}_i can be modeled by means of the following scalar-on-function regression model

$$y_i = \beta_0 + \sum_{p=1}^P \int_{\mathcal{T}} X_{ip}(t) \beta_p(t) dt + \varepsilon_i, \quad i = 1, \dots, n, \quad (1.1)$$

where $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P)$, with $\beta_p \in L^2(\mathcal{T})$, the space of square integrable functions, are the coefficient to be estimated, and $\varepsilon_1, \dots, \varepsilon_n$ are the error terms, which are assumed to be independent identically distributed. Estimation of these types of models requires to deal with the infinite dimensionality of the dataset.

1.3 Methods for statistical process monitoring of ship performance

First of all, we want to clarify the meaning of the word “monitoring” in this thesis work. According to Woodall (2017), because it better reflects the application of the methods, the use of control charts and other monitoring methods should be referred to as “statistical

process monitoring”, not “statistical process control.” The use of the word “control” implies that control actions based on an adjustment variable are part of the practice and theoretical framework in the sense of Box & Narasimhan (2010), however this is not the case in this thesis.

The statistical process monitoring literature makes a distinction between phase I and phase II monitoring (Vining 2009). Phase I control charts use a “base period,” which is the period used to estimate the control chart parameters. Phase I control charts are much more an exploratory data analytic tool. Phase II control charts treat the control chart parameters as known and in actual practice these are estimated based on the results from the phase I study. One legitimately may view a phase II control chart as a sequence of hypothesis tests. In this thesis, only phase II control charts are proposed. While there is always a phase I based on both statistical and engineering considerations to remove outliers that can affect estimation of model parameters and control chart limits, this part is not focus of this thesis and we use phase I only to estimate model parameters. The objective is to identify new voyages as anomalous after estimation of model parameters, then we perform only phase II monitoring. Therefore, in each chapter it is always assumed that the data used for model estimation have already been filtered and do not contain outliers.

There are two main aims related to statistical process monitoring of ship performance and fuel consumption and/or CO₂ emissions: (i) statistical monitoring of ship fuel consumption and CO₂ emissions to support shipping management to identify anomalies and (ii) quantifying fuel consumption reduction consequent to energy efficiency initiatives or dry-dock operations. In the naval literature, the most common method used to estimate fuel consumption and then CO₂ emissions is the so-called speed-power curve. These curves are used both for prediction and monitoring of any possible anomalies in the ship performance. However, despite its intuitive usage, this method is affected by large variability due to different sailing (e.g. trim, displacement, etc.) and weather conditions. Many of the available methods for prediction and monitoring have strong limitations when applied to high-dimensional and correlated data, or they do not fully exploit all of the available information.

There is a common framework among all the methods proposed in this thesis for statistical process monitoring, which can be summarized in the following steps:

1. the starting point is a regression model where the ship fuel consumption or the CO₂ emissions are the response variable to be predicted using one of the approaches described in the previous section;
2. since the predictor variable space is characterized by high or infinite dimensionality, dimensionality reduction is applied to both stabilize the estimation of regression coefficients in the prediction problem and to describe covariates in a more efficient and interpretable way, in a lower dimensional subspace; usually principal component analysis (possibly functional) and partial-least squares methods are used for this purpose;
3. the dimension reduction provides a nice split of the covariate space into two complementary subspaces: correspondingly, two control charts are used. Usually the Hotelling T^2 statistic is calculated on the variables obtained from the projection of covariates onto the subspace of the components retained in the model, while a squared prediction error statistic monitors the squared distance of the covariates in the original space from the projection subspace; finally a third monitoring statistic looks at the prediction error on the response variable based on the regression model;

4. on the basis of the monitoring statistics and a reference sample of in-control observations, control charts are built to monitor future observations;
5. once control charts are built (again, in all the research work proposed in this thesis, phase I focuses only on estimation of model parameters), we can use them to detect if a new observation is anomalous, in phase II;
6. when some control charts detect out-of-control observations, contribution plots are built in order to decompose the monitoring statistics as sums over the covariates and then to identify the variable(s) responsible of anomalies.

The next subsections illustrate the methodologies proposed in this thesis for monitoring ship performance, by following the increasing complexity of the data, i.e. we consider multivariate methods and functional data analysis methods. We want to specify that most of the thesis refers to a retrospective monitoring, which means that points can be plotted on control charts for each monitoring statistic only at the end of a ship voyage, even though variables for each voyage are stored as profiles. Then, the monitoring is performed at the end of each voyage by looking backwards at the data off-line. The real-time monitoring is indeed of primary interest and is in fact the newest part of this thesis work. It is presented in Chapter 7. However, we think that off-line monitoring is still interesting to give valuable information to the maritime engineers between consecutive voyages, moreover when profiles are analyzed it is important to have indication about which part of a voyage was potentially anomalous.

Multivariate methods

In Chapters 3 and 4, multivariate methods for monitoring ship performance are proposed. Chapter 4 uses partial-least squares (PLS) regression that is very useful because it is able to give good predictions of the response variable and also allows multivariate statistical process monitoring. The choice of the PLS in place of e.g., multiple linear regression has great potential of supporting the management to handle the great amount of data collected on board of modern ships that are usually noisy and strongly correlated. The residuals left by the PLS model are also naturally prone to be monitored at each new voyage through prediction error control chart, whereas the predictor variables can be monitored through the Hotelling's T^2 and squared prediction error control charts. When a point falls outside the upper control limit of at least one of the control charts, a possible problem may have occurred. The management is then urged to further investigate physical variables that have caused the out-of-control condition by exploring the corresponding contribution plot.

Another novel method proposed in this thesis is developed in Chapter 3 with the aim of enhancing fuel-speed curves accuracy when information is available on additional factors from multi-sensor systems. This is done by elaborating the orthogonal least-squares partial least-squares (LS-PLS) approach (Jørgensen et al. 2004, 2007). This method avoids the collinearity issue between additional factors (Montgomery et al. 2012), which usually bounds the number of additional factors to be included in the regression models. The LS-PLS estimation method makes interpretation easier with respect to the standard partial least-squares (PLS). The orthogonal LS-PLS algorithm is in fact used to estimate parameters of the model and to maintain the physical relation between fuel consumption (or CO₂ emissions) per hour and speed over ground expected by marine engineers that are used to classical fuel-speed curves. Besides, the LS-PLS method is expanded also for constructing opportune control charts for monitoring the predictor variables as well as the response variable.

Functional data analysis methods

As for the prediction problem case, while multivariate methods described above work with summary statistics (typically averages) for each voyage to perform statistical process monitoring, ship operating conditions can be described by sensor signals collected throughout each voyage and stored as profiles. The latter can be analyzed through appropriate techniques.

The main advantage of considering profiles instead of single observations per each voyage is the possibility to give real-time predictions and indications on possible anomalies in ship operating conditions during a voyage and on the instant at which anomalies may have occurred.

Two main approaches are presented in this thesis to profile monitoring. Both of them are able to monitor profiles with different lengths at different voyages (due to the different duration).

- The first one is presented in Chapter 6 and uses multi-way partial least-squares regression Nomikos & MacGregor (1995a) and a multilinear version of PLS proposed by Bro (1996), Smilde (1997), which is called three-way partial least-squares (tri-PLS), which requires a three-dimensional array that contains ship operational reference profiles at given domain points with the following three dimensions: the number of replications, the number of variables, and the number of evaluation points.
- The second one is presented in Chapter 7 and is based on functional data analysis, in particular on multivariate functional principal component analysis.

1.4 Outline of the thesis

Next chapters are based on papers that have been published in refereed journals. In particular, Chapter 2 is based on the paper:

- Lepore, A., Reis, M. S. d., Palumbo, B., Rendall, R. & Capezza, C. (2017), ‘A comparison of advanced regression techniques for predicting ship CO₂ emissions’, *Quality and Reliability Engineering International* **33**(6), 1281–1292.

Chapter 3 is based on the paper:

- Lepore, A., Capezza, C., & Palumbo, B. (2019), ‘Orthogonal LS-PLS approach to ship fuel-speed curves for supporting decisions based on operational data’, *Quality Engineering* **31**(3), 386–400.

Chapter 4 is based on the paper:

- Capezza, C., Coleman, S., Lepore, A., Palumbo, B. & Vitiello, L. (2019), ‘Ship fuel consumption monitoring and fault detection via partial least squares and control charts of navigation data’, *Transportation Research Part D: Transport and Environment* **67**, 375–387.

Chapter 5 is based on the paper:

- Reis, M. S. d., Rendall, R., Palumbo, B., Lepore, A. & Capezza, C. (2019), ‘Predicting ships’ CO₂ emissions using feature-oriented methods’, *Applied Stochastic Models in Business and Industry* **36**(1), 110–123.

Chapter 6 is based on the paper:

- Lepore, A., Capezza, C., & Palumbo, B. (2018), ‘Analysis of profiles for monitoring of modern ship performance via partial least squares methods’. *Quality and Reliability Engineering International* **34**(7), 1424–1436.

Chapter 7 is based on the paper:

- Capezza, C., Lepore, A., Menafoglio, A., Palumbo, B. & Vantini, S. (2020+), ‘Control charts for monitoring ship operating conditions and CO₂ emissions based on scalar-on-function regression’. To appear in *Applied Stochastic Models in Business and Industry*.

Finally, Chapter 8 introduces the R package `funcharts`, which provides the software for the methodology proposed in Chapter 7 and is available on GitHub at the link <https://github.com/unina-sfere/funcharts>.

In particular, Chapters 2, 4, and 5 are based on papers that are output of a collaboration agreement between The Industrial Statistics Research Unit, Newcastle University, United Kingdom, The Department of Industrial Engineering, University of Naples Federico II, Italy, The Process Systems Engineering Group of the Chemical Process Engineering and Forest Products, and Research Center (CIEPQPF), Department of Chemical Engineering, University of Coimbra, Portugal, whose title is “Big data in shipping for monitoring and benchmarking fuel efficiency and CO₂ emissions”, which has been signed with reference to years 2016, 2017, and 2018, and which has been extended to years 2019, 2020, and 2021.

Chapter 2

Advanced regression techniques for the prediction of ship CO₂ emissions

The new European Union Regulations urge shipping operators to set up systems for the monitoring, reporting and verification of CO₂ emissions. Indeed, new monitoring data acquisition systems installed on modern ships have brought a navigation data overload that needs to be correctly handled in order to make proper decisions about their operation. However, in today's market, there is no standard solution or method available that can be robustly adopted in real environments for the shipping industry. In view of the novel attempts for solving this issue proposed by statisticians, marine engineers and practitioners, this chapter presents an extensive comparison of several regression techniques that can exploit the navigation information usually available in modern ships: variable selection methods, penalized regression methods, latent variable methods and tree-based ensemble methods. The comparison is made by means of operational data collected on a Ro-Pax cruise ship owned by the Italian shipping company Grimaldi Group. The goal of this analysis is twofold: (i) to identify methodologies with more potential at analyzing the data collected from this shipping industry scenario; (ii) to develop a predictive model for CO₂ emissions with good characteristics of accuracy and robustness.

2.1 Introduction

The shipping industry is currently facing a new regulatory regime as well as new relevant market challenges. Environmental directives imposed by the Kyoto Protocol and the International Maritime Organization (IMO) are coming into force to strictly control greenhouse gases emissions, especially into the so-called emission controlled areas (Buhaug et al. 2009, IMO 2012*a,b,c,d*, European Commission 2013). Concurrently, sensor technology and automatic acquisition systems allow data to be uploaded and stored without manual intervention and in very short periods of time. Thus, the initial problem of acquiring and transferring data to a remote server (Chen 1989) now turns into a problem of correctly handling and processing the information hidden in the massive amount of operational data (Bertram 2011, Løvoll & Kadal 2014) available (e.g. vessel's resistance, power needed for propulsion, fuel consumption, speed, different routes, weather conditions, sea waves, displacement, draughts, trim, engine operation mode) and converting it into value.

This article addresses the critical issue of predicting CO₂ emissions in the shipping industry,

as a starting point to devise better operational strategies to mitigate this environmental aspect and decrease the ecological footprint of this activity. In this context, methods for estimating the relationship between operational data and some response (such as CO₂ emissions) are usually classified as *white-box*, when based only on principles of physics and *black-box*, when the relationship between response and predictor variables is purely data-driven. However, these two extreme categories are not so well distinct and *grey-box* is the most common label for methods that start from physics principles but also learn from the observed data. The marine engineering literature mainly relies on the use of white-box models. The most common example in this sense is represented by the speed-power curves (Van Manen et al. 1988, Schrady et al. 1996) which describe an ideal univariate relationship between the engine power and the vessel speed and are usually calibrated through dedicated tests. However, these approaches overlook other factors affecting the vessel during navigation and lead to poor predictions of fuel consumption and harmful emissions. In this context and in the recent years, a few black- and grey-box approaches have been proposed to circumvent these limitations. As an example of black-box approach, Perera & Mo (2016) drew empirical relationships between ship resistance and speed through data visualization methods in order to identify energy efficient operation conditions. A grey-box methodology was proposed by Bialystocki & Konovessis (2016) for a more accurate estimation of ship's speed-fuel curve on the basis of the major operational variables affecting it, namely ship's draught and displacement, weather force and direction, hull and propeller roughness. Petersen et al. (2012) investigated artificial neural networks and Gaussian Processes approaches for statistical modelling of fuel efficiency. Lu et al. (2013, 2015) also developed a semi-empirical ship operational performance predictive model to estimate the ship's added resistance considering specific operational variables, e.g. ship type, draughts, speeds, encounter angles, sea states, fouling effect and engine degradation conditions. Unfortunately, most of these methods are based only on physical deterministic relationships and/or classical univariate statistical tools and therefore, have strong limitations when applied to high-dimensional and correlated data.

The statistical and data science domains, in turn, offer potentially interesting alternatives to the methodologies mentioned above. A variety of multivariable methods satisfactorily solves the issue of simultaneously integrating the contribution of operational variables and their interactions. However, they may present sensitivity to outliers and cross-correlations (both in the variable and observation modes) as well as be affected by the well-known collinearity issue between predictors (Montgomery et al. 2012), which limits the accuracy of the parameters included in the models.

In this context, statisticians, engineers and data scientists are naturally called upon in order to propose and test alternative predictive techniques that might answer positively to the issues referred above, especially when applied to current complex data, which is typically less structured, larger in size, and requires merging several datasets from various sources. The aim of this chapter is to compare and discuss modern regression techniques and their ability for predicting ships' CO₂ emissions based on the navigation information available in modern ships and to develop a robust and accurate model for this critical aspect. The comparison is made by means of operational data collected on a Ro-Pax cruise ship owned by the Italian shipping company *Grimaldi Group*.

Table 2.1: Operational variables measured for each voyage.

	Variable	Description
1	SG_P	Shaft generator power (port) [kW]
2	SG_S	Shaft generator power (starboard) [kW]
3	ΔP	Power difference between port and starboard propeller shafts [kW]
4	ΔSG	Power difference between two shaft generators [kW]
5	h	Actual voyage navigation time
6	V	Speed Over Ground (SOG) [kn]
7	W_F	Following wind [kn]
8	W_H	Head wind [kn]
9	W_S	Side wind [kn]
10	T_{FD}	Departure draught (fore perpendicular) [m]
11	T_{AD}	Departure draught (aft perpendicular) [m]
12	T_{PD}	Departure draught (midship section—port) [m]
13	T_{SD}	Departure draught (midship section—starboard) [m]
14	T_{FA}	Arrival draught (fore perpendicular) [m]
15	T_{AA}	Arrival draught (aft perpendicular) [m]
16	T_{PA}	Arrival draught (midship section—port) [m]
17	T_{SA}	Arrival draught (midship section—starboard) [m]
18	σ_V^2	SOG variance [kn ²]
19	$Trim_D$	Departure Trim [m]
20	$Trim_A$	Arrival Trim [m]
21	Δ	Displacement [m]
22	M	Sailed Distance Over Ground [NM]

2.2 Materials and methods

Data description

Operational data from a Ro-Pax cruise ship were collected during a period of one year and stored for analysis in the current shipping industry scenario. The ship links four Mediterranean ports. However, their names and voyage dates are intentionally omitted for confidentiality reasons. The total CO₂ emissions at each voyage, which is the response variable, is calculated depending on fuel consumption through the fuel mass to CO₂ emission factor for heavy fuel oil (IMO 2009).

Table 2.1 describes the predictor variables used to describe the ship operating conditions and to build regression models. Further details can also be found in Bocchetti et al. (2015).

The observations for the predictor variable (PV) are obtained for each voyage as summary statistics of high-rate sensor measurements; they are acquired without human intervention by a large-scale network system patented by CETENA SpA. All PVs reported in Table 2.1 refer to the *actual voyage navigation time*, which is defined as the time between the “finished with engine order” (when the ship leaves the departure port) and the “stand by engine order” (when the ships enters the arrival port) IMO (2000). The *Sailed Distance Over Ground* is the distance travelled by the vessel during the actual voyage navigation time, measured in Nautic Miles (NM) and calculated by the on-board data acquisition system from GPS latitude and longitude using the Haversine formula Veness (2007). The *Speed Over Ground* (SOG) is obtained as the ratio between the sailed distance over ground and the actual navigation time.

The *SOG variance* is the variance on the actual voyage navigation time, of the five-minute SOG average. This variable takes into account SOG variation (acceleration). *Head, following* and *side wind* are defined on the basis of comprehensive engineering considerations on the wind component influence. Let V_{WT} denote the true wind speed and Ψ_{WT} the difference between the true wind angle (in the earth system) and the course over ground ITTC (2008). Then W_H , W_F and W_S (Table 2.1) are obtained as the five-minute averages of the following quantities:

$$\begin{aligned} \tilde{W}_H &= \begin{cases} 0 & \text{if } 90^\circ \leq \Psi_{WT} \leq 270^\circ \\ V_{WT} \cos(\Psi_{WT}) & \text{otherwise} \end{cases}, \\ \tilde{W}_F &= \begin{cases} -V_{WT} \cos(\Psi_{WT}) & \text{if } 90^\circ \leq \Psi_{WT} \leq 270^\circ \\ 0 & \text{otherwise} \end{cases}, \\ \tilde{W}_S &= -|V_{WT} \sin(\Psi_{WT})|. \end{aligned} \quad (2.1)$$

Port and *starboard* shaft generator power measure the average power delivered by the shaft generators for electrical supply on board. The *power difference between two shaft generators* allows us to take into account the different modes of navigation (constant and combinator mode). *Power difference between two propeller shafts* is useful for discovering anomalies or malfunctioning in the main engines. *Departure* and *arrival trim* are obtained through the inclinometer measurements and the geometric features of the ship. *Draughts* are measured both at departure and arrival ports by four submersible transmitters located at fore and aft perpendiculars, and at port and starboard midship sections. *Displacement* is derived from the hydrostatic data on the basis of the mean draught at midship and trim.

Regression Methods

A large number of regression methods are available in the literature and can be used in the present context of predicting CO₂ emissions. The following four different classes of regression methods were contemplated, representing a rich variety of modeling approaches usually considered for handling large datasets possibly presenting collinearity and sparsity issues: variable selection methods, penalized regression methods, latent variable methods and tree-based ensemble methods. Each class contains methods that share similarities in terms of prior assumptions regarding the PVs, i.e. the operational variables defined in Table 2.1, the response variable (CO₂ emissions) and the relationship between them. Multiple linear regression (Montgomery et al. 2012, Draper & Smith 2014) (MLR) was also considered as it represents one of the most tested and studied methods. The MLR estimate of the $p+1$ coefficients b_0, \dots, b_p of a linear model is obtained by minimization of the mean squared prediction error

$$\hat{\mathbf{b}}_{\text{MLR}} = \arg \min_{\mathbf{b}=[b_0 \dots b_p]^T} \left\{ \sum_{i=1}^n (y(i) - \hat{y}(i))^2 \right\}, \quad (2.2)$$

where $\hat{\mathbf{b}}_{\text{MLR}}$ is a vector of regression coefficients, $y(i)$ is the i -th observed response value and $\hat{y}(i)$ is the corresponding model prediction ($\hat{y}(i) = b_0 + \sum_{j=1}^p b_j x_{i,j}$), n is the number of observations ($n = 538$) and p is the number of PVs ($p = 22$). A known drawback of MLR

is that for collinear predictors, the solution might get unstable or even unfeasible. In this case, alternatives such as latent variable methods are preferred since they are able to extract uncorrelated sources of structured variability in a stable way. On the other hand, MLR has the advantage of being simple to interpret and implement. In order to explore the agreement between the method’s assumptions and the dataset, a large collection of statistical tests and residual analysis tools have been developed in the last decades. The four classes of methods considered in this work are briefly presented below to address the limitations of MLR.

Variable selection methods. The variable selection class assumes that only some predictors are relevant for the prediction model, i.e. some PVs carry relevant information regarding the response while others are irrelevant and should be discarded in order to obtain more simple, robust and parsimonious models. Forward stepwise regression (Andersen & Bro 2010, Montgomery & Runger 2010) (FSR) was selected as the representative method from this class. In this method, the relevance of each predictor is assessed based on the p -value of the partial F -test and the most important predictor is selected. Then, at each subsequent step, the relevance of other predictors is assessed, conditioned on the predictors already selected, and the most relevant one is included in the model. Variables already included in the model can also be removed if they are later found to be insignificant, which can happen due to the inclusion of new variables in the mean time.

Penalized regression methods. The class of penalized regression methods imposes a penalty on the magnitude of the regression coefficients, thus constraining their values to be small. The penalty increases the bias of the regression model, but stabilizes the estimator variance so that better estimates of the regression coefficients can be obtained. Note that the variance decrease often compensates the bias increase so that the overall prediction performance can be improved. In this class, four regression methods were selected: ridge regression (Draper & Smith 2014) (RR), least absolute shrinkage and selection operator (Tibshirani 1996, Rasmussen & Bro 2012) (LASSO), elastic net (EN) and support vector regression (Smola & Schölkopf 2004, Ahmed et al. 2010, Canu et al. 2005) (SVR).

The EN is a general method and contains both RR and LASSO as special cases. The EN model is obtained by solving the following optimization problem:

$$\hat{\mathbf{b}}_{\text{EN}} = \arg \min_{\mathbf{b}=[b_0 \dots b_p]^T} \left\{ \sum_{i=1}^n (y(i) - \hat{y}(i))^2 + \gamma \left(\alpha \sum_{j=1}^p |b_j| + \frac{1-\alpha}{2} \sum_{j=1}^p b_j^2 \right) \right\}, \quad (2.3)$$

where γ is a hyper-parameter that controls the bias-variance tradeoff, and α is a hyper-parameter that weighs the squared (b_j^2) and the norm $|b_j|$ penalties. These hyper-parameters are often selected by cross-validation (Efron & Gong 1983) in order to ensure an appropriate compromise between (b_j^2) and $|b_j|$. The RR is obtained by setting $\alpha = 0$, i.e. only the squared penalty is imposed. The solution obtained with RR usually contains many small but non-zero coefficients, so that each variable has a small contribution to predict the response. On the other hand, LASSO is obtained by setting $\alpha = 1$ and the norm penalty often yields a sparse regression coefficient vector (i.e. many regression coefficients are effectively set to zero) which means that some predictors are discarded from the regression problem. By combining RR and LASSO penalty terms, the EN can adapt to a wide range of scenarios and applications. In particular, the method is able to include small contributions from a group of predictors while eliminating other groups.

The last method from the penalized regression class is the SVR (Smola & Schölkopf 2004,

Ahmed et al. 2010, Canu et al. 2005). The SVR minimizes the sum of squared regression coefficients $\left(\sum_{j=1}^p b_j^2\right)$ in order to decrease model variance but also constrains the prediction errors to be smaller than a certain threshold (ϵ). Additionally, slack variables are introduced to allow errors above the threshold, relaxing the optimization problem. Samples with errors above the threshold constitute the support vectors and contribute actively to establish the model.

Latent variable methods. The third class of methods are based on the latent variable framework (Burnham et al. 1999) where a set of unmeasured quantities are considered to be responsible for the observed variability in both predictors and response variable. Linear combinations of the measured variables are used to estimate the latent variables and the model obtained can be used for predicting the response based on the available predictors. In this class, three regression methods were considered: principal component regression (PCR) (Jolliffe 2002, Jackson 2005, Wold et al. 1987), principal component regression with the scores added in a forward stepwise fashion (PCR_{FS}) and PLS regression (Wold et al. 2001, Geladi & Kowalski 1986, Wold et al. 1984). The latent variable model is described by the following equations:

$$\begin{aligned}\mathbf{X} &= \mathbf{TP}^\top + \mathbf{E}, \\ \mathbf{y} &= \mathbf{Tc} + \mathbf{f},\end{aligned}\tag{2.4}$$

where \mathbf{T} is a $n \times a$ orthogonal matrix of scores, \mathbf{P} is a $p \times a$ loading matrix, \mathbf{c} is an $a \times 1$ vector that relates scores to the response, and \mathbf{E} and \mathbf{f} correspond to residuals. PCR is applied after a principal component analysis (PCA) decomposition of the predictor space where most of its variability is explained by a number $a = \alpha_{PCR}$ of principal components, which is typically much smaller than p and is selected by cross-validation in order to avoid overfitting. Since only a few orthogonal columns of the score matrix \mathbf{T} (i.e. principal components) are retained in the model, the final covariance matrix is easily inverted. Then a model structure similar to MLR is adopted to relate these scores (i.e. principal components) with the response variable. Note that for PCR, the matrix \mathbf{T} is obtained by purely considering the variability in the predictor space. The PCR_{FS} is quite similar to the PCR but PCs are selected using the forward stepwise algorithm based on the p -value of the partial F -test (as described for FSR). Finally, PLS regression estimates directions with maximum covariance between predictors and the response variable. However, unlike PCR, PLS scores (\mathbf{T}) also contain information for describing the response variable and, at the same time, provide a good approximation of the predictor space. Again, the number of latent variables (α_{PLS}) is selected by cross-validation in order to choose a suitable model complexity.

Tree-based ensemble methods. The last class of regression methods is based on ensembles of regression trees (Hastie et al. 2009, Dietterich 2000, Strobl et al. 2009, Breiman et al. 1984). A regression tree is a model that splits the predictors' space into regions where the response variability is small. The algorithm starts by identifying a split variable and a split point so that the predictors' space is divided into two regions where the sum of the response variance is as small as possible. Then, a greedy search is used to identify more split points and split variables, further decreasing the response variability. A criterion must be defined in order to stop the splitting process. In this work, trees are built until a minimum of five samples are obtained for each region, a standard value that can also be optimized. Once the splitting procedure stops, each region represents a leaf in the final regression tree. When

predicting a new sample, the value of its predictors are used to map the sample into one of the tree’s leaves. Then, the mean response of the samples within the leaf is the predicted value. Ensemble methods are adopted to further decrease the variance of the trees, which is indeed very high when trees are deep and therefore could overfit the training sets. In this class, three regression methods were considered: bagging of regression trees (BRT), random forests (RF) and boosting of regression trees (BT). BRT applies bootstrap to the original dataset in order to generate many datasets. Each one of these datasets is used to build a regression tree and the average prediction from all trees in the ensemble constitute the predicted response. Similarly, RF uses bootstrap to generate datasets, which however only contain a smaller number (e.g. \sqrt{p}) of randomly selected predictors. For both BRT and RF, the number of trees (T_{BRT} and T_{RF} , respectively) in the ensemble is selected by cross-validation and controls the bias-variance tradeoff. Lastly, BT (Elith et al. 2008, Cao et al. 2010, Freund et al. 1996) iteratively fits regression trees to the residuals from previous trees so that each tree focuses on samples that were poorly modelled. In order to avoid overfitting, BT considers only a small fraction (u) of the response at each iteration. The fraction u , also known as the learning rate, and the number of trees are usually inversely related (i.e. a small learning rate requires a larger number of trees and vice-versa). In this work, the learning rate is fixed at a small value ($u=0.02$) and the number of trees (T_{BT}) is selected by cross-validation.

2.3 Comparison framework

A double cross-validation (Filzmoser et al. 2009) procedure was devised to compare the wide range of regression methods considered in this particular case study, as well as to obtain estimates of their prediction performance as measured by the root mean squared error of double cross-validation ($RMSE_{dcv}$). The proposed procedure has the advantage of providing variability estimates of $RMSE_{dcv}$ due to the different data splits.

The procedure starts by randomly partitioning the complete dataset into training and test datasets. In this work, 80% of the data constitute the training set whereas the remaining 20% represents the test set. The training set is used to build models and choose suitable values of the hyper-parameters, which are selected by 10-fold cross-validation. When the model building stage is complete, the trained models are used to predict the test set and the $RMSE_{dcv}$ is computed. Since the $RMSE_{dcv}$ is affected by the initial split of the complete dataset, the whole procedure is repeated in order to obtain a measure of its variability. Then, the complete dataset is again randomly partitioned into training and test sets: models are built based on the training set and used to compute the $RMSE_{dcv}$ over the test set. In this work, 40 iterations of double cross-validation allowed to have estimates of the mean and the variance of $RMSE_{dcv}$ and to characterize the performance of each regression method much better than the point estimates that would be obtained by a single dataset split.

In the model building stage, a 10-fold cross-validation scheme was adopted to select the most suitable hyper-parameters. This scheme has been extensively applied in many application domains and found to be robust to overfitting. Table 2 presents the values of the hyper-parameters tested for each regression method considered.

Finally, variables were centered and scaled because the results of regression methods can be sensitive to the measurement units. Since no prior knowledge was available to define suitable scaling factors for all variables, auto-scaling was adopted by mean-centering each variable (i.e. the average value of each variable is subtracted from the data) and dividing each variable by the sample standard deviation. In each iteration of double cross-validation,

Table 2.2: Hyper-parameters tested during model training. The suitable value for each method is selected by 10-fold cross-validation.

	Hyper-parameter(s)	Possible value(s)
MLR	-	-
FSR	p_{enter}	0.05
	p_{rem}	0.1
RR	α	0
	γ	0.002;0.02;0.2;2;20
LASSO	α	1
	γ	0.002;0.02;0.2;2;20
EN	α	0.001;0.01;0.1
	γ	0.002;0.02;0.2;2;20
SVR	ϵ	0.001;0.005;0.01;0.05;0.1
PCR	α_{PCR}	$[1 : \min(20, n, p)]$
PCR _{FS}	p_{enter}	0.05
	p_{rem}	0.1
PLS	α_{PLS}	$[1 : \min(20, n, p)]$
BRT	T_{BRT}	50; 100;500; 1000; 5000
RF	T_{RF}	50; 100;500; 1000; 5000
BT	T_{BT}	50; 100;500; 1000; 5000

variable centering and scaling were applied in a consistent way: the mean and the standard deviation of each variable to be used for scaling the test set were calculated based on the training set.

2.4 Results and discussion

The comparison framework described in the previous section was applied to the dataset collected from the Ro-Pax cruise ship and to all the regression methods considered, in order to assess their performance for predicting the CO₂ emissions. The results obtained from this comparison framework also allowed the identification of the PVs that most influence CO₂ emissions. All computations were performed in the Matlab platform (version 2015b, The Mathworks, Inc.).

In order to examine the available dataset and obtain preliminary insights, an exploratory data analysis was carried out prior to the application of the comparison framework. The number of operational variables available is relatively small (22 PVs). A simple way to measure the predictive power of individual variables is to compute the correlation between each variable and CO₂ emissions, as reported in Figure 2.1. From the correlation values obtained from all routes reported in Figure 2.1, it is trivial to observe that the actual voyage navigation time (PV 5) and sailed distance over ground (PV 22) are the variables with the highest correlation (very close to 1) with the CO₂ emissions.

Based on these high correlation values, one might expect the regression problem to be straightforward and even a simple ordinary least squares model to give satisfactory predictions. However, correlation does not provide information regarding the sample distribution. Therefore, a PCA model was built in order to better explore the sample space. Figure 2.2a

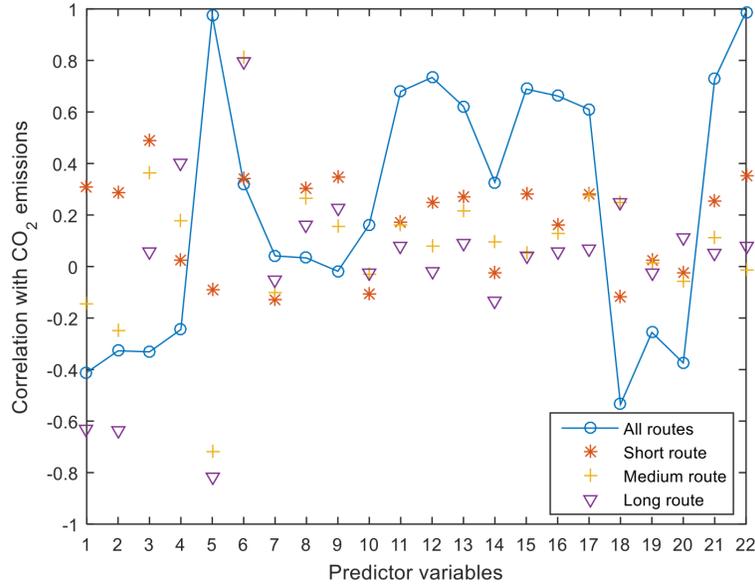


Figure 2.1: Correlation between predictors and CO₂ emissions.

presents the scores for the first two principal components. Analyzing Figure 2.2a, one can verify the existence of at least two clearly separable clusters. The presence of clusters is known to artificially increase correlation even when it is zero or near zero within each cluster. Further exploration of the dataset revealed that three clusters can be identified when plotting CO₂ emissions and the distance over ground (Figure 2.2b), corresponding to three different types of routes: short (average 120 NM), medium (average 150 NM) and large (average 400 NM). Note that response variable values were hidden for confidentiality reasons, but scale was left unchanged. When correlation is analyzed within each cluster (Figure 2.1), the values are generally smaller and an interesting behavior is observed for the actual voyage navigation time (PV 5): although the correlation is high and positive in the complete dataset, it becomes negative for the medium and long routes and is quite small for the shortest route.

In summary, exploratory data analysis revealed the presence of three clusters, each corresponding to a different route. In order to remove this effect and avoid inflated correlations, the performance of the considered regression methods was studied for each route separately.

Prediction of CO₂ emissions—short route

The short route constitutes the first cluster (218 samples) and the correlation between CO₂ emissions and the operational variables is rather weak (Figure 2.1, short route), suggesting that the regression methods might not be able to predict the response accurately. The $RMSE_{dev}$ obtained for all regression methods included in the comparison study are presented in Figure 2.3. From this figure, one can verify that most methods present similar performances, except for PCR_{FS}, BRT, RF and BT, which had prediction errors that were almost twice the errors obtained with the best methods. Thus, the class of tree-based ensembles does not appear to be particularly suitable for the short route. This may be justified by the piece-wise

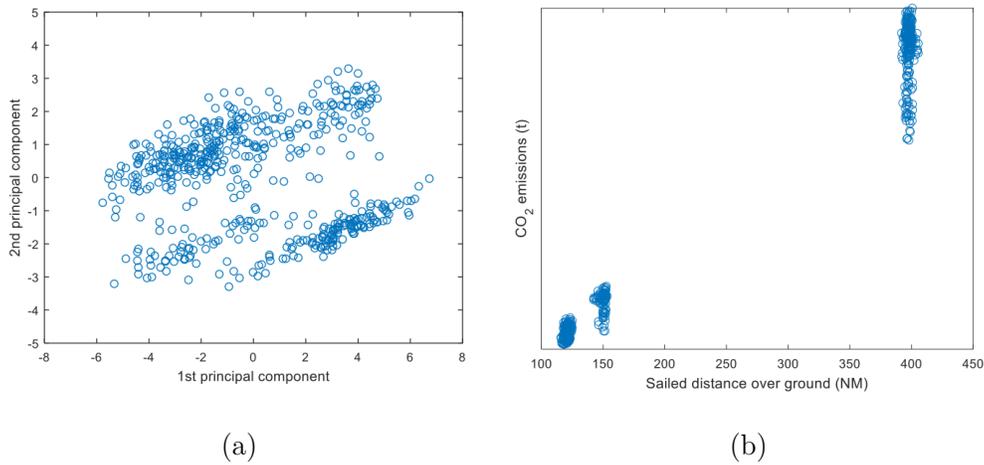


Figure 2.2: Analyzing the samples' space: (a) the scores from a PCA model reveal clusters in the data, which can be well identified by the sailed distance over ground (b). Analyzing the samples' space: (a) the scores from a PCA model reveal clusters in the data, which can be well identified by the sailed distance over ground (b).

nature of the regression trees. Indeed, it is expected that most operational variables have a continuous (linear or nonlinear) relation with the response and the approximation obtained with piece-wise functions does not seem to be effective at approximating this relationship.

In terms of regression methods with good performances, it can be seen in Figure 2.3 that the difference between the top methods is not pronounced nor statistically significant. The method with the minimum median $RMSE_{d_{cv}}$ is the RR. However, due to its similar performance to other regression methods, LASSO is the recommended approach for predicting CO₂ emissions for the short route. As is known, the parsimony principle (also called Occam's razor) (Seasholtz & Kowalski 1993) states that when two methods have the same performance for a regression problem, the simpler method has to be preferred. LASSO often generates solutions where some PVs are discarded due to the penalization term and the model tends to be more parsimonious. The LASSO model also has a smaller variance than the MLR model (due to correlation between predictors, the MLR regression coefficients were quite sensitive to the training set used in the double cross-validation procedure). Thus, although MLR may be more familiar to practitioners, the parsimony principle provides a robust argument for selecting LASSO since it presents similar prediction errors but has a smaller number of regression coefficients. This fact also facilitates interpreting the LASSO model since some PVs have zero regression coefficients and do not contribute to predict the response.

In terms of latent variable methods, an interesting fact was observed for both PLS and PCR: the hyper-parameters selected were very often at the higher boundary of possible values. For example, the number of latent variables used in PLS for all 40 iterations of double cross-validation was 20, the maximum number allowable as specified in Table 2. This result suggests that the structure underlying the dataset may not be that of a latent variable model.

In order to further explore results from LASSO, Figure 2.4a presents the regression

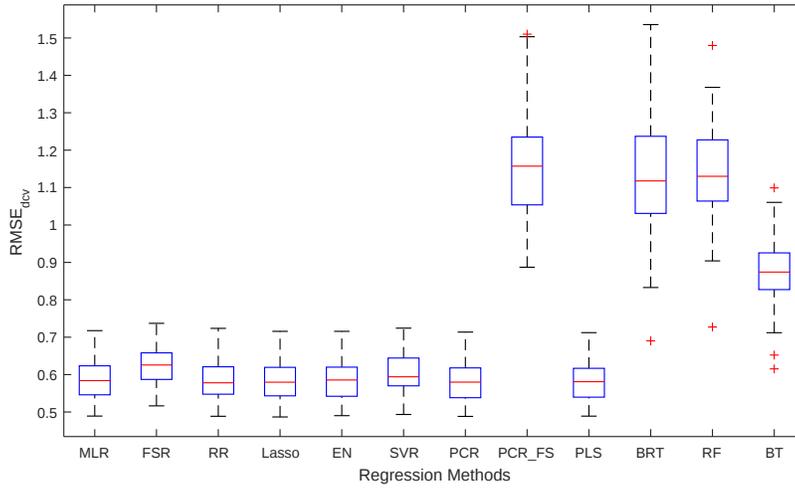


Figure 2.3: Distribution of $RMSE_{dcv}$ obtained when predicting CO_2 emissions for the short route.

coefficients obtained for all 40 iterations of double cross-validation, whereas Figure 2.4b presents the plot of predicted against observed CO_2 emissions. In Figure 2.4a, one can note that CO_2 emissions are mostly influenced by the actual voyage navigation time (PV 5), then by the power difference between two shaft generators (PV 13) and the arrival draught (midship section—starboard, PV 17). Furthermore, it is interesting to note that the first four variables of Table 2.1 do not contain relevant information regarding the response as well as the presence of some iterations affecting the variable arrival draught (midship section—port, PV 16). The latter has positive regression coefficients in some iterations of double cross-validation and negative coefficients in other ones. Figure 2.4b shows that the plot of predicted against observed CO_2 emissions follows the 1:1 line and residuals do not exhibit patterns.

Prediction of CO_2 emissions—medium route

With regard to the second cluster, which contains 85 samples, one can observe in Figure 2.1 that the correlation between CO_2 emissions and the actual voyage navigation time (PV 5) is negative, and large in magnitude. This observation agrees with theoretical considerations because higher navigation times correspond to lower speeds. Since fuel consumption per hour is expected to vary with the vessel's speed raised to the third power, lower speeds (i.e. longer navigation times) tend to have lower CO_2 emissions.

The $RMSE_{dcv}$ obtained in all 40 iterations of double cross-validation for the medium route are presented in Figure 2.5. In Figure 2.5, one can observe similar trends as those observed for the short route. In particular, methods based on regression trees have the highest prediction errors while the other classes achieve comparable ones. As in the previous case, using 10-fold cross-validation, the maximum allowable number of latent variables was selected for most iterations of double cross-validation.

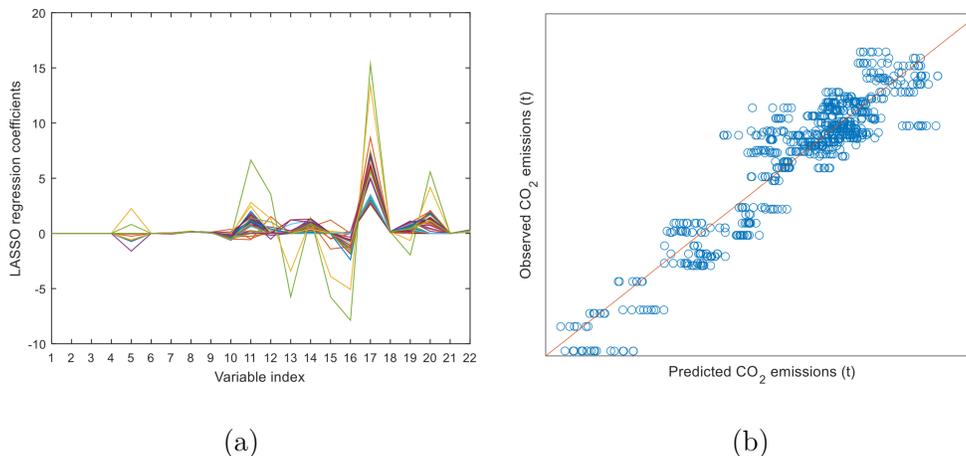


Figure 2.4: Results from LASSO model training for the short route: (a) regression coefficients and (b) the predicted and observed CO₂ emissions for all 40 iterations of double cross-validation.

Strictly speaking, the regression method with the smallest median $RMSE_{d_{cv}}$ is the SVR, although the variability around the median value is larger when compared to other good methods. As for the short route, the more parsimonious model and the smaller model variance indicates the LASSO method as the recommended approach. For further inspection, Figure 6 presents the regression coefficients as well as the predicted and observed CO₂ emissions in all 40 iterations of double cross-validation for the LASSO method. In particular, Figure 2.6a shows that arrival draught (midship section—starboard, PV 17) is the most important predictor, followed by arrival trim (PV 20) and departure draught (aft perpendicular, PV 11). From Figure 2.6a, it is also worth noting that many regression coefficients are zero (i.e. the corresponding predictors are not included in the model) which suggests a sparse structure of the dataset. Furthermore, the prediction errors tend to be relatively small over all iterations of double cross-validation. Figure 2.6b supports the successful prediction of CO₂ emissions for the medium route, reporting a satisfactory median R^2 (0.86).

Prediction of CO₂ emissions—long route

The last cluster contains 235 samples from the long route. Figure 2.7 presents the $RMSE_{d_{cv}}$ obtained for all 40 iterations of double cross-validation. As for the previous routes, the class of variable selection, penalized regression and latent variable methods (except for PCR_{FS}) present similar performances. However, the class of tree-based methods now have prediction errors which are in the same range as other classes and BT is actually the best regression method with a median $RMSE_{d_{cv}}$ of 2.67. This is confirmed, at 5% significance level, also by means of a multiple paired t-test conducted to compare its $RMSE_{d_{cv}}$ distribution with those obtained through the other regression methods. For all iterations of double cross-validation, Figure

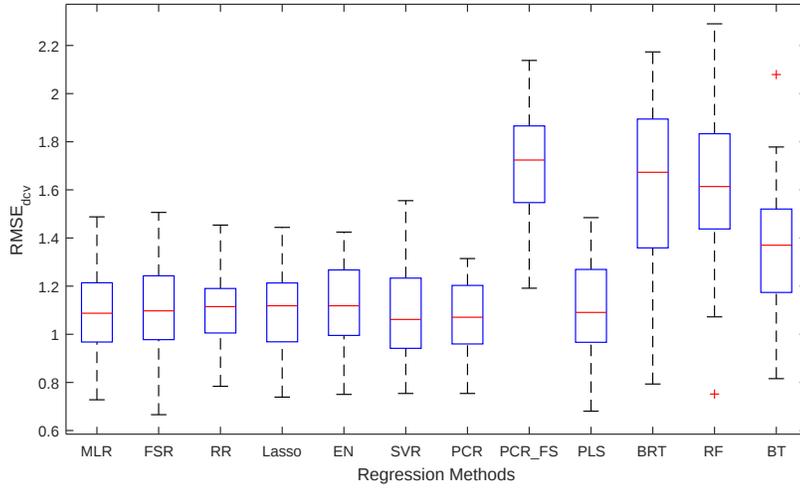


Figure 2.5: Distribution of $RMSE_{dcv}$ obtained when predicting CO_2 emissions for the medium route.

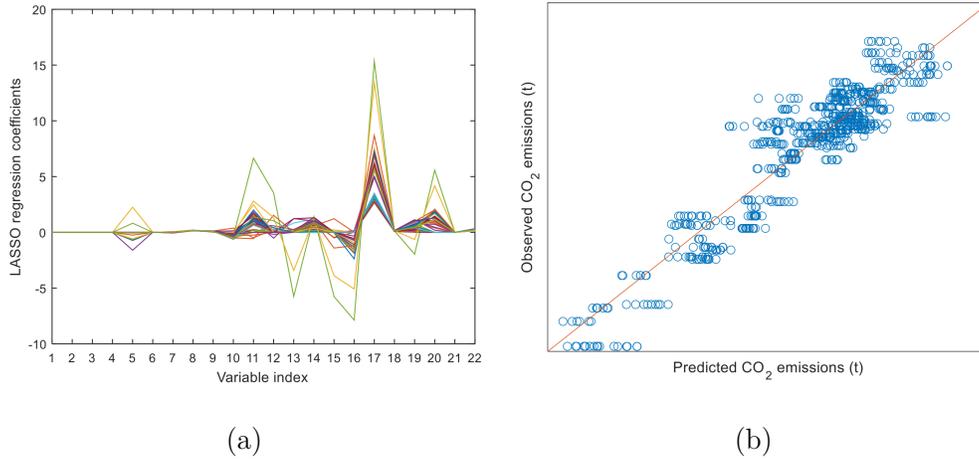


Figure 2.6: Results from LASSO model training for the medium route: (a) regression coefficients and (b) the predicted and observed CO_2 emissions for all 40 iterations of double cross-validation.

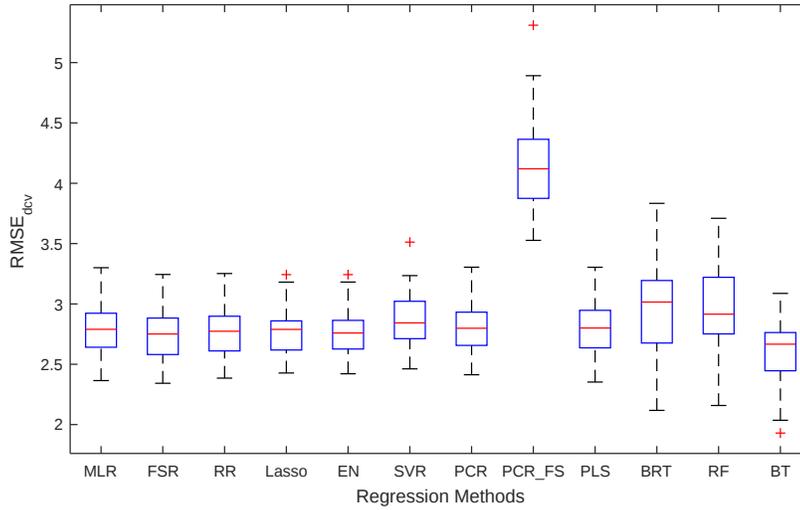


Figure 2.7: Distribution of $RMSE_{dev}$ obtained when predicting CO₂ emissions for the long route.

2.8a and Figure 2.8b present the predictors' importance and the plot of predicted vs observed CO₂ emissions, respectively. In particular, Figure 2.8a shows that actual voyage navigation time (PV 5) and speed over ground (PV 6) are the two most important predictors. The agreement between predicted and observed CO₂ emissions (Figure 2.8b) also confirms that BT accurately model the relation between the predictors and the response variable with median R^2 equal to 0.91.

As a summary, the comparison framework pointed to reliable regression methods that can be adopted for predicting CO₂ emissions. The fact that the recommended method depends on the route type corroborates the need to conduct a comparison study since the available a priori knowledge was not enough to select a suitable regression method. The absence of an overall best method is in accordance with theoretical considerations since the data generating mechanisms for the different route types ought not to be the same. Thus, by considering different classes of regression methods and a robust comparison scheme, a large number of prior assumptions regarding the data generating mechanisms were explored and a more consistent assessment of prediction performance was obtained.

2.5 Conclusions

The ongoing changes in regulations regarding the control of greenhouse gases emissions are having a great impact in the marine industry and are urging shipping operators to correctly measure and predict CO₂ emissions of their fleet. However, there is no incontrovertible method that can be satisfactorily adopted in real environments as the best one to turn the

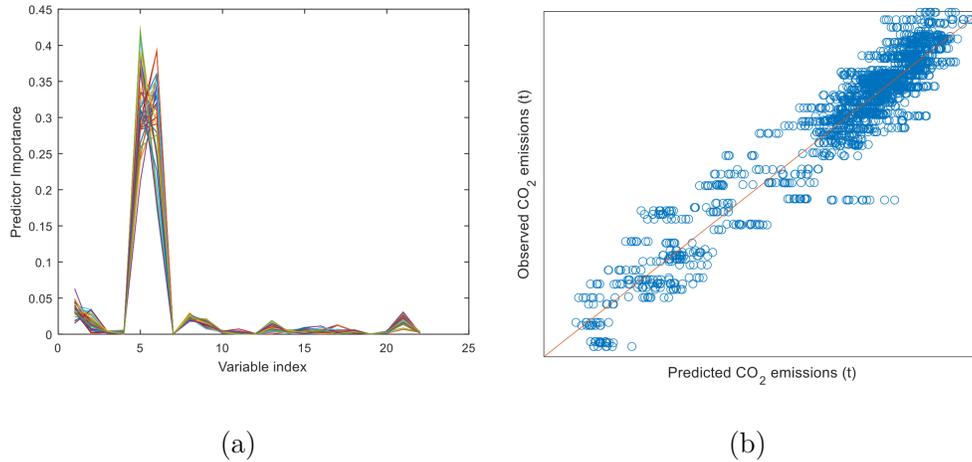


Figure 2.8: Results from BT model training for the long route: (a) regression coefficients and (b) the predicted and observed CO₂ emissions for all 40 iterations of double cross-validation.

operational information into value. In this context, a robust and thorough comparison is critical especially when applied to real world complex data. This work provided directions on the use of advanced regression techniques for predicting CO₂ emissions for a real shipping industry scenario.

The regression methods considered in this study were grouped in four classes, namely variable selection, penalized regression, latent variable methods and tree-based ensembles. Representative methods from each class were selected and a double cross-validation procedure was used to estimate their prediction errors. The analysis of one year’s worth of data collected on-board of a modern Ro-Pax cruise ship owned by the Italian company *Grimaldi Group* revealed clusters corresponding to three different route lengths—namely short (average 120 NM), medium (average 140 NM) and long (average 400 NM). In order to accurately model each route, double cross-validation was conducted for each cluster separately. For the short route: the class of tree-based ensembles and latent variable methods did not appear to be particularly suitable, whereas most of the regression methods presented similar good performances. However, LASSO (with a median $RMSE_{dcv}$ equal to 0.58) was preferred in the light of the parsimony principle. Voyage navigation time, power difference between shaft generators and arrival draught were shown to be the most influencing variables on the response. For the medium route: LASSO (with a satisfactory median $RMSE_{dcv}$ equal to 0.86) was the more parsimonious model with the smaller variance and therefore, the recommended technique also in this case. Arrival draught, arrival trim and departure draught (aft perpendicular) were the most important predictors over a noticeable sparse structure of the LASSO regression coefficients. For the long route: BT presented the best prediction results (with a median $RMSE_{dcv}$ equal to 0.91). The actual voyage navigation time and speed over ground were shown to be the two most important predictors for this route. This is in accordance with theoretical (white-box) approaches that expect CO₂ emissions to vary with the vessel’s speed raised to the third power. In fact, only for the medium and long routes, low speeds over ground achieve higher navigation times and therefore, significant

2. PREDICTING SHIP CO₂ EMISSIONS

CO₂ emission reductions.

Chapter 3

LS-PLS approach to ship fuel-speed curves

The shipping industry relies on ship fuel-speed curves to describe the fuel consumption (and CO₂ emissions levels) per hour as a function only of the vessel's speed over ground, based on dedicated test data. However, they are affected by additional factors in real cases. In this chapter, a novel method is developed elaborating the orthogonal least-squares partial least-squares (LS-PLS) approach to enhance fuel-speed curves accuracy when information is available on additional factors from multi-sensor systems. Through real data examples, the approach is shown capable of detecting anomalies in CO₂ emission levels and testing the effectiveness of ship energy efficiency initiatives.

3.1 Introduction

The maritime shipping has a great need for benchmarks and measures of fuel efficiency, which is the major factor in establishing the price required for faster travel (Gabrielli & von Karman 1950). In this context, fuel efficiency is usually calculated as the ratio between fuel consumption per hour (FCPH) and mean hourly vessel's speed over ground (SOG). In general, it represents the relationship between the distance travelled and the amount of fuel consumed. Maritime engineers are in fact crucially concerned with the use of fuel-speed curves (St Amand 2012), which are essentially two-dimensional graphs that represent the cubic physical-based relationship between fuel consumption and SOG (Ronen 1982, Schrady et al. 1996, Fagerholt et al. 2010, MAN Diesel & Turbo 2011). These curves are also popular for estimating levels of CO₂ emissions (i.e. carbon dioxide), as the latter are stoichiometrically related to fuel consumption.

It is worth noting that, in maritime engineering, fuel-speed curves are equivalent to the power-speed curves, because a direct proportional relationship also holds between engine power and fuel consumption through the power-based specific fuel oil consumption coefficient (Corbett & Koehler 2003, MAN Diesel & Turbo 2011, Lu et al. 2015, Bialystocki & Konovessis 2016). Hereinafter we refer only to fuel-speed curves, even though the obtained results can be trivially extended to power-speed curves.

So far, ship performance verification protocols refer to fuel-speed curves calibrated on data from dedicated speed trials under standard reference conditions (Bazari 2007) only. By means of these data, as common practice, fuel-speed curves are estimated through the ordinary

least squares (LS) method. However, this method may give inaccurate fuel predictions when based on operational data (i.e. under non-standard conditions), because of several additional factors that can in reality affect vessel performance (Molinero & Mitsis 1984, Lu et al. 2013, 2015).

Indeed, vessel operators are nowadays in the position to access massive additional operational data from modern multi-sensor systems, which provide great potential for managerial decision-making by improving fuel-speed curve accuracy. Efforts in this direction are still very sparse in the naval literature, however. The main exception is provided by Bialystocki & Konovessis (2016), who however simply suggest a univariate statistical approach for calculating the influence of ship draft and displacement, weather force and direction, and hull roughness on the fuel-speed curve. On the other hand, multiple linear regression (MLR) has been historically used to budget (Molinero & Mitsis 1984) and, more recently, to monitor ship fuel consumption (Bocchetti et al. 2015, Erto et al. 2015).

Monitoring of fuel consumption represents in fact a compelling task imposed by the EU regulation 2015/757 (European Commission 2015) that came into force from January 2018. In such a regulatory regime, there is also a managerial need for testing whether claims for ship fuel consumption reduction after a specific energy efficiency initiative (EEI) (e.g. hull form optimization, hull cleaning and propeller polishing, ultra-smooth coating, propulsion efficiency improvement, engine maintenance operation, power plant efficiency improvement) are true. Unfortunately, canonical applications of methods based on multivariate statistical techniques and artificial neural networks (see e.g., Petersen et al. (2012), Beşikçi et al. (2016)) do not appear to attract shipping practitioners, who are used to dealing with classical fuel-speed curves.

The aim of this chapter is then to mitigate the drawbacks of those curves when applied to operational data, and to give marine engineers the additional opportunity to use a familiar tool for monitoring and testing purposes of fuel efficiency and CO₂ emissions.

In order to do that, prediction accuracy of the fuel-speed curves is enhanced by exploiting the massive amounts of available operational data via the orthogonal least squares-partial least squares (LS-PLS) method (Jørgensen et al. 2004, 2007), firstly. This method avoids the collinearity issue between additional factors (Montgomery et al. 2012), which usually bounds the number of additional factors to be included in the MLR models. The main focus of Jørgensen et. al. was to point out that the LS-PLS estimation method makes interpretation easier with respect to the standard partial least-squares (PLS). Their orthogonal LS-PLS algorithm is in fact used in this chapter to estimate parameters of the model and to maintain the physical relation between FCPH and SOG expected by marine engineers. Besides, we newly expand the method also for constructing opportune control charts for monitoring the predictor variables as well as the response variable. In fact, two multivariate control charts are introduced to detect whether future observations move away from the normal operating conditions or not (Duchesne & MacGregor 2004). Voyages with unusual variations are further investigated through contribution plots corresponding to those control chart, which gives the opportunity of supporting the prognosis of faults, unfeasible via the classical fuel-speed curves. Otherwise, for voyages with unusual variations, the fuel consumption is monitored through the construction of pointwise prediction intervals for the proposed LS-PLS normalized fuel-speed curve.

Lastly, the enhanced prediction accuracy of the proposed LS-PLS normalized fuel-speed curve is exploited also for comparing graphically fuel efficiency before and after EEIs, and for estimating any consequent reduction in terms of fuel consumption and CO₂ emissions. This feature can be regarded as a relevant proposal to entitle shipping operators to directly

cash in carbon credits according to the international emission trading mechanism set by the Kyoto Protocol (Zhang & Folmer 1995, Ki-Moon 2008). Moreover, the proposed LS-PLS normalization approach may be used to overcome the indications of weakness reported in the ISO guidelines (ISO 2015) for normalization purposes (Bazari 2007) of the fuel-speed curve estimation based on real operational data.

The remainder of the chapter is as follows: in Section 3.2 the orthogonal LS-PLS approach to ship fuel-speed curves is introduced; in Section 3.3 the approach is applied to real operational data automatically acquired from a roll-on/roll-off passenger (Ro-Pax) cruise ship that is owned by the Italian shipping company *Grimaldi Group*. Finally, Appendix A validates the use of the orthogonal LS-PLS with respect to classical MLR method; Appendix B outlines the main steps of the PLS algorithm; Appendix C reports a simulation study to verify the coverage of the approximate pointwise prediction intervals for the proposed LS-PLS normalized fuel-speed curve.

3.2 The proposed approach

The section is structured in four subsections. In Section 3.2, a variant of the orthogonal LS-PLS method is presented and customized to enhance the ordinary LS ship fuel-speed curve, while maintaining the original physical-based relationship between the cubed SOG and the FCPH. Section 3.2 shows how to obtain the LS-PLS normalized fuel-speed curve and, accordingly, opportune control charts are provided and demonstrated to be relevant tools for monitoring FCPH and supporting the diagnosis of faults. Then, in Section 3.2, a hypothesis testing procedure is developed to assess a change in the mean fuel consumption per hour (e.g. before and after an EEI that aims to improve ship performance). Lastly, in Section 3.2, a practical solution is proposed for estimating fuel consumption reduction achieved in a given period that follows an effective EEI.

The orthogonal LS-PLS method

As outlined, from physical-based models, the relationship between fuel and speed is known to be linearized by considering the cubed SOG. Starting from this assumption, let us consider a response variable Y (e.g., the mean FCPH over a single voyage), a predictor variable X (e.g., the vessel's cubed SOG) and a vector $Z = (Z_1, Z_2, \dots, Z_{m-1})^\top$ of $m - 1$ additional factors. Moreover, let us consider n observations $\mathbf{y} = (y_1, \dots, y_n)^\top$ of Y , $\mathbf{x} = (x_1, \dots, x_n)^\top$ of X , and

$$\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_{m-1}) = \begin{pmatrix} z_{11} & \cdots & z_{1(m-1)} \\ \vdots & & \vdots \\ z_{n1} & \cdots & z_{n(m-1)} \end{pmatrix} \quad (3.1)$$

of Z . Each column of \mathbf{Z} , \mathbf{x} , and \mathbf{y} is mean centred and scaled to unit variance prior to performing the regression analysis. Then, the simple linear regression model without intercept of Y on X can be written as follows

$$\mathbf{y} = \mathbf{x}\beta + \boldsymbol{\varepsilon}_{LS}, \quad (3.2)$$

where the vector $\boldsymbol{\varepsilon}_{LS}$ contains the n random error terms, which are assumed independent identically distributed with zero mean and uncorrelated with X . The ordinary LS estimate

of β is $\hat{\beta} = (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathbf{y}$, and the ordinary LS prediction of \mathbf{y} is then obtained as

$$\hat{\mathbf{y}}_{LS} = \mathbf{x} \hat{\beta}. \quad (3.3)$$

Trivially note that the vector of residuals

$$\mathbf{e}_{LS} = \mathbf{y} - \mathbf{x} \hat{\beta} \quad (3.4)$$

is orthogonal to the predictor variable vector \mathbf{x} , i.e. $\mathbf{e}_{LS}^\top \mathbf{x} = 0$.

However, unless Z is uncorrelated with X or Y , if we consider a new model $\mathbf{y} = \mathbf{x}\beta + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, by simply adding a new term in Equation (3.2), then the ordinary LS estimator of β is biased. This is known as the omitted variable bias problem (Clarke 2005) However, if there is a linear relationship between X and Z , a new term can be added to Equation (3.2), while still maintaining the unbiasedness of the ordinary LS estimator $\hat{\beta}$. In order to prove that, let us consider the following multi-response model without intercept

$$= \mathbf{x} \boldsymbol{\delta}^\top + \mathbf{Z}_{orth}, \quad (3.5)$$

where $\boldsymbol{\delta} = (\delta_1, \dots, \delta_{m-1})^\top$ is a vector of parameters, and \mathbf{Z}_{orth} is a $n \times (m-1)$ matrix of errors, with zero mean and uncorrelated with X . The vector $\boldsymbol{\delta}$ is estimated as $\hat{\boldsymbol{\delta}} = (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathbf{Z}$. The model in Equation (3.5) allows to remove the linear dependence of Z on X . Then, the following linear regression model can be used to describe the relationship between the response variable Y and both the predictor variable X and the additional factors, and to apply the LS-PLS method (Jørgensen et al. 2004, 2007)

$$\mathbf{y} = \mathbf{x}\beta + \mathbf{Z}_{orth}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \quad (3.6)$$

The vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_{m-1})^\top$ contains the additional parameters to be estimated, whereas $\boldsymbol{\varepsilon}$ contains the n random error terms, independent identically distributed with zero mean and uncorrelated with X and \mathbf{Z}_{orth} (viz., also with Z). The model can be also written as

$$\mathbf{y} = \mathbf{x}(\beta - \boldsymbol{\delta}^\top \boldsymbol{\gamma}) + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \quad (3.7)$$

In any case, since X and \mathbf{Z}_{orth} are uncorrelated, $\hat{\beta}$ is still an unbiased estimator of β and is independent of the additional factors Z . Then, the matrix of the residuals

$$\mathbf{Z}_{orth} = \mathbf{Z} - \mathbf{x} \hat{\boldsymbol{\delta}} \quad (3.8)$$

of model (3.5) can be used as predictor of \mathbf{y} . Note that, although columns of \mathbf{Z}_{orth} are orthogonal to \mathbf{x} , i.e. they contain the orthogonal components of the additional $m-1$ factors, they are not necessarily orthogonal to each other. Therefore, the standard MLR technique cannot be applied and methods for dimension reduction are required in case of collinearity among variables (i.e. if the rank of \mathbf{Z}_{orth} is less than $m-1$, which happens when modern systems acquire data from many sensors). This is the reason for performing a PLS regression of \mathbf{y} on \mathbf{Z}_{orth} . PLS allows in fact to obtain orthogonal scores \mathbf{T} from the loading matrix \mathbf{P} and the weight matrix \mathbf{W} (see e.g. Helland (1988), Phatak & De Jong (1997)) as follows

$$\mathbf{T} = \mathbf{Z}_{orth} \mathbf{W} (\mathbf{P}^\top \mathbf{W})^{-1}. \quad (3.9)$$

Then, the regression of \mathbf{y} on the scores \mathbf{T} yields $\hat{\boldsymbol{\gamma}} = (\mathbf{T}^\top \mathbf{T})^{-1} \mathbf{T}^\top \mathbf{y}$ and the PLS prediction of \mathbf{y} is

$$\hat{\mathbf{y}}_{PLS} = \mathbf{T} \hat{\boldsymbol{\gamma}}_{PLS}. \quad (3.10)$$

By setting $\hat{\boldsymbol{\gamma}} = \mathbf{W}(\mathbf{P}^\top \mathbf{W})^{-1} \hat{\boldsymbol{\gamma}}_{PLS}$, we obtain an estimate of $\boldsymbol{\gamma}$ in Equation (3.6). A discussion on the properties of the PLS estimator is provided in Krämer (2007).

The number of columns of \mathbf{T} , i.e. the number of scores a , is determined through the leave-one-out cross-validation (Efron & Gong 1983, Geladi & Kowalski 1986). In this chapter, the evaluation of the number of scores is based on the prediction residual sum of squares (PRESS) statistic (Geladi & Kowalski 1986). The cross-validation procedure assigns to each number of latent variables a value of the PRESS statistic, in order to find the optimal number of components to retain in the model. Given a number of components, for each of the original n observations, the PLS model is built on the other $n - 1$ ones, and the prediction error on the observation left out from the model is computed. The PRESS statistic for that number of components is calculated as the sum of these squared prediction errors, and the optimal number of components is chosen as the one corresponding to the minimum PRESS value. Since \mathbf{x} and \mathbf{Z}_{orth} are orthogonal, the LS-PLS response prediction of \mathbf{y} can be written as

$$\hat{\mathbf{y}}_{LS-PLS} = \hat{\mathbf{y}}_{LS} + \hat{\mathbf{y}}_{PLS} \quad (3.11)$$

and the vector of the final LS-PLS residuals is given by

$$\mathbf{e}_{LS-PLS} = \mathbf{y} - \hat{\mathbf{y}}_{LS-PLS} \quad (3.12)$$

With this strategy, it is possible to achieve scores that are orthogonal each other and to \mathbf{x} simultaneously. In contrast with the direct application of the PLS method, Equation (3.11) expresses $\hat{\mathbf{y}}_{LS-PLS}$ as a combination of physical variables and scores that still preserves orthogonality among predictor variables. By keeping the information in the different matrices separated, Equation (3.11) then has a more straightforward interpretation with respect to the canonical PLS models. In the Appendix A, the PLS is shown to outperform the MLR in terms of predictive ability.

Fuel consumption monitoring through normalized fuel-speed curve

The fitted prediction values of FCPH $\hat{\mathbf{y}}_{LS}$, calculated by Equation (3.3), can be plotted against the vessel's cubed SOG observations \mathbf{x} to draw the ordinary LS fuel-speed curve. The $100(1 - \alpha)$ percent LS prediction interval for the future observation of the FCPH y_0 can then be calculated based on the corresponding cubed SOG x_0 through the limits given by Montgomery et al. (2012)

$$x_0 \hat{\beta}_0 \mp t_{\alpha/2; n-2} \sqrt{\frac{\mathbf{e}_{LS}^\top \mathbf{e}_{LS}}{n-2} \left(1 + \frac{1}{n} + \frac{x_0^2}{\mathbf{x}^\top \mathbf{x}} \right)}, \quad (3.13)$$

where $t_{\alpha/2; n-2}$ is the percentile of a Student's distribution with $n - 2$ degrees of freedom. The LS-PLS method illustrated in Section 3.2 naturally lends itself to be used in reducing the LS prediction interval width defined by Equation (3.13) while maintaining the original relationship between the cubed SOG and the FCPH. Let us then define the normalized FCPH by subtracting the term $\hat{\mathbf{y}}_{PLS}$ calculated in Equation (3.10) from the observations \mathbf{y} as follows

$$\mathbf{y}_N = \mathbf{y} - \hat{\mathbf{y}}_{PLS}. \quad (3.14)$$

Note that $\hat{\mathbf{y}}_{PLS}$ in Equation (3.10) depends explicitly on $\hat{\boldsymbol{\gamma}}$ and \mathbf{Z}_{orth} , i.e. on the complete operational dataset used to build the model. Upon substituting \mathbf{y} with \mathbf{y}_N as the response

variable to plot against SOG observations, the LS-PLS normalized fuel-speed curve can finally be obtained. Figure 3.1 illustrates the graphical scheme of the proposed approach to develop the model.

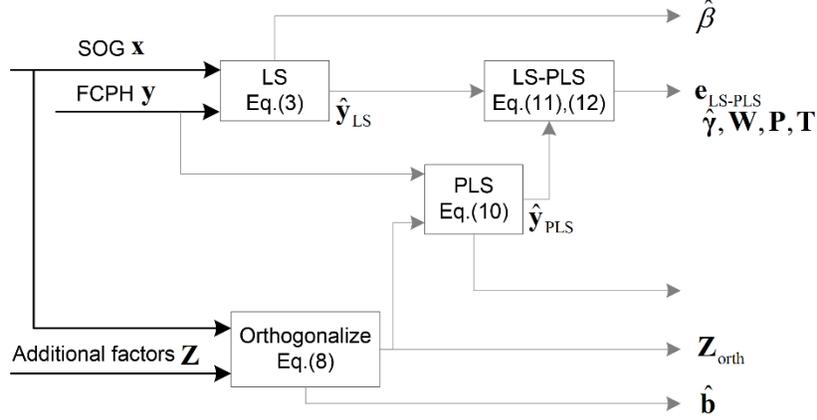


Figure 3.1: Graphical scheme of the proposed approach for model building. Bold lines refer to measured variables/factors.

Given a future observation of the cubed SOG x_0 and of the additional $m - 1$ factors arranged in the vector $\mathbf{z}_0 = (z_{10}, \dots, z_{(m-1)0})^\top$, the corresponding observation of the normalized FCPH, $y_{N,0}$, can be explicitly calculated through Equation (3.14) as

$$y_{N,0} = y_0 - \mathbf{z}_{0,orth}^\top \hat{\boldsymbol{\gamma}}, \quad (3.15)$$

where $\mathbf{z}_{0,orth} = \mathbf{z} - x_0 \hat{\boldsymbol{\gamma}}$. Based on a slightly different formula to the one given by Nomikos & MacGregor (1995a), the approximate limits of the $100(1 - \alpha)$ percent LS-PLS pointwise prediction interval can be calculated as

$$x_0 \hat{\beta}_0 \mp t_{\alpha/2; n-a-2} \sqrt{\frac{\mathbf{e}_{LS-PLS}^\top \mathbf{e}_{LS-PLS}}{n-a-2} \left(1 + \frac{1}{n} + \mathbf{t}_0^\top (\mathbf{T}^\top \mathbf{T})^{-1} \mathbf{t}_0 \right)}, \quad (3.16)$$

where $t_{\alpha/2; n-2}$ is the $100\alpha/2$ percentile of a Student's distribution with $n - a - 2$ degrees of freedom and

$$\mathbf{t}_0^\top = \mathbf{z}_{0,orth}^\top \mathbf{W} (\mathbf{P}^\top \mathbf{W})^{-1}, \quad (3.17)$$

as in Equation (3.9). Note that the limits in Equation (3.16) are only approximated because \mathbf{T} depends non-linearly on \mathbf{y} Nomikos & MacGregor (1995a), but allow to develop a simple expression. In the Appendix C, we carried out a simulation study to confirm the coverage of the prediction intervals in Equation (3.16). In order to explain the operative meaning of the FCPH normalization, note that regression of \mathbf{y}_M on \mathbf{x} has the amazing property of yielding the same coefficient $\hat{\beta}$ obtained through LS regression of \mathbf{y} on \mathbf{x} , because $\hat{\mathbf{y}}_{PLS}$ is, by construction, orthogonal to \mathbf{x} . Thus, the vector of fitted values $\hat{\mathbf{y}} = \mathbf{x} \hat{\beta}$ is equal to $\hat{\mathbf{y}}_{LS}$ in Equation (3.3). Therefore, the LS-PLS fuel-speed curve (in the normalized FCPH-SOG plane) coincides with the ordinary LS one (in the usual FCPH-SOG plane). Therefore, with abuse of terminology, we sometimes simply refer to fuel-speed curve. However, it is easy

to verify that normalized FCPH deviations $\mathbf{y}_N - \hat{\mathbf{y}}_N$ are equal to the LS-PLS residuals \mathbf{e}_{LS-PLS} of Equation (3.12). In other words, even if normalized FCPH does not represent actual FCPH, its deviation from the fuel-speed curve still represents deviation in terms of the actual FCPH (from the LS-PLS FCPH prediction $\hat{\mathbf{y}}_{LS-PLS}$). This allows monitoring FCPH through the LS-PLS pointwise prediction intervals of the normalized FCPH defined in Equation (3.16). Moreover, from Equation (3.15), note that the observation y_0 are equal to $y_{N,0}$ when $\mathbf{z}_0 = x_0 \hat{\boldsymbol{\delta}}$, i.e. the additional factors do not bring supplementary information, because each is equal to its expected value at the given SOG. Conversely, when the additional factors bring further information, $\hat{\mathbf{y}}_{PLS}$ contributes to reduce the LS-PLS residual sum of squares $\mathbf{e}_{LS-PLS}^\top \mathbf{e}_{LS-PLS}$ with respect to the LS residual $\mathbf{e}_{LS}^\top \mathbf{e}_{LS}$. Therefore, the higher the uncorrelated information brought by the additional factors for predicting the FCPH, the smaller the prediction interval width.

In short, let us note that using the LS-PLS method instead of the direct application of the PLS technique can be regarded as forcing the direction of the cubed SOG \mathbf{x} to be the first latent variable direction. This means that the first component was not selected by maximizing its covariance with \mathbf{y} . In general, this impacts on the number of columns of \mathbf{T} , which may indeed increase as the direction of the predictor variable \mathbf{x} (fitted through ordinary LS) deviates from the first component that would be obtained through the direct application of the PLS technique. However, the proposed approach mitigates the drawbacks of the classical fuel-speed curves and gives marine engineers the additional opportunity to use a familiar tool for monitoring purposes. In fact, the normalized FCPH observation for each voyage can be compared in the fuel-speed plot with the corresponding curve, as well as with the pointwise prediction interval (Equation (3.16)) at a given significance level. Points that fall outside the LS-PLS prediction interval warrant further investigation for fault diagnosis.

Strictly speaking, as with the canonical application of PLS methods (Nomikos & MacGregor 1995b, Kourti & MacGregor 1996), LS-PLS normalization of FCPH is allowed if and only if factors show no unusual variation during a given voyage. More precisely, the two following multivariate control charts have to be used to monitor the respective variations inside and outside the hyperplane defined by the model in the predictor variable space.

The first control chart is based on the T^2 statistic for a new observation

$$T^2 = \left(\frac{x_0^2}{\mathbf{x}^\top \mathbf{x} / (n-1)} + \mathbf{t}_0^\top \mathbf{S}^{-1} \mathbf{t}_0 \right) \frac{n(n-a)}{a(n^2-1)}, \quad (3.18)$$

where \mathbf{S} is the estimated covariance matrix of the scores \mathbf{T} introduced in Equation (3.9). Note that Equation (3.18) represents the LS-PLS modification of Hotelling's statistic for scores (Tracy et al. 1992, Nomikos & MacGregor 1995b) and involves, as expected, both the cubed SOG x_0 and the latent variable values \mathbf{t}_0 . The T^2 statistic follows an F -distribution with $a+1$ and $n-a-1$ degrees of freedom. Therefore, it can be used opportunely to define an upper control limit to test whether a new observation remains within the normal operating region in the projection space or not.

The second, namely the squared prediction error (SPE_X) control chart, detects the occurrence of any new event that causes the ship's operating conditions to move away from the latent variable space. A valid specification involves not only monitoring the scores of the additional factors from new observations, but also requires monitoring the residuals or the distance from the PLS model (Duchesne & MacGregor 2004). The SPE_X for the additional factors is defined for a new observation as

$$SPE_X = (\mathbf{z}_{0,orth} - \hat{\mathbf{z}}_{0,orth})^\top (\mathbf{z}_{0,orth} - \hat{\mathbf{z}}_{0,orth}), \quad (3.19)$$

3. LS-PLS APPROACH TO SHIP FUEL-SPEED CURVES

where $\hat{\mathbf{z}}_{0,orth} = \mathbf{P}\mathbf{t}_0$ is the value of the new observation for the additional $m - 1$ factors predicted by the latent variable model. In other words, it represents the squared perpendicular distance of an observation from the latent variable space and gives a measure of how close the observation is to it (Kourti & MacGregor 1996). Note that in contrast with T^2 , the statistic does not involve the cubed SOG, because there is no prediction error for the non-projected predictor variable. The upper control limit for this statistic is calculated assuming that it is approximately distributed as a weighted chi-squared (Jackson 2005, Nomikos & MacGregor 1995a). Following Jackson (2005), expression of the control limit with a significance level equal to is

$$Q_\alpha = \theta_1 \left(\frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} + 1 \right)^{1/h_0}, \quad (3.20)$$

where $\theta_i = \text{trace}(\mathbf{V}^i)$ is calculated from the covariance matrix \mathbf{V} of the residuals matrix $\mathbf{E} = \mathbf{Z}_{orth} - \mathbf{TP}^T$, $h_0 = 1 - 2\theta_1\theta_3/3\theta_2^2$, and z_α is the $100(1 - \alpha)$ percentile of the standard normal distribution. When an observation for T^2 or the SPE_X statistic falls outside of control limits, contribution plots can be drawn to detect, through visual inspection, the variable(s) with the highest deviation(s), which should be investigated. This relevant support to domain experts in fault diagnosis is unfeasible via the ordinary LS approach to fuel-speed curves. The LS-PLS method used for monitoring is summarized in the graphical scheme reported in Figure 3.2.

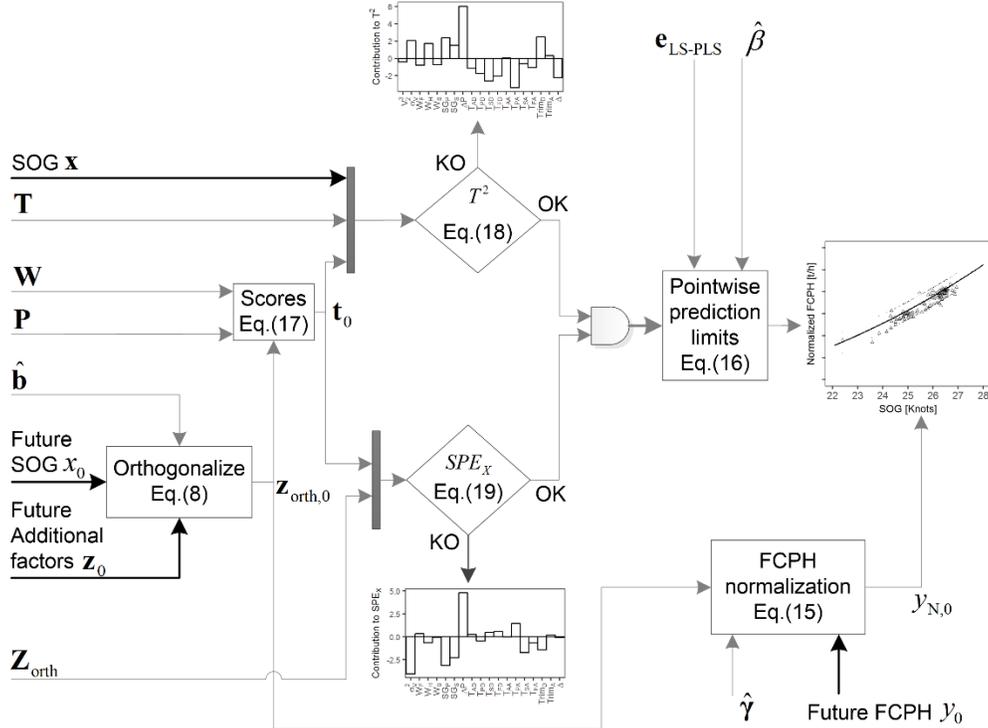


Figure 3.2: Graphical scheme of the proposed approach for monitoring FCPH. Bold lines refer to measured variables/factors.

Hypothesis testing on the effectiveness of energy efficiency initiatives

The LS-PLS fuel-speed curve gives a useful support when a benchmark comparison of data trends is needed before and after each specific EEI. From a formal viewpoint, let us consider the matrices \mathbf{x} and \mathbf{Z}_{orth} of the predictor variable observations corresponding to a proper reference period before EEI. We now combine them into the matrix $\mathbf{M}_1 = [\mathbf{x} \quad \mathbf{Z}_{orth}]$ and the regression coefficients into the vector $\boldsymbol{\theta}_1^\top = [\beta \quad \boldsymbol{\gamma}^\top]$. Then, the sample model corresponding to Equation (3.6) can be written into the more compact form

$$\mathbf{y}_1 = \mathbf{M}_1 \boldsymbol{\theta}_1 + \boldsymbol{\varepsilon}_1, \quad (3.21)$$

where \mathbf{y}_1 contains the FCPH observations before the EEI and $\boldsymbol{\varepsilon}_1$ contains the corresponding errors.

Similarly, let us consider only data after the EEI and arrange observations into matrices $\mathbf{M}_2 = [\mathbf{x}_2 \quad \mathbf{Z}_{orth,2}]$ and \mathbf{y}_2 . Then, the corresponding model is as follows

$$\mathbf{y}_2 = \mathbf{M}_2 \boldsymbol{\theta}_2 + \boldsymbol{\varepsilon}_2, \quad (3.22)$$

where $\boldsymbol{\theta}_2^\top = [\beta_2 \quad \boldsymbol{\gamma}_2^\top]$ is the new vector of regression coefficients and $\boldsymbol{\varepsilon}_2$ contains the corresponding errors.

Note that, by setting $\boldsymbol{\theta}_\Delta = \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1$, Equation (3.22) can be rearranged as follows

$$\mathbf{y}_2 = \mathbf{M}_2(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_\Delta) + \boldsymbol{\varepsilon}_2, \quad (3.23)$$

which means that

$$\mathbf{y}_2 - \mathbf{M}_2 \boldsymbol{\theta}_1 = \mathbf{M}_2 \boldsymbol{\theta}_\Delta + \boldsymbol{\varepsilon}_2, \quad (3.24)$$

The left-hand side represents the error $\boldsymbol{\varepsilon}_{21}$, obtained by using the model in Equation (3.21) estimated by reference data collected before EEI, with observations \mathbf{y}_2 and \mathbf{M}_2 collected after EEI. Thus, from Equation (3.24) a new model readily follows

$$\boldsymbol{\varepsilon}_{21} = \mathbf{M}_2 \boldsymbol{\theta}_\Delta + \boldsymbol{\varepsilon}_2. \quad (3.25)$$

From Equation (3.25), the effectiveness of a specific EEI can be tested by determining whether the vector $\mathbf{M}_2 \boldsymbol{\theta}_\Delta$ is zero, i.e. the model in Equation (3.25) does not account for a significant portion of the variation in the $\boldsymbol{\varepsilon}_{21}$ variable. The appropriate hypotheses are

$$\underbrace{H_0 : \mathbf{M}_2 \boldsymbol{\theta}_\Delta = 0}_{\text{effective EEI}}, \quad \underbrace{H_1 : \mathbf{M}_2 \boldsymbol{\theta}_\Delta \neq 0}_{\text{not effective EEI}} \quad (3.26)$$

These hypotheses relate to the significance of the model in Equation (3.25) and may be tested through analysis of variance (ANOVA), which is summarized in Table 3.1. If the null hypothesis holds, the ratio between the mean square due to regression (MSR) and the mean squared error (MSE) follows an F distribution with $a + 1$ and $n - a - 2$ degrees of freedom (Searle & Khuri 2017, Nomikos & MacGregor 1995a).

The proposed procedure can be trivially extended to different benchmark comparisons (e.g. between sister ships or between speed trials and operational data).

Quantification of fuel consumption and CO₂ emission reduction

If the null hypothesis is rejected, the model in Equation (3.21) can no longer be utilized for monitoring purposes. However, it can be useful for the quantification of savings in

Table 3.1: ANOVA table for hypothesis testing of the effectiveness of an EEI.

Source of variation	Degrees of freedom	Sum of squares	MSE	F-value
Regression	$a + 1$	$\hat{\mathbf{y}}_{LS-PLS}^\top \hat{\mathbf{y}}_{LS-PLS}$	$MSR = \frac{\hat{\mathbf{y}}_{LS-PLS}^\top \hat{\mathbf{y}}_{LS-PLS}}{a+1}$	$\frac{MSR}{MSE}$
Residuals	$n - a - 2$	$\mathbf{e}_{LS-PLS}^\top \mathbf{e}_{LS-PLS}$	$MSE = \frac{\mathbf{e}_{LS-PLS}^\top \mathbf{e}_{LS-PLS}}{n-a-2}$	

terms of fuel consumption or CO₂ emissions. In this regard, unfortunately, the international regulations (Milligan et al. 2006, European Commission 2015) are not rigorous in properly defining the periods before the EEI (to be used for curve estimation) and after the EEI (to be used for quantification of the fuel saving).

Regardless of this, let us consider for a given voyage j after the EEI, the FCPH y_j and its prediction \hat{y}_j based on the ship operating conditions x_j and $\mathbf{z}_{orth,j}^\top$, using the model developed before the EEI

$$\hat{y}_j = x_j \hat{\beta} + \mathbf{z}_{orth,j}^\top \hat{\gamma}. \quad (3.27)$$

The total fuel consumption reduction (FCR) can be calculated as

$$FCR = \sum_{j=1}^k (y_j - \hat{y}_j) h_j, \quad (3.28)$$

where $y_j - \hat{y}_j$ represents the saving achieved for the voyage j , h_j is the actual voyage sailing time, and k is the number of voyages considered after the EEI. Calculation of FCR also allows estimating CO₂ emission reduction (Yuan et al. 2016) through the emission factor for marine fuel oil, which is equal to 3.114 grams of CO₂ per gram fuel (Smith et al. 2015).

3.3 Case study: a Grimaldi Group Ro-Pax cruise ship

In the following treatment, real operational data from the Ro-Pax cruise ship mentioned in the introduction are applied in order to illustrate the proposed approach. Table 3.2 summarizes the variables chosen as response and first predictor, as well as the additional factors used in the LS-PLS model. Further details on these factors can be found in Bocchetti et al. (2015) Two applications are shown in the two following subsections. In particular, Section 3.3 shows the LS-PLS approach used for monitoring both predictor variable and additional factors (through the and control charts), and response variable (through the LS-PLS fuel speed curve) for new voyages. Section 3.3 evaluates the effectiveness of an EEI in terms of fuel consumption reduction.

Using the LS-PLS approach for monitoring new voyages

A Phase I sample of 59 reference observations is collected during a ten months' period, while Phase II monitoring is applied on 46 new voyages over the following six months. Figure 3.3 shows the PRESS statistic calculated for the PLS model, as result of the leave-one-out cross-validation procedure, which achieves its minimum when the number of components is nine.

Table 3.2: Operational variables measured for each voyage.

	Symbol	Description
Response variable	y	FCPH
Predictor variable	V^3	cubed SOG
Additional factors	σ_V^2	SOG variance
	W_H	mean head wind
	W_F	mean following wind
	W_S	mean side wind
	SG_P	mean port shaft generator power
	SG_S	mean starboard shaft generator power
	ΔP	mean power difference between two propeller shafts
	$Trim_D$	departure Trim
	$Trim_A$	arrival Trim
	T_{FD}	departure draught (fore perpendicular)
	T_{AD}	departure draught (aft perpendicular)
	T_{PD}	departure draught (midship section—port)
	T_{SD}	departure draught (midship section—starboard)
	T_{FA}	arrival draught (fore perpendicular)
	T_{AA}	arrival draught (aft perpendicular)
	T_{PA}	arrival draught (midship section—port)
	T_{SA}	arrival draught (midship section—starboard)
	Δ	displacement

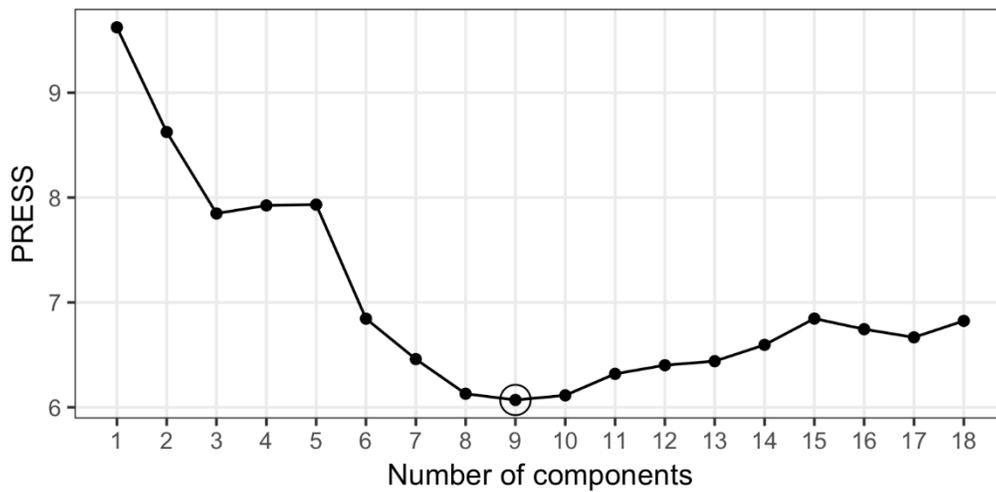


Figure 3.3: The PRESS statistic used to determine the number of components to retain in the PLS model.

Estimating the LS-PLS fuel-speed curve

The LS and LS-PLS fuel-speed curves are reported in Figure 3.4, together with the reference observations of FCPH and normalized FCPH (obtained by Equation (3.14)), used for their construction, respectively. For confidentiality reasons, y-axis tick labels are hidden, but the vertical scale is left unchanged. As already indicated in Section 2.2, note that the LS-PLS fuel-speed curve is identical to the ordinary LS one, but the points in Figure 3.4a are much more scattered than the points in Figure 3.4b, because the LS residual sum of squares is larger than the LS-PLS residual.

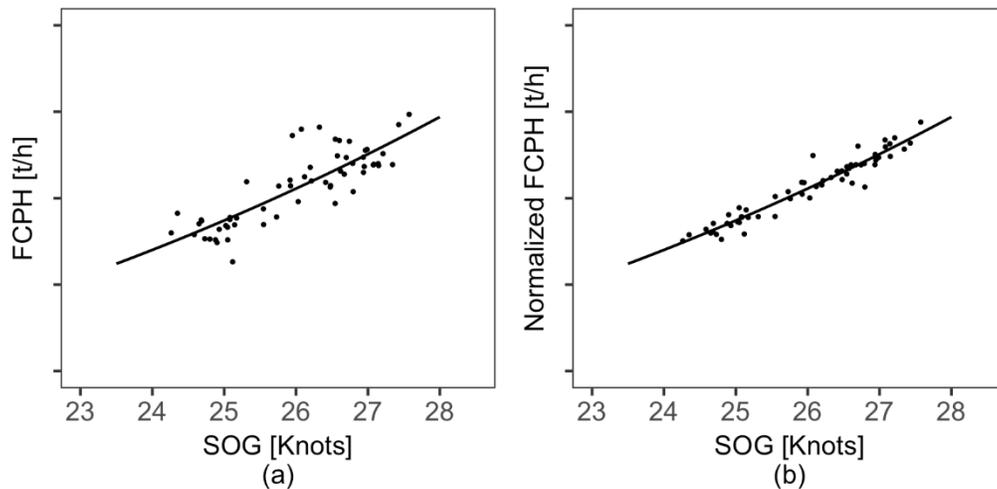


Figure 3.4: (a) FCPH and (b) normalized FCPH reference observations (points) used for the construction of (a) ordinary LS and (b) LS-PLS fuel-speed curves. Vertical scale unit is 1 t/h .

Monitoring the predictor variables

As described in Section 3.2, fuel consumption Phase II monitoring through LS-PLS fuel-speed curve requires and (defined in Equation (3.18) and (3.19)) to be in control. For this reason, Figure 3.5 and Figure 3.6 report T^2 and SPE_X control charts, respectively, together with the corresponding upper 99% limits calculated based on the Phase I sample. Voyages are identified by the progressive voyage number (VN) and are analysed in chronological order below. In those figures, the labelled voyages are outside the control limits and thus may require further technological investigation on the SOG or the additional factors (Table 3.2) for the diagnosis of unusual variations.

In this case, there is a group of three consecutive voyages, VN 18, 19, and 20 that are out of control in both the T^2 and SPE_X charts. By exploring the corresponding contribution plots, which are reported in Figure 3.7 and Figure 3.8, respectively, the power difference between propeller shafts (ΔP) appear as the major variable contributing to both the T^2 and SPE_X statistics. The contribution to the T^2 statistic also identifies as anomalous the cubed SOG (V^3) and the arrival trim ($Trim_A$).

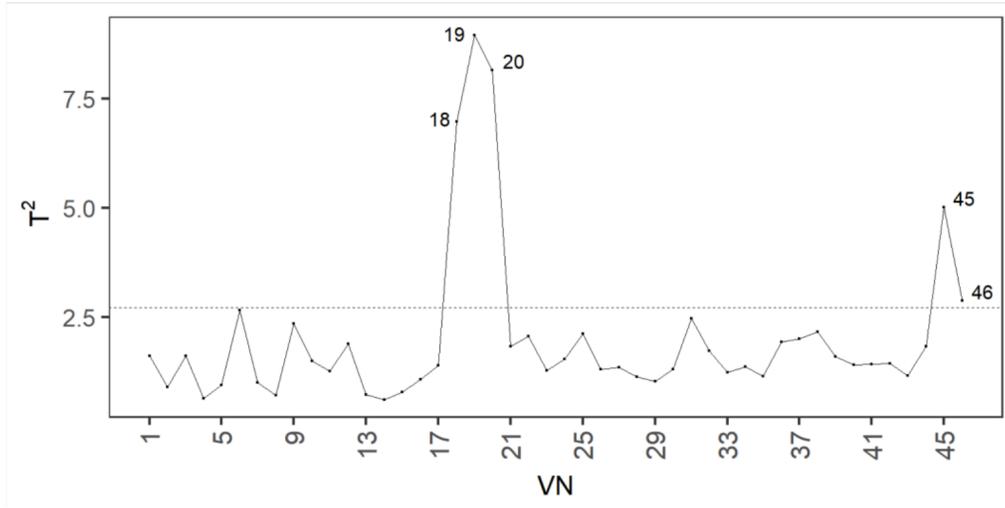


Figure 3.5: T^2 control chart for monitoring voyages with 99% control limit (dashed line).

The VN 43 is out of control only in the SPE_X control chart (Figure 3.6) that is again plausibly due to the power difference between propeller shafts (ΔP) as indicated by the corresponding contribution plot (Figure 3.7).

Unfortunately, the consecutive VN 45 and 46, which are out of control in the T^2 control chart (Figure 3.5), do not present dominant factors as almost all show high contributions (Figure 3.7) to T^2 .

As already noted in Section 3.2, the predictor variable V^3 does not contribute to the SPE_X statistic because it is not involved in PLS latent model building; thus it does not appear in contribution plots in Figure 3.8.

Monitoring the response variable through the LS-PLS fuel-speed curve

For monitoring the FCPH of those voyages that do not fall outside the limits of either the T^2 or SPE_X control charts (Figure 3.5 and Figure 3.6), the fuel-speed curves (Figure 3.4) and LS and LS-PLS pointwise 95% prediction intervals based on Phase I sample are obtained through Equation (3.13) and (3.16) and are reported in Figure 3.9a and Figure 3.9b, respectively.

We note again that the LS-PLS pointwise prediction intervals of the FCPH depend not only on SOG, but also on the additional factors. Then, in 3.9b it is not surprising that voyages with the same SOG do not (necessarily) have the same prediction interval width.

By comparing 3.9a and 3.9b it is, however, evident that the LS-PLS method achieved, as expected, narrower prediction intervals. In particular, the 95% prediction intervals obtained through the ordinary LS method (3.9a) are at least 78% wider than LS-PLS pointwise ones (3.9b) and are thus useless for managerial decision-making. In those voyages with FCPH outside the limits of the 95% pointwise prediction intervals, a fault may have occurred caused by factors external to those considered in Table 3.2.

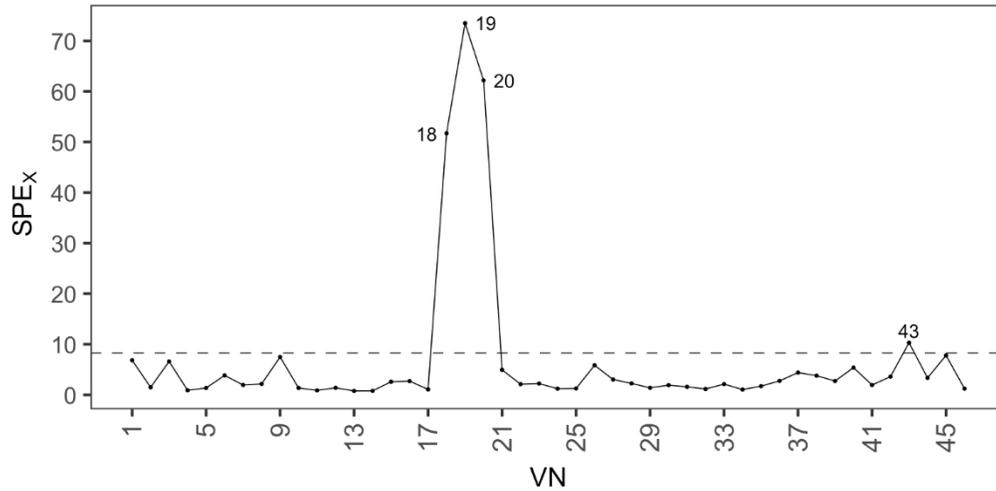


Figure 3.6: SPE_X control chart for monitoring voyages with 99% control limit (dashed line).

Using the LS-PLS approach for testing the effectiveness of an EEI

The proposed approach lends itself to be used in testing the effectiveness of a specific EEI, through a benchmark comparison of FCPH before and after the EEI, which, for the considered ship, consisted in a silicone foul release coating of the hull. In order to do this, two different groups of data are considered. The first group of data is made of 66 observations available before the EEI, whereas the second group of data is made of 61 observations collected after the EEI. Both the fuel speed curves that are plotted in Figure 3.10 are estimated on the basis of the 66 observations of the first group. Then, FCPH and normalized FCPH observations of the second group are reported together with the corresponding LS and LS-PLS pointwise 95% prediction intervals in Figure 3.10a and Figure 3.10b, respectively. In particular, the latter is based on five PLS components as resulted from the leave-one-out cross-validation procedure. The advantage of the proposed LS-PLS approach is in correctly identifying a larger number of points that fall under the lower prediction limit at the same level (95%) as graphically confirmed by Figure 3.10. In more formal terms, the effectiveness of the considered EEI is also tested as described in Section 3.2 and summarized in the ANOVA table reported in Table 3.3.

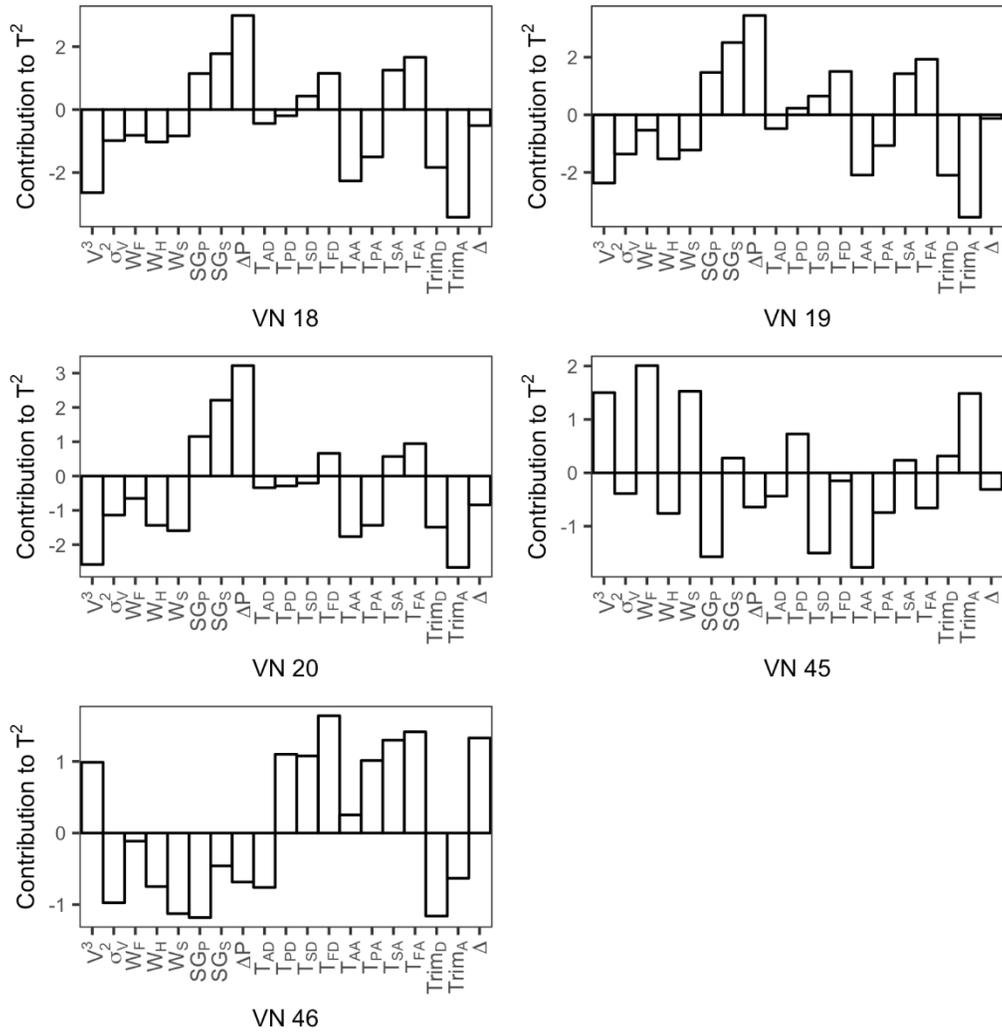


Figure 3.7: Contribution to T^2 for VN 18, 19, 20, 45, and 46.

3. LS-PLS APPROACH TO SHIP FUEL-SPEED CURVES

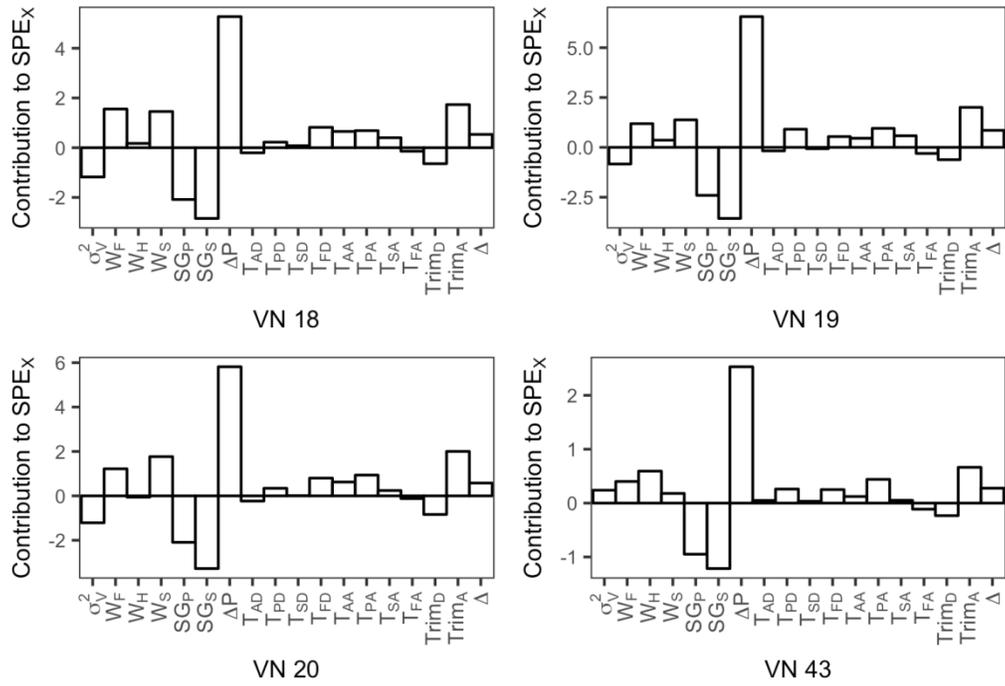


Figure 3.8: Contribution to SPE_X for VN 18, 19, 20 e 43.

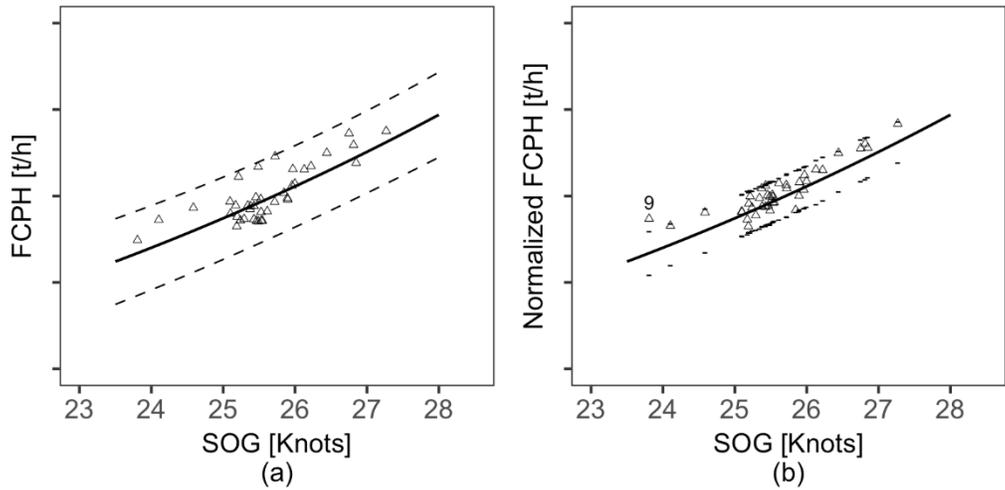


Figure 3.9: Monitoring of voyages (triangles) through (a) ordinary LS fuel-speed and (b) LS-PLS fuel-speed curve. Vertical scale unit 1 t/h.

Table 3.3: ANOVA table for hypothesis testing of the effectiveness of an EEI.

Source of variation	Degrees of freedom	Sum of squares	MSE	F -value	p -value
Regression	6	1.030	0.172	18.92	<0.001
Residuals	54	0.490	0.009		

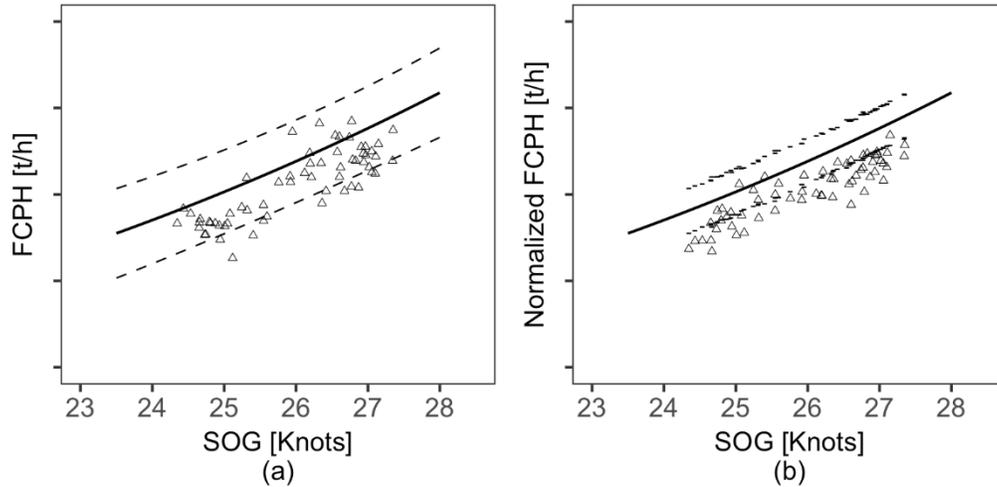


Figure 3.10: Benchmark comparison of 61 new voyages (triangles) after the EEI via (a) 95% LS prediction intervals and (b) 95% LS-PLS pointwise prediction intervals of the corresponding fuel-speed curve obtained based on 66 reference observations before the considered EEI. Vertical scale unit 1 t/h .

The null hypothesis of not effective EEI is rejected at a significance level $\alpha = 0.01$. This allows explicitly stating the effectiveness of the considered EEI and thus, estimating fuel consumption and CO_2 emission reduction as explained in Section 3.2. Through Equation (3.28), the FCR achieved by the considered 61 voyages after EEI is equal to 96.9 t of heavy fuel oil. By the conversion coefficient for this fuel type, this corresponds to a CO_2 emission reduction of 302 t .

Then, the rejection of the null hypothesis allows the mean FCR for a future voyage conditional to a given SOG to be visualized through the difference, at the given SOG, between the fuel-speed curve depicted in Figure 3.9b, depicted in Figure 3.11, that are constructed based on the observations collected before (thin line) and after (bold line) the EEI.

3.4 Conclusions

Fuel-speed curves are very attractive for shipping practitioners, who are used to dealing with classical curves when estimating and testing for fuel efficiency improvement. Motivated by real operational data automatically acquired from a Ro-Pax cruise ship operating in the

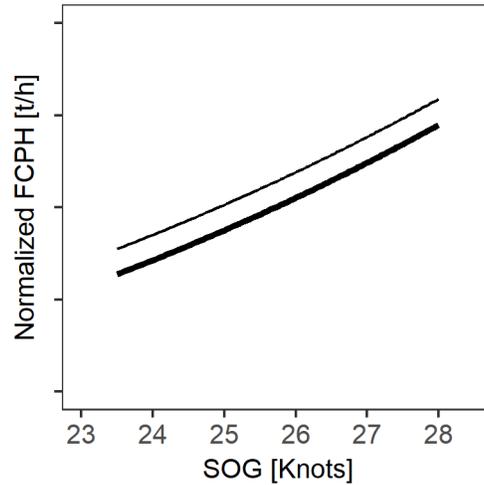


Figure 3.11: Comparison between fuel-speed curves estimated based on reference observations collected before (thin line) and after (bold line) the EEL. Vertical scale unit 1 t/h .

Mediterranean Sea, a novel approach has been proposed by elaborating the orthogonal LS-PLS method. This approach allows estimating fuel-speed curves that exploit the contribution of orthogonal components constructed as linear combinations of additional factors collected on-board by modern multi-sensor systems. The LS-PLS normalized fuel-speed curve is shown to possess the highly useful property of maintaining the theoretical relationship between fuel consumption per hour and mean hourly vessel's speed over ground, as well as achieving narrower prediction intervals for future observations than those obtained through the ordinary LS method. This makes the fuel-speed curve suitable for monitoring fuel consumption per hour in real cases.

Accordingly, the LS-PLS modification of Hotelling's control chart for scores and the classical squared prediction error control chart are used to detect whether or not future observations move away either inside or outside the normal operating region of the latent variables, respectively. Voyages with unusual variations have been further technologically investigated through contribution plots of the corresponding control chart in order to support the diagnosis of fault causes. This support would have been unfeasible via the ordinary LS approach to fuel-speed curves.

Prediction intervals for future observations are also utilized to compare graphically fuel consumption before and after energy efficiency initiatives. Moreover, a specific hypothesis test is implemented and allows for the estimation of any fuel consumption reduction due to a specific energy efficiency initiative. In particular, for the Ro-Pax cruise ship considered in the case study, fuel consumption and CO_2 emission reduction were estimated after silicone foul release coating of the hull. These results entitled the shipping company to directly cash in the carbon credits, according to the international emission trading mechanism set by the Kyoto Protocol.

Lastly, the proposed approach deserves the attention of ship owners, managers, and

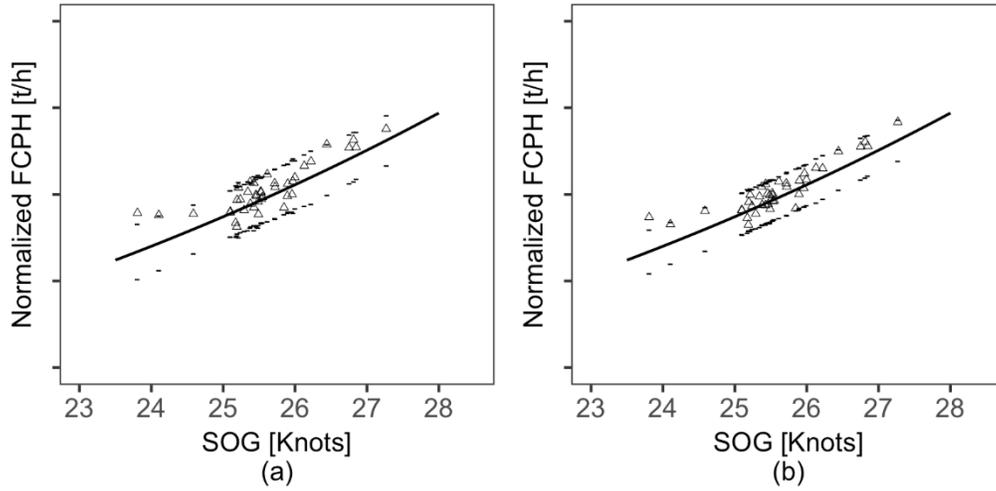


Figure 3.12: Fuel-speed curve (solid line) and 95% pointwise prediction interval of the normalized FCPH (dashes) estimated by (a) MLR and (b) PLS methods for voyages monitored in Phase II (triangles).

operators who are called to make decisions based on massive amounts of operational data on their fleet.

3.5 Appendix A: predictive ability comparison of PLS and MLR

The predictive ability of the LS-PLS approach can be validated by comparing the PLS method that enters into force in Equation (3.8) to decompose \mathbf{Z}_{orth} with respect to the use of the standard MLR. In general, the two methods are equivalent when the number of components used in the PLS method equals the number of additional factors. Let us consider the same Phase I sample used in Section 3.3, as well as the same voyages monitored in Phase II. The PRESS statistic has already been reported in Figure 3.3 in the case of the PLS method. The minimum value (6.07) is achieved at nine components and therefore, is smaller than the PRESS achieved by the MLR (6.82). This shows that the PLS method has better than or at least the same predictive ability of MLR. This is also validated on the Phase II by looking at Figure (3.12) that shows 95% pointwise prediction intervals obtained by MLR (Figure (3.12)a) are wider than those obtained by PLS (Figure (3.12)b). Accordingly, the sum of squared prediction errors achieved by MLR (1.09) is larger than that achieved by PLS (0.80).

3.6 Appendix B: main steps of the PLS algorithm

In this Appendix, the main steps of the PLS algorithm are outlined. Suppose you have a $n \times 1$ vector \mathbf{y} representing a scalar response variable and a $n \times m$ matrix \mathbf{X} representing the predictors. The linear relationship between \mathbf{y} and \mathbf{X} , which is $\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a vector of errors. Suppose that the variables are mean centered. In this work, the nonlinear

iterative partial least-squares (NIPALS) algorithm for the linear regression of \mathbf{y} on \mathbf{X} is used. The algorithm is initialized by setting $\mathbf{X}_0 = \mathbf{X}$ and $\mathbf{y}_0 = \mathbf{y}$. Then, at each iteration $i = 1, \dots, a$:

1. calculate the weight vector $\mathbf{w}_i = \mathbf{X}_{i-1}^\top \mathbf{y}_{i-1} / (\mathbf{y}_{i-1}^\top \mathbf{y}_{i-1})$;
2. scale \mathbf{w}_i to be unit length;
3. calculate the score vector as linear combination of columns of \mathbf{X}_{i-1} as $\mathbf{t}_i = \mathbf{X}_{i-1} \mathbf{w}_i$;
4. regress \mathbf{y}_{i-1} on \mathbf{t}_i obtaining the coefficient $b_i = \mathbf{y}_{i-1}^\top \mathbf{t}_i / (\mathbf{t}_i^\top \mathbf{t}_i)$;
5. calculate the loading vector $\mathbf{p}_i = \mathbf{X}_{i-1}^\top \mathbf{t}_i / (\mathbf{t}_i^\top \mathbf{t}_i)$;
6. deflate the matrices $\mathbf{X}_i = \mathbf{X}_{i-1} - \mathbf{t}_i \mathbf{p}_i^\top$ and $\mathbf{y}_i = \mathbf{y}_{i-1} - \mathbf{t}_i b_i$ and start from the first step to find a new component.

At the end of the algorithm, define the matrices $\mathbf{T} = (\mathbf{t}_1, \dots, \mathbf{t}_a)$, $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_a)$, $\mathbf{P} = (\mathbf{p}_1, \dots, \mathbf{p}_a)$, and the vectors $\hat{\boldsymbol{\gamma}}_{PLS} = (b_1, \dots, b_a)^\top$. The vector $\hat{\boldsymbol{\gamma}}_{PLS}$ allows to make predictions on the response variable, in fact $\hat{\mathbf{y}} = \mathbf{T} \hat{\boldsymbol{\gamma}}_{PLS}$. However, we are interested in expressing the prediction of the response variable as a linear combination of columns of \mathbf{X} , i.e. $\hat{\mathbf{y}} = \mathbf{X} \hat{\boldsymbol{\gamma}}$. Note that, since $\mathbf{T} = \mathbf{X} \mathbf{W} (\mathbf{P}^\top \mathbf{W})^{-1}$, we can write $\hat{\mathbf{y}} = \mathbf{T} \hat{\boldsymbol{\gamma}}_{PLS} = \mathbf{X} \mathbf{W} (\mathbf{P}^\top \mathbf{W})^{-1} \hat{\boldsymbol{\gamma}}_{PLS}$. Therefore, we have $\hat{\boldsymbol{\gamma}} = \mathbf{W} (\mathbf{P}^\top \mathbf{W})^{-1} \hat{\boldsymbol{\gamma}}_{PLS}$.

3.7 Appendix C: simulation study on the coverage of the approximate pointwise prediction intervals

This Appendix deals with a simulation study on the coverage of the approximate pointwise prediction intervals for the proposed LS-PLS normalized fuel-speed curve, whose limits are reported in Equation (3.16). To this aim, we firstly consider Equation (3.5) to introduce the relationship between (observations of) the predictor variable X and the vector of the $m - 1$ additional factors Z .

The observations \mathbf{x} of X are generated as pseudo-random observations from the standard normal distribution. The number of additional factors is set equal to $m - 1 = 9$ and the matrix \mathbf{Z} is obtained according to Equation (3.5), with $\boldsymbol{\delta} = (1, \dots, 1)^\top$. The matrix \mathbf{Z}_{orth} is then generated in order to introduce multicollinearity, as in (Gunst & Mason 1977). In particular, the first five columns $\mathbf{z}_{orth,1}, \dots, \mathbf{z}_{orth,5}$ of \mathbf{Z}_{orth} are generated as independent observations from the standard normal distribution; whereas each of the remaining columns is obtained as a nearly linear combination of the first five ones, i.e.,

$$\mathbf{z}_{orth,j} = \frac{1}{5} \mathbf{z}_{orth,1} + \dots + \frac{1}{5} \mathbf{z}_{orth,5} + \boldsymbol{\varepsilon}_{\mathbf{z}_{orth,j}}, \quad (3.29)$$

where $\boldsymbol{\varepsilon}_{\mathbf{z}_{orth,j}}$ is a vector of independent observations normally distributed with zero mean and standard deviation equal to 1. Observations of the response variable are generated according to Equation (3.7), with $\beta = 1$ and $\boldsymbol{\gamma} = (1, \dots, 1)^\top$, whereas the errors $\boldsymbol{\varepsilon}$ are independently generated from the standard normal distribution.

According to the proposed approach, four sets of reference observations, with different sample size $n = 30, 60, 200, 1000$ have been generated. For each set of reference observations,

1. the number of PLS components are chosen by leave-one-out cross-validation;

3.7. Appendix C: simulation study on the coverage of the approximate pointwise prediction intervals

Table 3.4: Mean, median, and 95% confidence interval of the approximate prediction interval coverage.

Sample size n	mean	median	95%-confidence interval
30	0.908	0.920	0.800–0.978
60	0.929	0.933	0.875–0.972
200	0.944	0.946	0.918–0.966
1000	0.949	0.949	0.935–0.963

2. a new independent set of 1000 observations has been further generated;
3. model parameters are estimated through the proposed orthogonal LS-PLS approach;
4. the prediction interval coverage is estimated as the proportion of new observations of the response variable within the corresponding prediction interval.

This procedure has been repeated 1000 times to get the empirical distribution of the approximate pointwise prediction interval coverage. The main results are summarized in Table 3.4. In particular, for each sample size, Table 3.4 reports the sample mean, the sample median as well as the corresponding 95%-confidence interval of the approximate prediction interval coverage, calculated as the 5th and 95th sample percentiles from the empirical distribution of the approximate prediction interval coverage. The smaller the sample size, the larger the 95% -confidence interval, the more the distribution of the coverage estimates is left-skewed (i.e., the mean is smaller than the median). Indeed, this is quite reasonable since the coverage is constrained in $(0, 1)$ and its mean and median are close to one. Accordingly, as the sample size increases, prediction intervals become much narrower and symmetric, and the sample mean and median get very close to 0.95. This clearly shows that the larger the sample size, the more precise the coverage of the approximate prediction interval.

Chapter 4

Ship fuel consumption monitoring and fault detection via control charts based on partial least squares regression

New regulations in the shipping sector aim to give greater transparency to operations and public access to CO₂ emissions data. EU regulation 2015/757 became mandatory in January 2018 and urges shipping companies to set up systems for daily monitoring, reporting and verification (MRV) of emissions for individual ships. Manual acquisition and handling of emissions data may be allowed (e.g. bunker fuel delivery note, bunker fuel tank monitoring), but is adversely affected by uncertainty due to human intervention and will eventually be unusable for monitoring purposes. However, the massive amounts of navigational data acquired by multi-sensor systems installed on-board modern ships have great potential to aid compliance with regulations but their use is hampered by the lack of effective analytical methods in maritime literature. This work demonstrates a statistical framework and automatic reporting system for fuel consumption monitoring that addresses the MRV requirements needed to comply with the regulations. The framework has been applied to the Grimaldi Group's Ro-Ro Pax cruise ships and is shown, in addition, to be capable of supporting fault detection as well as verifying CO₂ savings achieved after energy efficiency initiatives.

4.1 Introduction

Regulatory background

In recent years, the increase of greenhouse gas (GHG) emissions such as carbon dioxide (CO₂) has determined global warming and climate change. They are considered one of the biggest challenges of our time and prompt solutions have to be adopted to avoid severe consequences for society. According to the Kyoto Protocol (Ki-Moon 2008), several institutions have focused their attention on this problem. The International Maritime Organization (IMO), through the Maritime Environmental Protection Committee (MEPC), has developed initiatives to reduce GHGs from ships, urging shipping companies to adopt a set of technical and operational measures to improve energy efficiency of ships not only during operation but also in the design phase (IMO 2009, 2012*b,d*). The Energy Efficiency Design Index (EEDI) (IMO 2012*d*), and the Ship Energy Efficiency Management Plan (SEEMP) (IMO 2012*b*), are the tools used

to monitor these improvements. Alongside IMO guidelines, the Regulation EU 2015/757 of the European Parliament forces shipping companies, operating with their fleet in the EEA (European Economic Area) regardless of the flag state or port registry, to monitor and report all harmful emissions (European Commission 2015) from 1 January 2018.

The basic measure adopted by shipping companies to reduce CO₂ emissions and fuel consumption is the sailing speed reduction, since it does not require any energy efficiency operation that has a cost. In fact, the ship's speed has also implications in terms of energy efficiency and costs for owners and ship operators. Small changes in speed can significantly improve the energy efficiency as well as the productivity and revenue of the ship (Smith et al. 2015). The problem of optimizing the ship speed on a route in order to minimize the total fuel cost while satisfying the calling time window constraints at the calling ports has been faced by Kim et al. (2014, 2016). Because of the well-known difficulties in directly measuring CO₂ emissions, the MRV regulation (European Commission 2015) provides for indirect monitoring through the ship's fuel consumption. The calculation of CO₂ emissions can, in fact, be retrieved on the basis of the amount of fuel consumption, through the emission factor in accordance to Annex VI (European Commission 2012).

Literature review

In the naval literature, the most common method used to estimate fuel consumption and then CO₂ emissions is the so-called speed-power curve. This curve is drawn by exploiting the univariate relationship between the engine power and the vessel speed (Lewis 1988). However, despite its intuitive usage, this method is affected by large variability and may lead to poor predictions due to different sailing (e.g. trim, displacement, etc.) and weather conditions.

Several methods have been proposed to improve the estimate of speed-power curves, by exploiting any information from additional operational variables. In particular, Bialystocki & Konovessis (2016) firstly consider ship's draught and displacement, weather force and direction, hull and propeller roughness. Perera & Mo (2016) draw the empirical relationship between fuel consumption and the main operational variables using a graphical data analysis of performance and navigation parameters to support the management in analyzing the energy flow path. Petersen et al. (2012) proposed a statistical method based on artificial neural networks and Gaussian Processes to predict fuel consumption through sailing and environmental conditions such as vessel's speed over ground, vessel's speed through water, trim, displacement, wind force and directions. Lu et al. (2013) estimated the ship resistance considering as operational variables the ship type, draughts, speeds, encounter angles, sea states, fouling effect and engine degradation conditions. Meng et al. (2016) proposed a method based on regression models to estimate the fuel consumption rate of container ships. Trodden et al. (2015) analyzed fuel consumption and speed over the ground using a continuous data stream from a tug boat. Murphy et al. (2012) used fuel consumption data and engine load from sea trials to investigate reduction in fuel consumption. Zaman et al. (2017) presented a statistical analysis to automatically detect the vessel operational modes (port, manoeuvring, sailing) based on sensor data acquired on board as a pre-cursor to modelling fuel consumption in transit mode.

Bocchetti et al. (2013, 2015) proposed a statistical method based on multiple linear regression to predict ship fuel consumption and build prediction intervals for each voyage considering the operational and sailing conditions including ship speed, sailed distance, wind speed, wind direction, cumulative docking time, displacement, stabilizer fin operating time,

and engine operation mode. In particular, this method exploits the massive amount of sensor data acquired on board of two Ro-Ro Pax ships that link three ports in the Mediterranean Sea.

The proposed approach

Unfortunately, most of the presented methods have strong limitations when applied to high-dimensional and correlated data, or they do not fully exploit all of the available information. New data acquisition technologies in fact have brought massive navigation data and a call for shipping management to adopt new methodologies to fully exploit them. For this purpose, an engineering approach based on partial least squares (PLS) regression is introduced to develop a model for ship fuel consumption prediction and monitoring based on the massive navigation data automatically acquired on-board of the modern ships. In particular, the proposed model is based on summary statistics of each voyage deduced from the sensor signals that relate to the actual navigation time and therefore, neglects manoeuvre time and stay in port at departure and at arrival. The fuel consumed in this phase is known to contribute more than the 90 percent to the total consumption.

The aim of the proposed approach is twofold: *(i)* statistical monitoring of ship fuel consumption (and thus CO₂ emissions) to support shipping management to identify anomalies and *(ii)* quantifying fuel consumption reduction (FCR) consequent to energy efficiency initiatives (EEIs) (hereinafter referred to also as *dry-dock* operations). The rest of the chapter is organized as follows. A brief introduction on the acquisition data system and the variables used in this procedure is given in Section 4.2. Section 4.3 details the statistical approach and the monitoring tools utilized. Lastly, in Section 4.4 a real-case study is presented to illustrate the applicability and the effectiveness of the proposed approach.

4.2 Data

The navigation data are collected on board of four Ro-Ro Pax ships owned by Grimaldi Group that operate in the Mediterranean Sea. Ships, port names and dates are intentionally omitted for confidentiality reasons as well as actual numeric values but axis scales in the figures are left unchanged. However, even if we have currently been implementing the model on all the four Ro-Ro Pax ships mentioned above, for reasons of brevity, in what follows we show only the most relevant results achieved for two ships (hereinafter referred to as *Ship 1* and *Ship 2*). Therefore, we used two different ships to present different scenarios in which the proposed approach can be useful. A first scenario is meant to show the monitoring performance of the proposed approach, while in the second case the ship under the energy efficiency initiative did not show out-of-control voyages in the considered period.

Each ship is equipped with a sensor network and a data acquisition device (DAQ). Every five minutes the DAQ device collects the values of a set of physical variables used as predictors, the complete set of physical variables of engineering interest used as predictor variables to monitor the ship operating conditions, ship fuel consumption, and CO₂ emissions is illustrated in Table 4.1. Further information on predictor variables and their descriptions are given in the following sub-section.

4. SHIP FUEL CONSUMPTION MONITORING AND FAULT DETECTION VIA CONTROL CHARTS
BASED ON PARTIAL LEAST SQUARES REGRESSION

Table 4.1: Physical variables acquired at each voyage considered in the proposed approach.

	Symbol	Description
Response variable	y	Average fuel consumption per hour [Mt/h]
Predictor variables	V^3	cubed SOG [kn^3]
	σ_V^2	SOG variance [kn^2]
	W_H	head wind [kn]
	W_F	following wind [kn]
	W_S	side wind [kn]
	SG_P	port shaft generator power [kW]
	SG_S	starboard shaft generator power [kW]
	ΔP	power difference between two propeller shafts [kW]
	ΔSG	power difference between two shafts generators [kW]
	T_{FD}	departure draught (fore perpendicular) [m]
	T_{AD}	departure draught (aft perpendicular) [m]
	T_{PD}	departure draught (midship section—port) [m]
	T_{SD}	departure draught (midship section—starboard) [m]
	T_{FA}	arrival draught (fore perpendicular) [m]
	T_{AA}	arrival draught (aft perpendicular) [m]
	T_{PA}	arrival draught (midship section—port) [m]
	T_{SA}	arrival draught (midship section—starboard) [m]
	$Trim_D$	departure Trim [m]
	$Trim_A$	arrival Trim [m]
Δ	displacement [t]	
Other variables	UTC	Date and time UTC
	h	SOG variance [kn^2]
	P^T	shaft propeller power [kW]
	P^E	electrical power [kW]
	M	Distance travelled [NM]
	h_{ij}	Running hours (i -th engine, j -th engine set) [h]
	$SFOC_{ij}$	Specific fuel oil consumption (i -th engine, j -th engine set)
	P_{ij}	Power (i -th engine, j -th engine set) [kW]
	η_j^m	Gearbox mechanical efficiency (j -th engine set)
	η_j^e	Shaft generator electrical efficiency (j -th engine set)

Variable Definition

As previously stated, each variable summary statistic is available only at the end of the actual navigation time (h), that is defined as the time interval between the *Finished with Engine* order at departure port and *Stand by Engine* order at arrival port (IMO 2000). The actual navigation time is calculated on the data and time UTC Coordinated Universal Time acquired through the Saab R4 GPS Navigation Sensor.

The Speed Over Ground (V) and the variance of SOG (σ_V^2) represent respectively the mean value and the variance value of the averages obtained from the 5 minutes interval observations. In particular, the SOG is obtained as

$$V = \frac{M}{h} \quad (4.1)$$

where h is the actual navigation time and M is the distance travelled, calculated from the latitude and longitude data collected by the Saab R4 GPS Navigation Sensor. The distance travelled is calculated through the haversine formula according to (Veness 2007) as

$$M = r\theta \quad (4.2)$$

where r is the radius of the Earth, i.e. 6378.14 km and θ is defined as an angular distance in radians. The angle θ is calculated via the haversine formula as follows

$$\text{hav}\theta = \text{hav}\Delta\varphi + \cos\varphi_A \cos\varphi_B \text{hav}\Delta L, \quad (4.3)$$

where $\Delta\varphi$ is the difference in latitudes, ΔL is the difference in longitudes, while φ_A and φ_B are the latitude values at point A and point B, respectively, and $\text{hav}\theta = \sin^2(\theta/2)$ is the haversine function.

Specific resistance tests carried out in a towing tank within a speed range of 20 to 25 kn show that, for the considered ship type, the hydrodynamic resistance is proportional to the third power of SOG. Therefore, the proposed approach adopts a SOG that is cubed in order to obtain V^3 .

The wind speed and direction are considered through the head (W_H), side (W_S) and following (W_F) wind components. According to (ITTC 2017), Figure 4.1 shows the sign convention.

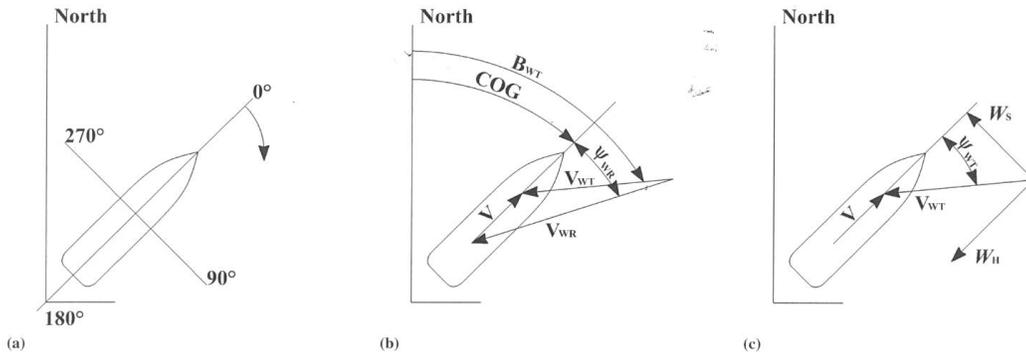


Figure 4.1: Wind components.

The head wind component is calculated as

$$\tilde{W}_H = \begin{cases} 0 & \text{if } 90^\circ \leq \Psi_{WT} \leq 270^\circ \\ V_{WT} \cos(\Psi_{WT}) & \text{otherwise} \end{cases}, \quad (4.4)$$

where the V_{WT} is the true wind speed and Ψ_{WT} the difference between the true wind angle in earth system (B_{WT}) and course over ground (COG) averaged by the DAQ device every 5 minutes. V_{WT} and B_{WT} are automatically calculated by a Thies Clime anemometer (Adolf Thies GmbH & Co. KG, Gottingen, Deutschland) based on the sensor measurements of the relative wind speed V_{WT} and direction Ψ_{WT} , as well as COG and SOG data. The wind speed sensor has an accuracy of $\pm 2.5\%$ and a resolution of 0.05 m, while the wind direction sensor has an accuracy of $\pm 2.5\%$ and a resolution of 2.5° . The side wind component W_S is defined once a voyage is completed as the mean of

$$\tilde{W}_S = |V_{WT} \sin(\Psi_{WT})|, \quad (4.5)$$

while the following wind component W_F is defined as the mean value of

$$\tilde{W}_F = \begin{cases} -V_{WT} \cos(\Psi_{WT}) & \text{if } 90^\circ \leq \Psi_{WT} \leq 270^\circ \\ 0 & \text{otherwise} \end{cases}. \quad (4.6)$$

The engine variables are port and starboard shaft generator power variables SG_P and SG_S and they allow taking into account the different modes of navigation (constant and combinator mode).

The Power difference between two propeller shafts ΔP allows discovering anomalies or malfunctions in the main engines, while the Power difference between two shaft generator powers ΔSG allows discovering if one of the two shaft generators is out of order.

The Displacement variable Δ is defined as the mean value of the two displacements obtained, respectively, at departure and arrival port. In particular, each of these two displacement values are derived from the hydrostatic data based on draughts at amidships and trim. The former is obtained by averaging the portboard and the starboard draught in the amidships section at the departure and arrival port, i.e. T_{PD} , T_{SD} , T_{PA} , and T_{SA} ; the latter in accordance with ITTC (2008) is defined as the mean value of difference between the draught fore and the draught aft at the arrival port, i.e. T_{FA} and T_{AA} . The draught variables T_{FD} , T_{AD} , T_{PD} , T_{SD} , T_{FA} , T_{AA} , T_{PA} , and T_{SA} do not refer to the entire voyage, but each variable is measured both at departure and at arrival ports by four submersible transmitters located at fore and aft perpendiculars, and at port and starboard amidships sections. These data are acquired by four Vegawell 52 draught gauges (pressure transmitters; VEGA, Schiltac, Deutschland), each with a maximum deviation of 0.2%. These measurements are collected in port when SOG is less than 0.3 Knots because this sensor acquires the hydrostatic pressure. At high speed these measurements are affected by errors.

Departure and arrival trim ($Trim_D$ and $Trim_A$) are obtained through the inclinometer measurements and the geometric features of the ship.

Hourly Fuel consumption and CO₂ emission calculation

The response variable object of this study is the average fuel consumption per hour for each voyage (i.e. the ratio between fuel consumed and actual navigation time in hours for each voyage) during navigation

$$Y = \frac{Q}{h}, \quad (4.7)$$

where Q is the total fuel consumption for the voyage during the actual navigation time and h is the navigation time in hours. A brief overview of the ship engine room layout is presented to identify how Q is calculated. All the cruise ships monitored in this chapter have two engine sets, each with two Wartsila 12V46D main engines for propulsion with a variable pitch propeller and a Marelli shaft generator (keyed on a gearbox) for electric power; Figure 4.2 outlines the Engine room layout.

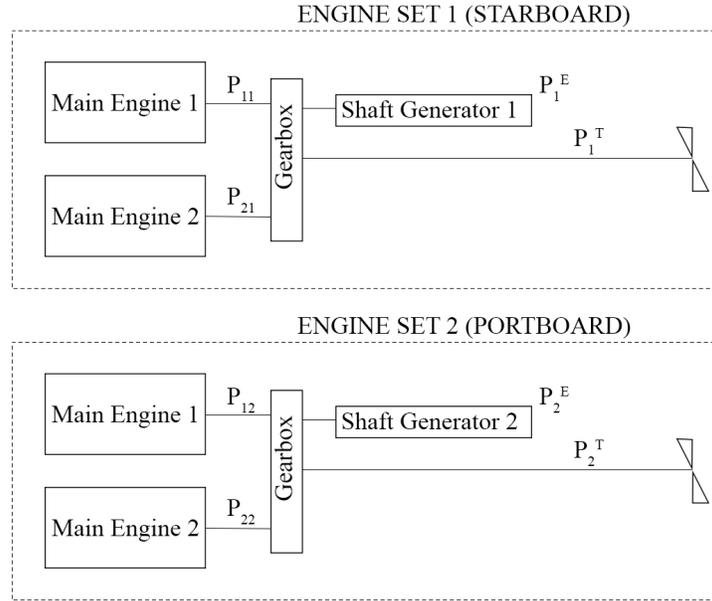


Figure 4.2: Engine room layout.

As detailed by Bocchetti et al. (2015), on the j -th engine set ($j=1, 2$) the DAQ device collects the thrust power P_j^T on the shaft propeller and the electrical power P_j^E on the shaft generator. The powers P_j^T and P_j^E are the only measurements available for calculating the main engine power. Note that when the engine operation mode is combinator, the shaft generator is necessarily powered off ($P_j^E = 0$). In this case, the electrical power is supplied by three diesel generators, which are intentionally not considered in the following fuel consumption calculation.

The actual fuel consumption Q related to the main engines is calculated through the following relation

$$Q = \sum_{i=1}^2 \sum_{j=1}^2 P_{ij} h_{ij} SFOC_{ij}, \quad (4.8)$$

where h_{ij} is the number of running hours of the i -th engine of the j -th engine set (with $i, j=1, 2$), $SFOC_{ij}$ is the specific fuel oil consumption of the i -th engine of the j -th engine set. Then,

the power P_{ij} of the i -th main engine of the j -th engine set can be calculated as follows

$$P_{ij} = \begin{cases} 0 & \text{if } x_{ij} = 0 \\ \frac{x_{ij}}{\sum_{i=1}^2 x_{ij}} P_j & \text{otherwise} \end{cases}, \quad (4.9)$$

where x_{ij} assumes the value 0 if the i -th main engine of the j -th engine set is powered off and 1 otherwise. For each voyage, the main engine power P_j of the j -th engine set is calculated as the mean value of

$$P_j = \frac{P_j^E}{\eta_j^e \eta_j^m} + \frac{P_j^T}{\eta_j^m}, \quad (4.10)$$

where η_j^m and η_j^e are the gearbox mechanical efficiency and the shaft generator electrical efficiency, respectively.

The proposed method can utilize CO₂ emissions as response variable, in particular, according to Annex I of MRV regulation (European Commission 2015), the calculation of CO₂ emissions can be performed exploiting the amount of ship fuel consumption through the following formula

$$\text{CO}_2 \text{ emission} = \text{fuel consumption} \times \text{emission factor}. \quad (4.11)$$

For each fuel type, a different value of the emission factor is available, according to the Intergovernmental Panel for Climate Change (IPCC) as reported in Annex VI (European Commission 2012).

Timeline and Maintenance Intervals

As already stated, for reasons of brevity, we show the most relevant results achieved for two of the four Ro-Ro Pax ships owned by Grimaldi Group. In particular, the application of the proposed approach for the on-line monitoring of fuel consumption and fault detection (aim (i)) is illustrated by means of data acquired on *Ship 1*; whereas, data acquired on *Ship 2* are used to show the capability of the proposed approach to assess and quantify the fuel consumption reduction related to a dry-dock operation (aim (ii)). In what follows, the data used to estimate the model are referred to as *calibration dataset*; whereas those used for aims (i) and (ii) are referred to as *monitoring dataset*. Accordingly, for *Ship 2* the calibration and monitoring datasets refer to data collected before and after EEI, respectively. The periods to which calibration and monitoring datasets refer, as well as those indicating dry-dock operations, are outlined in Figure 4.3 and Figure 4.4 for *Ship 1* and *Ship 2*, respectively. For the sake of completeness, for *Ship 1*, the calibration and the monitoring period include 606 and 720 voyages, respectively; for *Ship 2* the calibration (before EEI) and the monitoring (after EEI) datasets contain 329 and 462 voyages, respectively. In particular, for *Ship 1*, 11 months' worth of data collected right after EEI operation have been certified as reference data for the model calibration, in the extent of capturing all the typical operating conditions; whereas, for *Ship 2*, 10 months' worth of data between two dry-dock operations have been certified as reference data for the model calibration, Anomalous voyages have been identified by proper statistical procedures and are left out in the reference dataset only if being confirmed as exceptional by technical engineers.

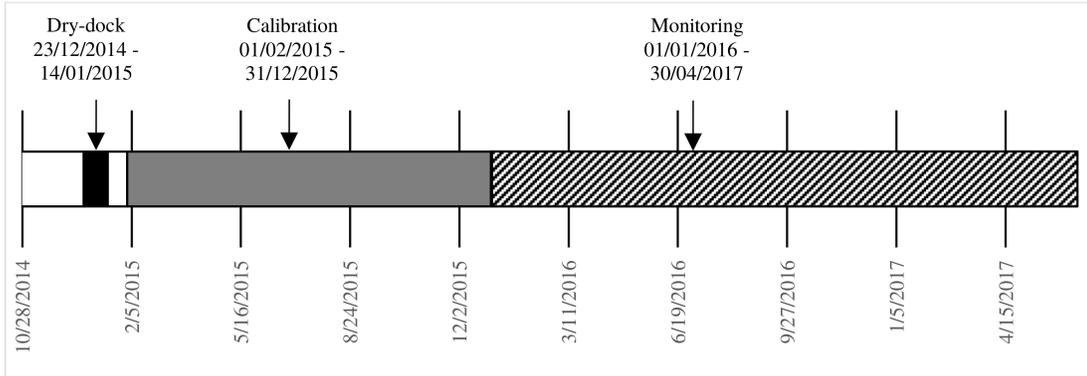


Figure 4.3: Ship 1 timeline.

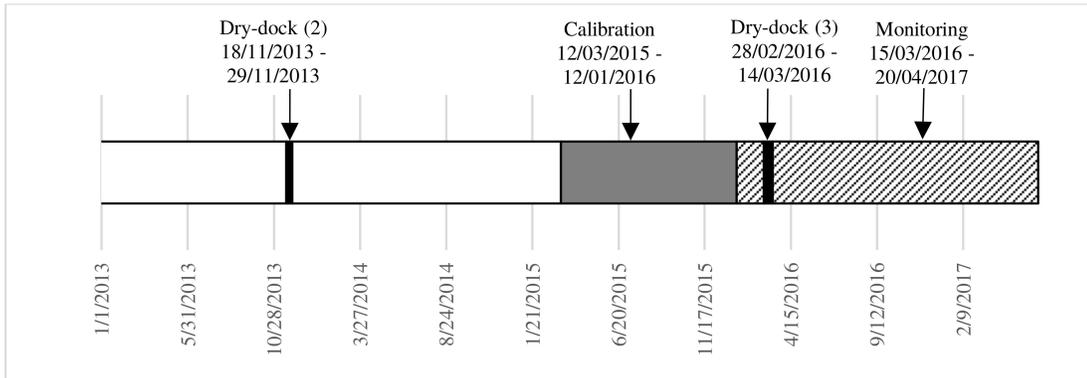


Figure 4.4: Ship 2 timeline.

4.3 The Statistical Approach and Monitoring Tools

In this chapter, PLS regression is used to evaluate the parameters of the statistical model for predicting and monitoring ship fuel consumption and emissions of CO₂. The choice of the PLS in place of e.g., multiple linear regression (Erto et al. 2015) has great potential of supporting the management to handle the great amount of data collected on board of modern ships that are usually noisy and strongly correlated. The residual left by the PLS model are also naturally prone to be monitored at each new voyage through *prediction error control chart*, whereas the predictor variables are monitored through the *Hotelling's T²* and *SPE_x* control chart, as detailed below. When a point falls outside the upper control limit of at least one of the latter control charts, a possible problem may have occurred. The management is then urged to further investigate physical variables that have caused the out-of-control condition by exploring the corresponding *contribution plot* (MacGregor & Kourti 1995).

From a mathematical point of view, the two statistics monitor different anomalies that may occur during voyages. In particular, a value exceeding the control limits of the Hotelling's *T²* chart indicates that the corresponding observation presents extreme values in one or more

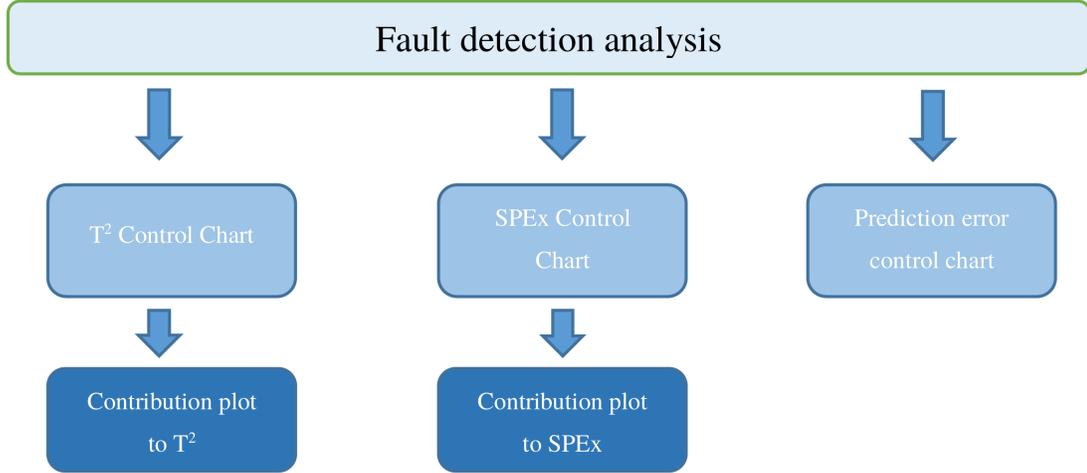


Figure 4.5: Fault detection analysis diagram.

physical variables, but plausibly maintains the same correlation structure as in the reference dataset (high variability *inside* the PLS model). In opposition, values exceeding the control limits in the SPE_x control chart are related to observations that have a different structure with respect to the reference data (*outside* the PLS model). Finally, if both T^2 and SPE statistics are out of control for an observation, we can say that there is some value in one or more physical variables that is extreme with respect to the reference dataset in the projected subspace (high variability inside the PLS model) and that, at the same time, this or other variables show a different correlation structure with respect to the reference dataset in the residual space (high variability outside the PLS model).

In Figure 4.5 the main steps of the statistical procedure proposed for monitoring of fuel consumption and diagnosis of faults is outlined. The first step is to set up the monitoring control chart of the T^2 statistic and identify the anomalous voyages for which the statistic falls outside the upper control limit. For each of these voyages, in order to identify the physical variables that have determined the largest value, the contribution plot to T^2 is built. The second step regards the SPE_x control chart. Similarly, to the T^2 statistic, for all voyages with the SPE_x statistic falling above the upper control limit, a contribution plot to the SPE_x statistic is performed. Finally, the prediction error control chart is built up. More information is given in the next section, which shows the procedure applied operatively to a real case study from a Ro-Ro Pax ship.

The Hotelling T^2 chart

Note that the detailed explanation of the statistical considerations about the monitoring statistics introduced in the previous subsection are not in the scope of this article, readers are referred to (Nomikos & MacGregor 1995b) for more discussion on this topic. When dealing with a new voyage to be monitored, let us denote by \mathbf{x}_{new} the vector of observations of the predictor variables from which the (multivariate) sample mean of the reference data is subtracted and with \mathbf{t}_{new} the vector of the corresponding observations of the latent variables (or scores) for this voyage. Moreover, let us denote with \mathbf{S} the sample covariance matrix of

the scores of the reference observations.

The Hotelling T^2 chart is a monitoring chart reporting for a single voyage to be monitored the relative statistic T^2 , which is calculated as

$$T^2 = \frac{\mathbf{t}_{new}^\top \mathbf{S}^{-1} \mathbf{t}_{new} N(N-R)}{R(N^2-1)}, \quad (4.12)$$

where N is the number of reference observations and R is the number of latent variables included in the model. Figure 4.6 shows the graphical meaning of T^2 statistics. Since T^2 has the F -distribution with R and $N-R$ degrees of freedom, the upper control limit T^2_{limit} is defined as follows

$$T^2_{limit} = F_{R, N-R, \alpha}, \quad (4.13)$$

where $\alpha = 0.01$ is the significance level. When the T^2 statistics for a new voyage falls above T^2_{limit} , a possible problem in a physical variable may have occurred. In particular, the value of one or more physical variables is unusual and gives a high contribution to the T^2 statistics that have determined an out of control signal in the T^2 statistic. It is possible to identify the physical variables that have the highest contribution to the out-of-control signal by calculating the contribution to the T^2 statistic according to MacGregor & Kourti (1995), as the elements of the vector

$$\text{Contribution}_{T^2} = \mathbf{t}_{new} \mathbf{p}_j^\top, \quad (4.14)$$

where \mathbf{p}_j is the j -th column of the loading matrix \mathbf{P} . Graphically, the plot of the contributions to T^2 is a bar plot, each bar displaying the corresponding physical variable used in the statistical model. The physical variables that have determined the largest contribution to the T^2 statistic are identified by the highest bars in the contribution plot. On the vertical axis, the values of the physical variables contribution are reported.

The SPE_x chart

For each voyage, the statistic reported in the SPE_x control chart is as follows MacGregor & Kourti (1995)

$$SPE_x = (\mathbf{x}_{new} - \hat{\mathbf{x}}_{new})^\top (\mathbf{x}_{new} - \hat{\mathbf{x}}_{new}), \quad (4.15)$$

where $\hat{\mathbf{x}}_{new} = \mathbf{t}_{new} \mathbf{P}^\top$ is the prediction of \mathbf{x}_{new} based on the PLS latent variable model. Accordingly, the upper control limit is defined as

$$SPE_{sup} = \theta_1 \left[\frac{c_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} + 1 \right]^{1/h_0}. \quad (4.16)$$

A point that falls outside of this limit may indicate that a variable has assumed an atypical value with respect to the other variables, and an exceptional cause may have occurred with respect to the baseline used to calibrate the model, as shown in Figure 4.7. As for the T^2 statistic, the contribution to the SPE_x

$$\text{Contribution}_{SPE} = \mathbf{x}_{new} - \hat{\mathbf{x}}_{new} \quad (4.17)$$

can be plotted to have indications on physical variables that may have concurred to the out-of-control.

Variable contributions to T^2 and SPE_x statistics are such that large contributions in absolute value lead to large values of the control statistics. Then, when an out of control signal is detected in one of the two control charts, one can identify the responsible variables by looking at the larger absolute contribution values.

The prediction error control chart

As the residuals left by the PLS model are also naturally prone to be monitored at each new voyage a *prediction error control chart* can be defined by plotting the predicted hourly fuel consumption \hat{y}_{new} and the $100(1 - \alpha)$ prediction interval calculated as in Nomikos & MacGregor (1995a)

$$\hat{y}_{new} \mp t_{\alpha/2;n-R-1} \sqrt{\frac{\mathbf{e}^\top \mathbf{e}}{n-R-1} [1 + \mathbf{t}_{new}^\top (\mathbf{T}^\top \mathbf{T})^{-1} \mathbf{t}_{new}]}. \quad (4.18)$$

The matrix \mathbf{T} is the matrix of the scores based on the reference data, \mathbf{e} is the vector of the prediction error of the response variable for the reference observations, and $t_{\alpha/2;n-R-1}$ is the $100\alpha/2$ percentile of a Student's distribution with $n-R-1$ degrees of freedom. Note that from an engineering point of view, the use of the prediction error control chart is discouraged when the T^2 and SPE_x control charts signal an out of control, since this can cause the problem of *extrapolation*, i.e., using the estimator \hat{y}_{new} in Equation (4.18) beyond the boundary of the predictor space (Montgomery et al. 2012). On the other hand, anomalies detected only by the prediction error control chart cannot be addressed to any of the monitored variables and therefore, they plausibly pertain to a change among factors not included in the set of predictor variables. Moreover, the usefulness of the prediction error control chart is the possibility to detect anomalous trends of the response variable over subsequent voyages or shifts from zero of the prediction error mean, which are not observable by neither the T^2 nor SPE_x control charts.

4.4 Real-case study

However, even if we have currently been implementing the model on all the four Ro-Ro Pax ships mentioned above, for reasons of brevity, in what follows we show only the most relevant results achieved for two ships (hereinafter referred to as *Ship 1* and *Ship 2*). In particular, Section 4.4 show the capability of this procedure to support management to identify anomalous fuel consumption (and thus CO₂ emissions) and the physical variable (prognosis of fault) that give the highest contribution to the out-of-control are identified for *Ship 1*. whereas, Section 4.4 illustrates how the approach is able to quantifying FCR consequent to a silicone foul release coating of the hull of *Ship 2*.

Prognosis of faults

In this section, for illustrative purposes a relevant case study is presented to underline the ability of the proposed procedure to discover anomalous voyages and to support management in making suitable decisions to solve the problem that occurred on *Ship 1*. Note that each voyage is identified with a unique voyage number (VN).

In this case, a monitoring window from VN 2034 to 2053 of *Ship 1* is considered. With respect to this monitoring window, Figure 4.6, Figure 4.7 and Figure 4.8 show the Hotelling T^2 control chart, the SPE_x control chart, and the prediction error control chart, respectively. In particular, VN 2035 exceeds the upper control limit in all the three charts:

As previously explained, according to the procedure displayed in the flow chart of Figure 4.5, when an observation falls outside the upper control limits in both the T^2 and SPE_x charts, to further investigate the cause that occurred, the contribution plots to T^2 and SPE_x are produced (Figure 4.9 and Figure 4.10).

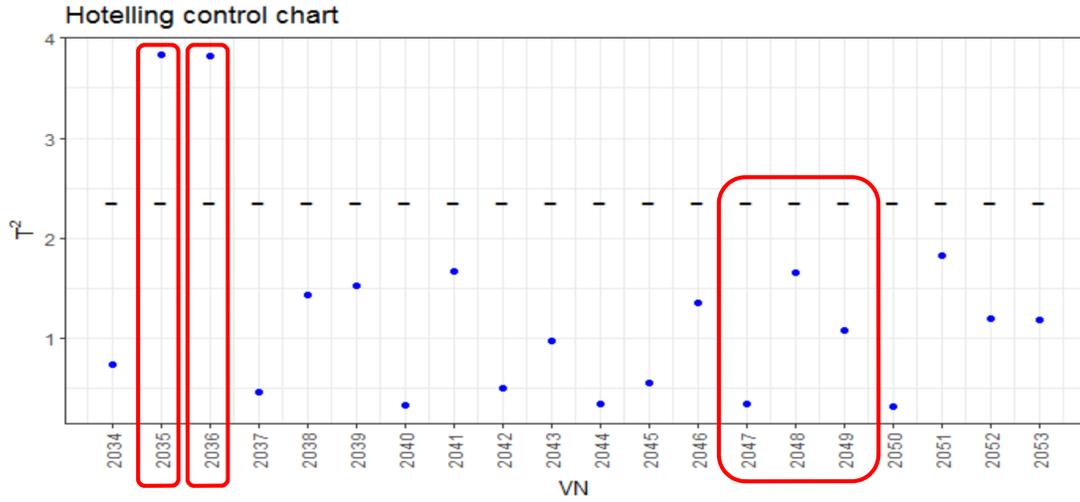


Figure 4.6: Hotelling T^2 control chart for voyages 2034–2053 of Ship 1.

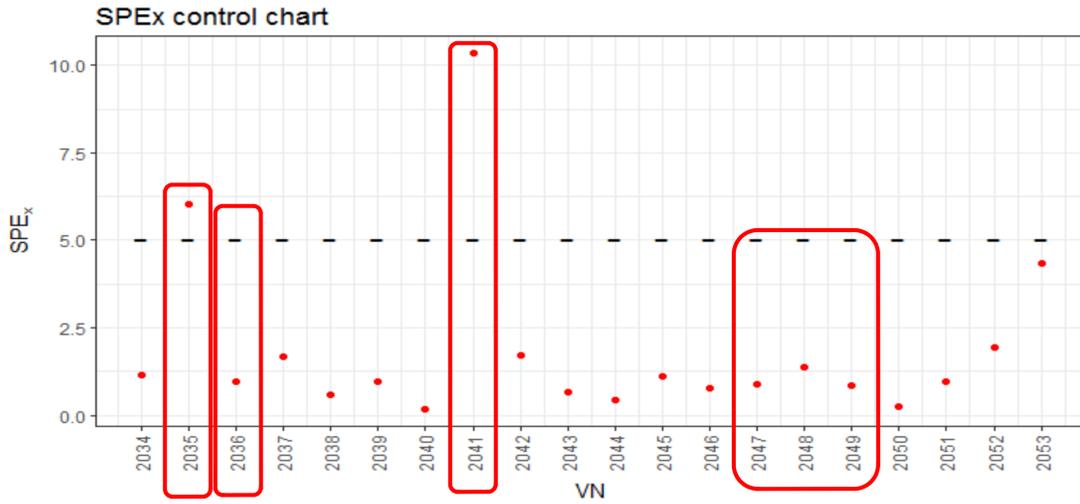


Figure 4.7: SPE_x control chart for voyages 2034–2053 of Ship 1.

4. SHIP FUEL CONSUMPTION MONITORING AND FAULT DETECTION VIA CONTROL CHARTS BASED ON PARTIAL LEAST SQUARES REGRESSION

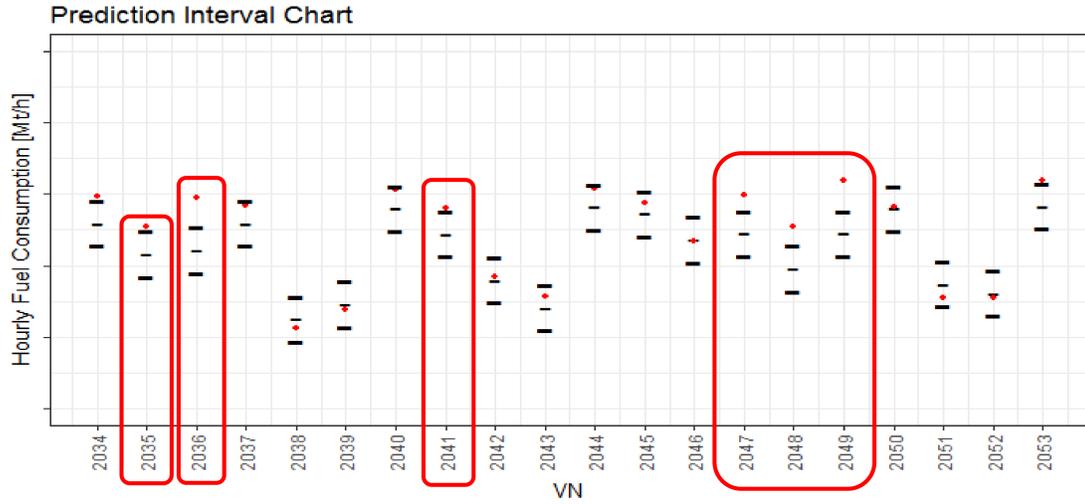


Figure 4.8: Prediction error control chart for voyages 2034–2053 of Ship 1.

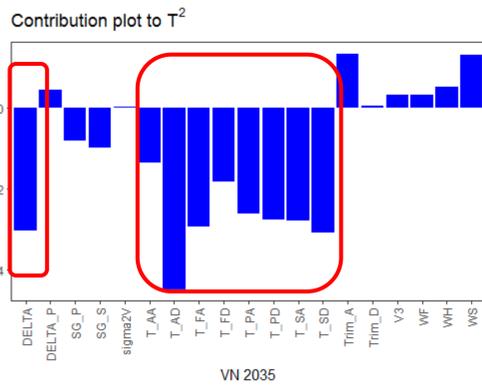


Figure 4.9: Contribution of the variables to the Hotelling T^2 statistic, for the voyage 2035 of Ship 1.

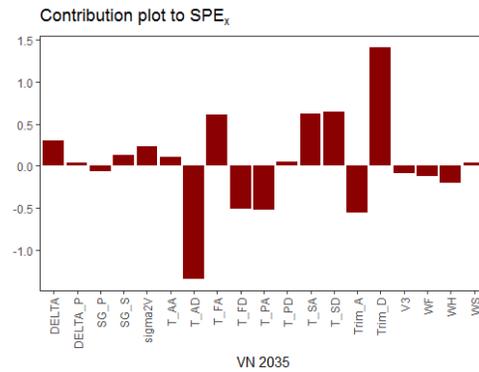


Figure 4.10: Contribution of the variables to the SPE_x statistic, for the voyage 2035 of Ship 1.

Draughts and displacement for voyage 2035 display the largest contribution to the T^2 statistic (Figure 4.9), which support technical investigations in detecting an error signal in the draft gauge sensor. The large contribution to SPE_x of the departure trim in Figure 4.10 is due to an unusual value of draught variables. In fact, trim and draught variables are related and an error signal in the draft gauge sensor turn into an error in the trim value accordingly.

Referring to Figure 4.6, for VN 2036 the Hotelling T^2 statistic falls outside the control limit, therefore the respective contribution plot is reported in Figure 4.11.

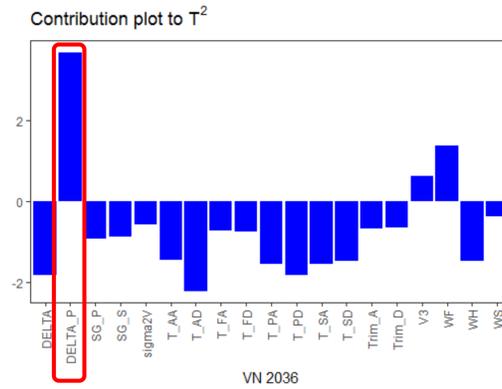


Figure 4.11: Contribution of the variables to the Hotelling T^2 statistic, for the voyage 2036 of Ship 1.

The physical variable responsible of the out of control in the Hotelling T^2 control chart is ΔP , i.e. the power difference between shaft propellers; further investigations revealed that the ship sailed with three engines in service during one hour of the total sailing time.

VN 2041 is out of control both in the SPE_x control chart (Figure 4.7), and in the prediction error control chart (Figure 4.8). For this voyage, a contribution plot to SPE_x is built in Figure 4.12.

4. SHIP FUEL CONSUMPTION MONITORING AND FAULT DETECTION VIA CONTROL CHARTS BASED ON PARTIAL LEAST SQUARES REGRESSION

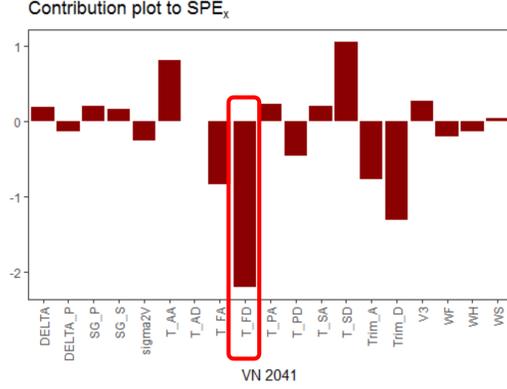


Figure 4.12: Contribution of the variables to the SPE_x statistic, for the voyage 2041 of Ship 1.

The physical variable that gives the highest contribution of SPE_x is the draught in fore at departure. The cause is an error in the draft gauge sensor.

Instead, from VN 2047 to 2049 both the monitored statistics T^2 and SPE_x are in control (Figure 4.6 and Figure 4.7), while the prediction error control chart (Figure 4.8) underlines for these voyages an extra-consumption of fuel. In this case, contribution plots are not needed because the statistics T^2 and SPE_x do not exceed their upper control limits. The causes of extra-consumption are to be detected outside the variables considered in the statistical model.

For graphical reasons, Figures 4.6 through 4.8 show only 20 relevant voyages (from 2034 to 2053) out of the 720 included in the monitoring dataset of *Ship 2*.

Energy efficiency quantification

The statistical approach presented is also able to quantify the energy saving after an energy efficiency improvement (EEI) operation. In particular, the fuel consumption reduction and therefore the CO₂ emission reduction can be quantified through the saving z_i of a new voyage i defined as

$$z_i = (\hat{y}_i - y_i)h_i, \quad (4.19)$$

where \hat{y}_i is hourly fuel consumption predicted for the new voyage, y_i is the actual hourly fuel consumption for the same voyage and h_i is the sailing time. As consequence, the quantity $\hat{y}_i - y_i$ represents the hourly saving after the EEI operation. To quantify the total fuel consumption reduction (FCR) after an EEI operation, the set voyages to be considered refers to the monitoring period that immediately follows the EEI operation. Only voyages that are in control in both Hotelling T^2 and SPE_x control chart are considered.

The FCR can be calculated as

$$FCR = \sum_{i=1}^{N_{EEI}} z_i \quad (4.20)$$

where N_{EEI} is the number of new voyages considered after the EEI to evaluate its effectiveness.

A relevant case study follows. Considering the *Ship 2*, the EEI operation that was set up for this ship is the installation of silicone foul release coating on the wetted surface of the hull. This is a non-toxic hull paint based on fluoropolymer and siloxane (silicone polymer) coating. The main property of this class of painting is its “non-stick” surface that prevents the attachment of marine organisms to the hull, therefore it avoids an increase in the resistance, which requires additional power and consequently additional fuel consumption (and CO₂ emissions) to maintain the same vessel speed. Setting up this kind of EEI operation, shipping management is able to determine the actual improvement in shipping performance as well as economic and environmental savings consequent to reduction in fuel consumption and CO₂ emissions. One year’s worth of data of voyages after the dry-dock operation (VN 427 to 888), from 15 March 2016 to 14 March 2017 are monitored to quantify the FCR.

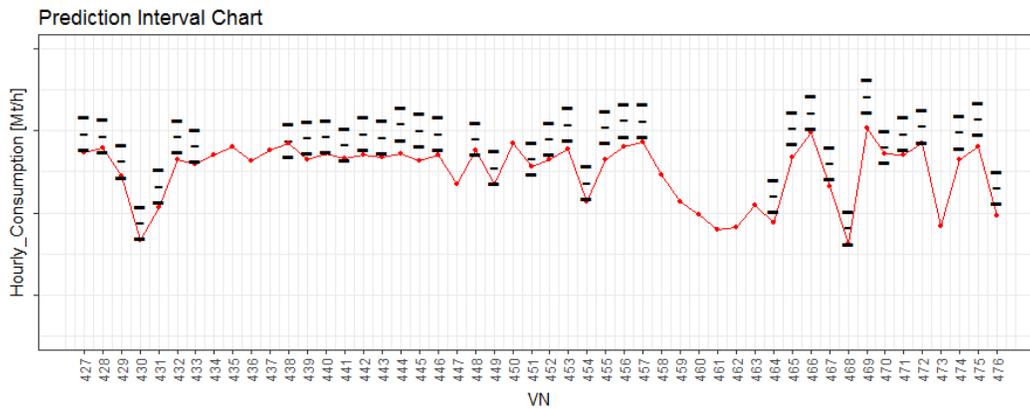


Figure 4.13: Prediction error control chart for voyages 427–476 of Ship 2, which follow an energy efficiency operation.

For graphical reasons, Figure 4.13 shows the prediction error control chart for voyages 427—476 of Ship 2 to illustrate the improvement and the consequent saving gained after the installation of silicone foul release coating only for the first month following this operation. Note that the actual hourly fuel consumption is lower than the prediction hourly fuel consumption for all the voyages, underlying that the saving has been effectively obtained thanks to the EEI. Moreover, voyages displayed without their relative prediction error control chart are those outside upper control limit in Hotelling T^2 and/or SPE_x control chart, therefore according to Nomikos & MacGregor (1995a) their prediction interval is not calculated and not considered for the quantification of FCR. In particular, the percentage of FCR quantified on one year of observations is 9.25%. This is a profitable result for the shipping management.

4.5 Conclusion

The shipping industry is facing a new regulatory regime that aims to give public access to CO₂ emissions data and this challenge can be addressed by making better use of the massive

4. SHIP FUEL CONSUMPTION MONITORING AND FAULT DETECTION VIA CONTROL CHARTS BASED ON PARTIAL LEAST SQUARES REGRESSION

amounts of sensor data now available. The International Maritime Organization—through the Energy Efficiency Design Index and the Ship Energy Efficiency Management Plan—and the European Union—through the application of the EU regulation 2015/757, which is mandatory from January 2018—urge shipping companies to set up a system for daily monitoring of emissions from each ship. Engineers in the naval sector traditionally rely on deterministic relationships among the physical variables of interest in order to make decisions. Moreover, even if manual acquisition of emission data is allowed by all the international regulations (e.g. bunker fuel delivery note, bunker fuel tank monitoring), it is yet affected by uncertainty due to human intervention and thus will be eventually unusable for monitoring purposes. On the other hand, the massive amount of navigation data acquired by multi-sensor systems installed on-board of modern ships have a great potential to naturally comply with those regulations but are hampered by the lack of effective methods in the maritime literature. Despite this advantage, many shipping companies limit their analyses to only calculating simple summary statistics at each voyage. The range of tools needs to be extended to include statistical methods for detecting patterns in data. The main contribution of this work is to provide statistical tools that effectively support the management. Without those statistical tools, the management cannot easily investigate and diagnose faults responsible for increase in fuel consumption (and CO₂ emissions). The proposed statistical approach is therefore able to support managerial decision making by setting up suitable actions to improve ship performance as well as to quantify consumption/emission savings after energy efficiency improvement operations (e.g., hull form optimization, hull cleaning and propeller polishing, ultra-smooth coating, engine maintenance operation, propulsion and power plant efficiency improvement). As is known, this is particularly profitable for shipping companies in order to claim for carbon credit.

The proposed statistical approach is not only useful to maritime engineers and shipping companies, but also to the international organizations responsible to adopt regulations related to the CO₂ emission monitoring problem. In fact, the same techniques can be adopted to verify that shipping companies satisfy the regulatory requirements. Currently, however, these statistical tools are difficult to implement and spread among shipping companies because such organizations are not ready to be forced to use automatic systems for the statistical analysis of emission data in order to verify that CO₂ emissions are coherent with the ship characteristic. The purpose of this chapter is to describe the methodology and demonstrate its applicability. Ongoing implementation is expected to yield further evidence of the methodology's benefits and usability.

Chapter 5

Feature-oriented methods for the prediction of CO₂

Shipping companies are forced by the current EU regulation to set up a system for monitoring, reporting, and verification (MRV) of harmful emissions from their fleet. In this regulatory background, data collected from on-board sensors can be utilized to assess the ship's operating conditions and quantify CO₂ emission levels. The standard approach for analysing such datasets is based on summarizing the measurements obtained during a given voyage by the average value. However, this compression step may lead to significant information loss since most variables present a dynamic profile that is not well approximated by the average value, only. Therefore, in this work, we test two feature-oriented methods that are able to extract additional features, namely, profile-driven features (PdF) and statistical pattern analysis (SPA). A real dataset from a Ro-Pax ship is then considered to test the selected methods. The dataset is segregated according to the voyage distance into short, medium, and long routes. Both PdF and SPA are compared to the standard approach and the results demonstrate the benefits of employing more systematic and informative feature-oriented methods. For the short route, no method is able to predict CO₂ emissions in a satisfactory way, whereas for the medium and long routes, regression models built using features obtained from both PdF and SPA resulted in improved prediction performance. In particular, for the long route, the standard approach failed to provide reasonably good predictions.

5.1 Introduction

The extensive and increasingly demanding air pollution programs (IMO 2012*b,d,c,a*, Smith et al. 2015) make predicting ships' CO₂ emissions not only a strategic, but also a mandatory task for shipping companies. At the European level (European Commission 2015), they are in fact forced to set up a system for daily monitoring, reporting, and verification (MRV) of harmful emissions for their fleet. Concurrently, automatic multi-sensor acquisition systems installed on-board of modern ships facilitate collecting operational data (i.e., ship operating conditions) from a massive number of variables (e.g., vessel's resistance, propulsion power, fuel consumption, speed, routes, weather conditions, sea weaves, displacements, draughts, trim, engine operation mode, etc.). The profile of these variables is typically complex, unstructured, intrinsically collinear and with non-stationary behaviour. In this scenario, data scientists are called upon in order to develop new predictive techniques that might answer

positively to the need of predicting CO₂ emissions and have a tangible impact in today's maritime industry and policy-making (Buhaug et al. 2009, Løvoll & Kadal 2014). More specifically, operational data show great potential to reveal, by means of suitable analysis, relevant and timely information (Bertram 2011) to be used for assessing the ship's efficiency in terms of CO₂ emissions. On this regard, classical approaches used in the naval literature may fail or are at least suboptimal, since they are limited to the analysis of averages per voyage. However, in spite of the easier interpretability, compressing a variable profile into a single average value may lead to significant information loss and to discarding most of the relevant dynamic patterns.

In the opposite spectrum of complexity, multivariate statistical methods commonly used for monitoring batch processes have also appeared in the shipping industry statistical literature (Lepore et al. 2018). However, they usually require the implementation of data pre-processing techniques that constitute an additional challenge for practitioners and may hamper their practical usability. For example, data needs to be correctly unfolded to handle its three-way structure, resulting in a very large number of pseudo-variables and model parameters. Furthermore, complex synchronization methods are required in order to ensure that the voyages' major landmarks are aligned and that all voyages have the same number of observations.

The batch process monitoring literature is vast and another class of approaches that is growing in importance is the class of feature-oriented methods. These methods are simpler to apply because they do not require synchronization and tend to be more parsimonious since the number of model parameters is smaller. Examples of feature-oriented methods include profile-driven features (PdF), recently proposed by Rendall et al. (2017), and statistical pattern analysis (SPA), proposed by He & Wang (2011). These techniques compress each variable into a small number of features that can be utilized for data-driven model building. In our previous work (Lepore et al. 2017), we utilized the standard approach in the shipping industry based on average values (AV) of each variable as predictors of CO₂ emissions. Therefore, it is now opportune to assess whether additional information, in the form of features, can actually be extracted from the variables' profiles for improving prediction performance. Accordingly, in this work, we study the suitability of both PdF and SPA to extract additional information from the on-board sensor data that are collected during each voyage and we compare them to the standard AV approach, which will be used as the benchmark method for this study. A real case study on shipping data acquired on-board of a Ro-Pax cruise ship, owned by the shipping company Grimaldi Group, is presented to illustrate the predictive performances of the proposed methods.

The remaining of this chapter is organized as follows. Section 5.2 details the dataset considered in this work as well as the feature-oriented methods tested. Section 5.3 describes the comparison framework employed for assessing and comparing the different feature-oriented methods, while Section 5.4 presents the results obtained, followed by their discussion. Finally, Section 5.5 summarizes the main conclusions of this work.

5.2 Materials and Methods

In this section, we provide details on the variables collected during each voyage of the Ro-Pax cruise ship mentioned in the introduction and on the feature-oriented methods utilized for converting the on-board sensor data into features. These features constitute the predictors that are used later on for developing regression models for predicting CO₂ emissions.

Data description

The dataset consists of operational data acquired during 679 voyages performed by a Ro-Pax cruise ship during a period of one year and three months. The name of the ship and voyage dates are not reported for confidentiality reasons. Voyages are grouped by length into short, medium and long routes. In particular, there are $N = 302$ short-route voyages, $N = 78$ medium-route voyages, and $N = 299$ long-route voyages.

The scalar response is the cumulative CO₂ emissions at the end of each voyage. The variables that are measured during each voyage and that were selected for analysis are presented in Table 5.1. Further details can be found in Bocchetti et al. (2015). In the following paragraphs, we provide more details regarding the variables collected during each voyage.

A large-scale IT network system acquires observations from each of the variables shown in Table 5.1. The collected values refer to the actual voyage navigation time, which is defined as the time between the “finished with engine order” (when the ship leaves the departure port) and the “stand by engine order” (when the ships enter the arrival port). The Speed Over Ground (SOG), denoted by V , is obtained as the ratio between the sailed distance over ground and the actual navigation time. The Sailed Distance Over Ground is the distance travelled by the vessel during the actual voyage navigation time, measured in Nautic Miles (NM) and calculated by the on-board data acquisition system from GPS latitude and longitude using the Haversine formula.

The power difference between port and starboard propeller shafts, denoted by ΔP , is useful for discovering anomalies or malfunctioning in the main engines. The power difference between two shaft generators, denoted by ΔSG , takes into account the different operation modes of navigation (constant speed mode and combinator mode).

Longitudinal and side wind, denoted by W_L and W_S , respectively, are calculated based on data acquired by the anemometer sensors and are defined on the basis of the true wind speed, denoted by V_{WT} , and the difference between the true wind angle (in the earth system) and the course over ground, Ψ_{WT} . That is, $W_L = V_{WT} \cos \Psi_{WT}$ and $W_S = -|V_{WT} \sin \Psi_{WT}|$.

The air temperature variables, denoted by T_1 , T_2 , T_3 , and T_4 , respectively, measure the air temperature of the four main engines. The sailing time h indicates the cumulative navigation time during the voyage. The trim variable, denoted by $Trim$, is obtained through the inclinometer measurements and the geometric characteristics of the ship. The percentage of total miles (PM) indicates the voyage completion proportion and is defined as the ratio of travelled distance to the total travelled distance.

Feature-oriented methods

A feature-oriented method is a procedure for transforming raw measurements into a set of potentially useful features that summarize the relevant information contained in the measurements. In particular, they can be viewed as a mapping from a measurement space (M), where data is collected, to a feature space (F) and finally to a decision space (D) Pal & Mitra (2004):

The mapping from M to F is made using a feature-oriented method (also called a feature dictionary) that defines the algorithmic procedure to obtain a finite set of features from measured data. In the context of CO₂ prediction, the measurement space contains the trajectories of the variables \mathbf{X} that are measured during each voyage and has dimensions $N \times J \times K_n$, where N is the number of voyages, J ($J = 12$) corresponds to the number of

Table 5.1: Variables used as covariates in the feature-oriented approach.

	Variable	Description
1	V	Speed Over Ground (SOG) [kn]
2	ΔP	Power difference between port and starboard propeller shafts [kW]
3	ΔSG	Power difference between two shaft generators [kW]
4	W_L	Longitudinal wind [kn]
5	W_S	Side wind [kn]
6	T_1	Air temperature of port engine 1 [$^{\circ}C$]
7	T_2	Air temperature of port engine 2 [$^{\circ}C$]
8	T_3	Air temperature of starboard engine 1 [$^{\circ}C$]
9	T_4	Air temperature of starboard engine 2 [$^{\circ}C$]
10	h	Sailing time [h]
11	$Trim$	Trim [m]
12	PM	Percentage of total miles [%]

measured variables, and K_n is the number of samples collected for the n^{th} voyage, with $n = 1, \dots, N$). Therefore, a feature dictionary transforms the three-way array $\underline{\mathbf{X}}$ into a feature matrix \mathbf{F} (with dimensions $N \times n_f$, where each row contains the features for a given voyage, and n_f is the number of features extracted by the selected dictionary). The decision space, D , contains the response value for each sample/voyage in M and corresponds to the cumulative values of CO₂ emissions per voyage. The effect of applying a feature-oriented method is graphically depicted in Figure 5.1.

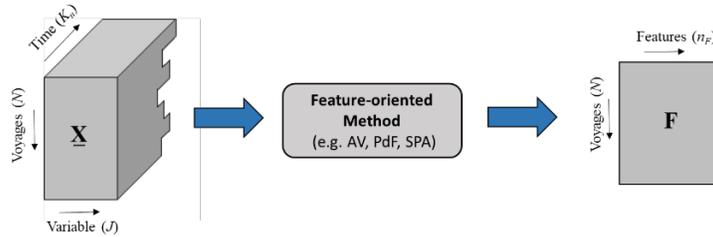


Figure 5.1: The application of a feature-oriented method transforms the input three-way structure $\underline{\mathbf{X}}$ into a feature matrix \mathbf{F} . AV, PdF and SPA stand for average values, profile-driven features, and statistical pattern analysis, respectively.

As stated before, the feature-oriented methods contemplated in this chapter, namely, PdF (Rendall et al. 2017) and SPA (He & Wang 2011, Wang & He 2010), are compared with the standard method based on the AV.

The AV approach consists in using the average value of each variable over a voyage as a feature and, therefore, the number of features per voyage corresponds to the number of variables J . In summary, the measured data $\underline{\mathbf{X}}$ is transformed into a feature matrix \mathbf{F}_{AV} with dimensions $N \times J$. This standard approach has been successfully utilized in our previous paper (Lepore et al. 2017), where we considered a larger set of variables. Nevertheless, compressing all the information contained in the profile of a variable into a single numerical

Table 5.2: Object-profiles and their respective features considered in the PdF dictionary.

Object-Profile	Features
Constant 	Average value Variance Area
Linear 	Slope Intercept SSR* Residual variance Area
Step 	Step occurrence Average value before and after step Variance before and after step SSR* Area
Pulse 	Pulse Beginning Average value before, after and in the pulse Variance before, after and in the pulse SSR* Area

*SSR corresponds to the sum of squared residual errors between estimated and measured profile.

value may be suboptimal. In particular, some variables present dynamic patterns that are not well represented by the average value, incurring in significant information loss. The use of PdF and SPA attempt to overcome this issue and lead to larger sets of features that capture, besides the mean values, other characteristics of the variables' profile, such as transients, stochastic variation and slopes, which may be relevant for predicting CO₂ emissions.

PdF is a methodology based on a set of archetype profiles (also called object-profiles in the original paper) that represent typically found trajectories. There are specific features associated with each archetype, characterizing the main dynamic patterns contained in the variable's trajectory. The object-profiles utilized in this work are presented in Table 5.2. In order to apply PdF, each variable collected from the on-board sensors is first assigned to a given profile archetype by visual inspection or automatically (the former alternative tends to be preferred since it is more reliable and can be easily conducted alongside a first stage of exploratory data analysis). The assignment is based on the similarity between the trajectory of a variable over all voyages and the archetype. Following the assignment step, an estimated trajectory is obtained using the measured data and the respective features are computed. Therefore, PdF transforms the measured data \mathbf{X} into a feature matrix \mathbf{F}_{PdF} . Further details of PdF can be consulted in the original paper (Rendall et al. 2017).

SPA is a feature extraction procedure based on the variables' statistical moments. In this context, typical features include the mean, variance, covariance between variables, kurtosis, and skewness. These features are few in number and describe many aspects of a variable's statistical distribution. SPA converts the three-way structure \mathbf{X} into a feature matrix \mathbf{F}_{SPA} (with dimensions $N \times (J(J+7))/2$). One can also note that this approach contains AV as default features and, therefore, it is expected to be at least as good as the approach based on AV.

The feature matrices \mathbf{F}_{AV} , \mathbf{F}_{PdF} , and \mathbf{F}_{SPA} , which contain the features extracted from the measured data \mathbf{X} by the three aforementioned approaches, are considered separately in order to develop regression models for predicting CO₂ emissions. For this task, a predictive analytics comparison (PAC) (Rendall & Reis 2018) framework is applied to compare, in a rigorous way, the predictive content of each feature set, as well as to identify the most promising regression methods.

5.3 Predictive analytics comparison framework

The predictive analytics comparison framework (PAC) is a robust methodology for assessing and comparing the prediction performance of different regression methods. It takes, as input, a dataset arising from a regression problem and produces a report that highlights the more promising regression method (and/or class of regression methods), signals important predictor variables, and can be used to infer combinations of methods that may be desirable. In the particular context of predicting CO₂ emissions, PAC is applied to each feature matrix separately (\mathbf{F}_{AV} , \mathbf{F}_{PdF} or \mathbf{F}_{SPA}) and allows the identification of the feature set that is richer in terms of predictive content: PAC has four main components: the analytics domain, the data domain, the comparison engine, and the results report. Each component is briefly described next.

Analytics domain. The analytics domain is composed of different classes of regression methods that are used for developing models that relate features and CO₂ emissions. Four classes of regression methods are considered: variable-selection methods, penalized regression methods, latent variable methods, and tree-based ensembles. Each class has different *a priori* assumptions and their suitability is dependent on whether or not they find good adherence to the data generating mechanisms. The class of variable selection assumes that only a few relevant variables contain information regarding the response, therefore, these methods employ various strategies in order to select important features and remove irrelevant ones. Forward stepwise regression (FSR) (Andersen & Bro 2010, Montgomery & Runger 2010) is selected as a representative method from this class. The class of penalized regression employs a penalty to the magnitude of the regression coefficients, constraining their values to be low and may even eliminate some features (when the estimated regression coefficient is zero). In this class, ridge regression (RR) (Hoerl & Kennard 1970), least absolute shrinkage and selector operator (LASSO) (Tibshirani 1996), elastic net (EN) (Zou & Hastie 2005), and support vector regression (SVR) (Smola & Schölkopf 2004, Ahmed et al. 2010) were included. Latent variable methods are based on the assumption that a few underlying unobservable variables govern the variability in the data. Therefore, they estimate the latent structure and the model can also be used to predict the response. Principal component regression (PCR) (Jolliffe 2002, Jackson & Mudholkar 1979, Wold et al. 1987) and partial least squares (PLS) regression (Wold et al. 2001, Geladi & Kowalski 1986) were included as representative methods for this class. Finally, tree-based ensemble methods use ensembles of regression trees in order to approximate the relationship between predictors and response variable. A regression tree (Strobl et al. 2009, Breiman et al. 1984) splits the input space into regions where the response variable tends to be constant, therefore, it provides a piecewise-constant approximation for the response variable. The predictions obtained from a single tree tend to present high variance, and ensembles are needed in order to decrease the variance and obtain stable and reliable estimates. Compared to the other classes, tree-based ensembles have the ability to model non-linear relationships between features and CO₂ emissions, broadening

the applicability of PAC to non-linear problems. In this class, bagging of regression trees (BaRT), random forests (RF), and boosting of regression trees (BoRT) (Freund et al. 1999, Cao et al. 2010) were considered. A more detailed description of the methods considered here can be consulted in the original paper (Rendall et al. 2017) or in its contextualized application to the prediction of CO₂ emissions (Lepore et al. 2017).

Data domain. The data domain provides the factual evidence that is used to assess the prediction performance of the regression methods considered in the analytics domain. This component is problem-specific and is often structured as a tabular dataset, where rows represent samples and columns correspond to different predictor variables. In the particular case of CO₂ prediction, the dataset is composed of features that are extracted following the procedure described for each dictionary, and each row corresponds to a voyage. Thus, this component contains \mathbf{F}_{AV} , \mathbf{F}_{PdF} , \mathbf{F}_{SPA} , (as predictors, considered one at a time) as well as the cumulative CO₂ emissions for each voyage (response variable).

Comparison engine. This component is related to assessing and comparing the performance of the regression methods considered in the analytics domain. In PAC, a robust comparison strategy based on double cross-validation is utilized, which is an enhanced approach when compared to single cross-validation. In brief terms, double cross-validation splits a dataset into a training and a validation sets. Models are built using the training set and their hyper-parameters are tuned using standard cross-validation (this constitutes the inner loop of double cross-validation). The models are then utilized to predict the validation set and the root mean squared error in the validation set ($RMSE_{dcv}$) is computed as the base metric of performance. The whole procedure of dataset splitting and model building can be repeated for a pre-defined number of iterations (this constitutes the outer loop of double cross-validation). The distribution of $RMSE_{dcv}$ obtained in the end, characterizes the prediction errors incurred by each method.

Results report. The last component of PAC concerns the presentation of results in a way that is user-oriented, highlighting methods that have lower $RMSE_{dcv}$. One can also assess whether a given class of methods is more promising than others and conjecture possible combination of methods that may be suitable or opportune to consider. In a more detailed analysis, important features can be identified for the top regression methods, improving the knowledge of the prediction problem at hand.

5.4 Results and Discussion

This section presents the results obtained for predicting CO₂ emissions based on the three methods considered, namely, AV, PdF, and SPA. As previously stated, the results are segregated according to the voyage distance into short, medium, and long routes in order to allow for a more detailed analysis.

Short route

The dataset collected for the short route contains $N = 302$ samples. PdF and SPA generate 37 and 114 features, respectively, whereas AV trivially considers 12 features (one average value for each variable of Table 5.1). The results obtained after applying PAC to these three feature dictionaries are presented in Figure 5.2. As can be observed in Figure 5.2, there is no practical difference and all dictionaries lead to similar predictions in the validation set for all iterations of double cross-validation. In other words, there appears to be no added value in this case from extracting additional features besides the average values. This is

coherent with results of our previous work (Lepore et al. 2017), where CO₂ emissions for the short route were also poorly predicted because, it was conjectured, the short distance of the route does not allow the main factors to significantly contribute to CO₂ emissions. Instead, the systematic variability is small, and the methods cannot capture any significant trends in the data. This is further corroborated here, as including additional features does not improve prediction performance. Even if there are statistical differences between the different combinations of regression methods and feature dictionaries, these differences are rather small for practical purposes. Therefore, no method is highlighted for the short route as the best one and the predictions are considered not to be reliable, which limits the interpretive insights that can be obtained from the developed models. This is also confirmed when analyzing the distribution of the coefficients of determination over different runs of double cross-validation (R_{dcv}^2), which is typically close to zero or even negative in some iterations (the results are not shown here for simplicity). Thus, no further analysis is made to identify the most relevant features.

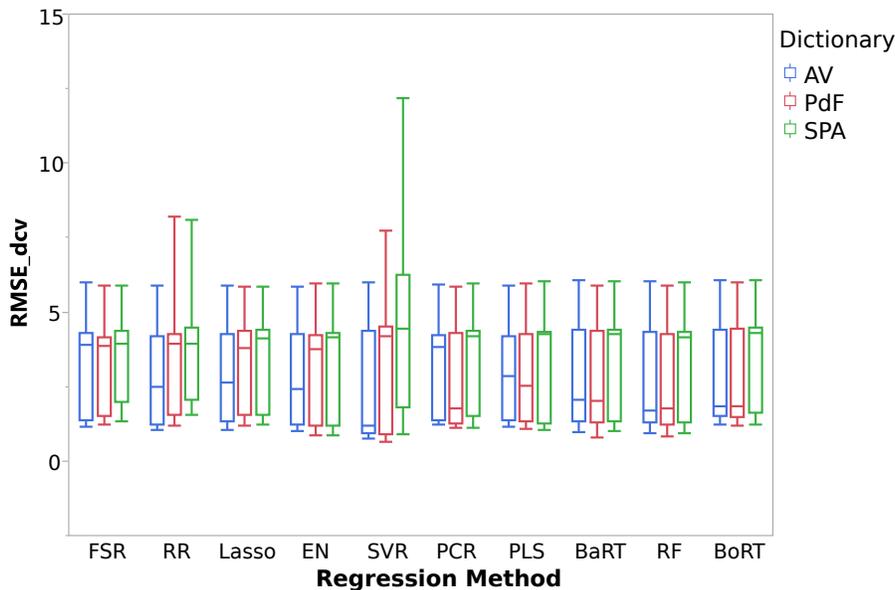


Figure 5.2: Distribution of $RMSE_{dcv}$ for 50 iterations of double cross-validation for the short route.

Medium route

The dataset for the medium route contains $N = 78$ voyages and the application of PdF and SPA generate 38 and 114 features, respectively. The results obtained for predicting CO₂ emissions are presented in Figure 5.3, where one can verify that the predictions obtained using PdF lead to better predictions, when compared to the other dictionaries (AV and SPA) for all the regression methods considered. Therefore, it suggests that PdF is able to generate more informative features that better characterize the profiles of the variables over different medium-route voyages, by accounting for their dynamic patterns. Moreover, one can also observe in Figure 5.3 that AV has the worst predictive performance, which again supports

the point that additional information contained in the trajectories can be used for predicting the CO₂ emissions.

The best combination of dictionary and regression method is obtained when an elastic net EN model is developed for features extracted by PdF. This is also confirmed by pairwise comparisons of all methods based on the t -test, which highlights EN as the best method overall, followed by BoRT. EN belongs to the class of penalized regression methods and has the implicit ability to remove irrelevant features (i.e., features that do not convey information about the response). In order to further explore these results, Figure 5.4a presents the features' importance (i.e., the regression coefficients) for the EN model, where one can see the features colored by their type. As can be observed in Figure 5.4a, other features besides the average values (also referred to as mean, feature # 18) are important for predicting CO₂ emissions. In particular, features # 6 (the area of the speed over ground) and # 29 (the variance of the air temperature of starboard engine 2) are the second and third most relevant features. These results show again that more systematic feature extraction methods are beneficial. Accordingly, Figure 5.4b shows that the predicted and observed values of CO₂ emissions for the EN model fall rather close to the 1:1 line.

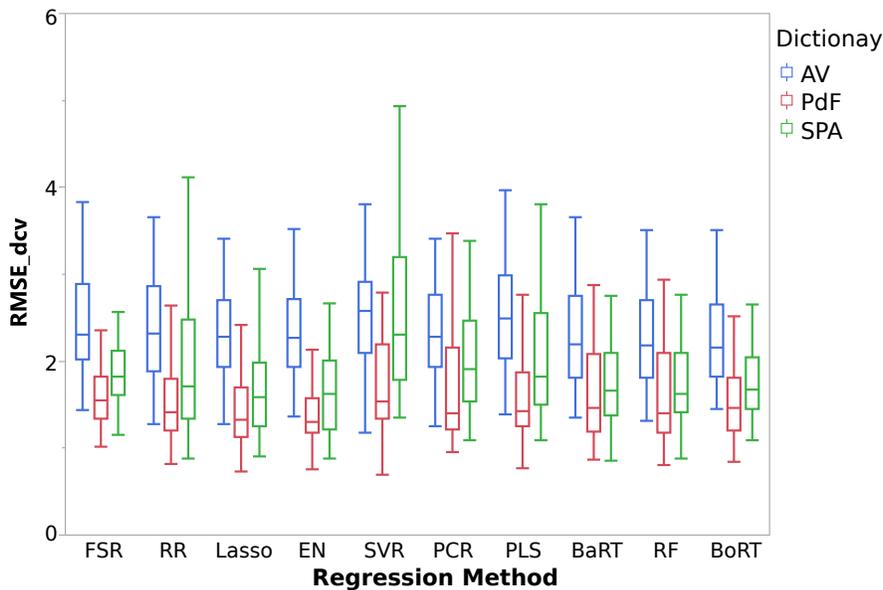


Figure 5.3: Distribution of $RMSE_{dcv}$ for 50 iterations of double cross-validation for the medium route.

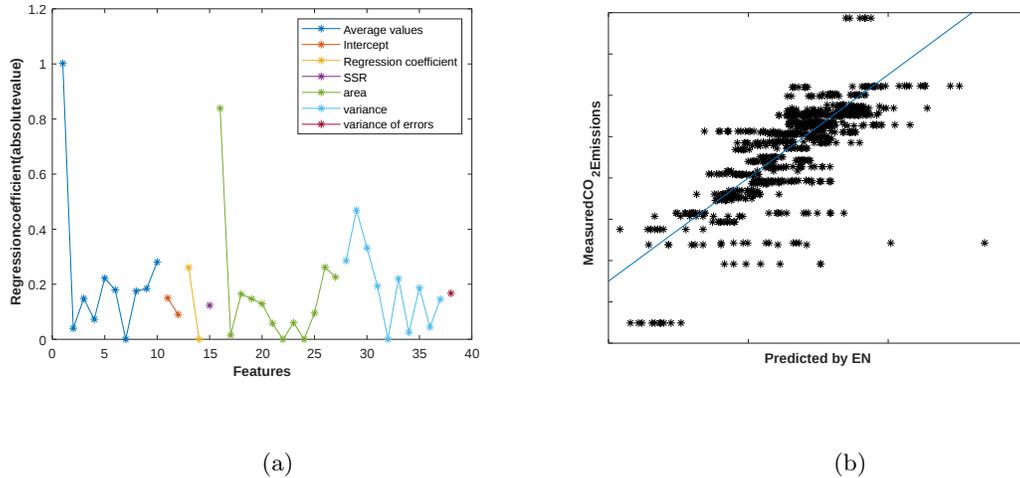


Figure 5.4: Elastic net (EN) modeling results with PdF features for the medium route showing (a) the feature importance for different types of features, and (b) predicted vs measured CO₂ emissions.

Long route

The long route contains $N = 299$ samples and the AV dictionary generates 12 features, whereas PdF and SPA generate 40 and 114 features, respectively. The distribution of $RMSE_{dcv}$ is presented in Figure 5.5 and confirms again that PdF and SPA are significantly better than AV, irrespectively of the regression method considered. This is a considerable difference and shows that using only AV leads to very poor predictions of CO₂ emissions. In contrast, PdF and SPA have lower prediction errors and pairwise comparisons, based on the t -test, identifies the best combinations as SPA with LASSO and PdF with BoRT. Since LASSO models are more parsimonious and simpler to interpret than BoRT, Figure 5.6a shows the regression coefficients for the combination of LASSO and SPA. Figure 5.6a clearly shows that some features are excluded from the model, as their regression coefficients are zero. However, multiple additional features besides the AV contribute to the regression model. In particular, features # 44 (the kurtosis of air temperature of starboard engine 1), # 68 (the covariance between the power difference between port and starboard propeller shafts and Trim), and # 72 (the covariance between power difference between two shaft generators and air temperature of port engine 1) are important for the prediction of CO₂ emissions, even though they are not considered by the AV dictionary. The predicted and observed CO₂ emission values are presented in Figure 5.6b, which displays that the main trends are captured.

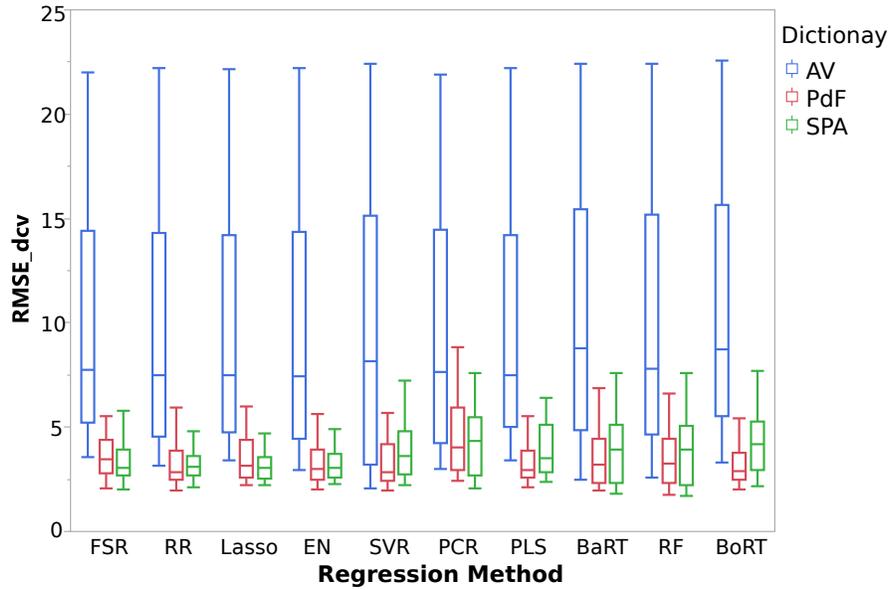
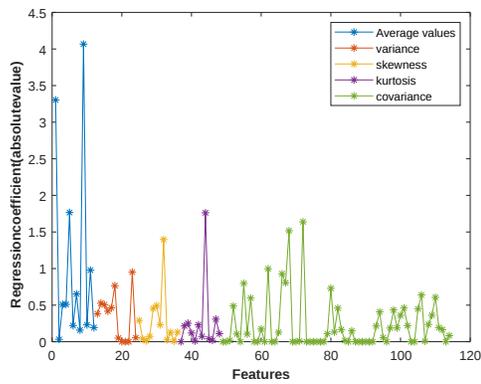
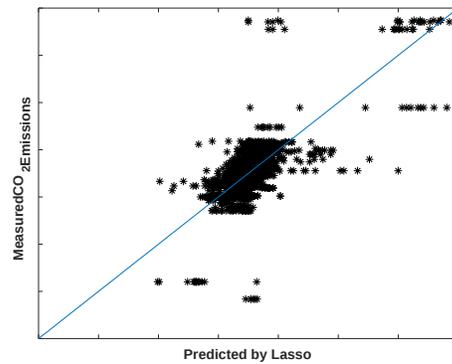


Figure 5.5: Distribution of $RMSE_{dcv}$ for 50 iterations of double cross-validation for the long route.



(a)



(b)

Figure 5.6: LASSO modeling results with SPA features for the long route showing (a) the feature importance for different types of features, and (b) the predicted and measured CO_2 emissions.

5.5 Conclusions

The prediction of CO₂ emissions based on on-board multi-sensor data has the potential to support the shipping industry efforts to comply with current EU regulations. In fact, both shipping companies and policy makers, that are responsible to adopt regulations related to the CO₂ emission monitoring problem, have two compelling needs: better predictions of CO₂ emissions and interpretable models that allow identifying which variables (and which of their features) are important for the prediction. In fact, on the one hand, shipping companies can take advantage of better support to decision-making and, on the other hand, international regulations can adopt stricter requirements if they are able to recognize with lower uncertainty that ships can emit lower CO₂ emissions under appropriate conditions. However, despite the opportunity provided by the availability of data collected from modern on-board multi-sensor systems, many shipping companies still limit their analysis to variables' averages per voyage, due to the intrinsic complexity of the methods needed to handle multivariate profiles, and thus overlook any information contained in the dynamic patterns of variables' trajectories within each of the voyages.

The main contribution of this work, with respect to the applicability of the proposed framework in the maritime engineering, is the possibility to implement solutions that do not require difficult data pre-processing and implementation of complex techniques, while still guaranteeing predictive performance and interpretability.

In this context, this work presented a comparison study of feature-oriented methods that convert the collected data from on-board sensors into features that are potentially useful for the prediction of ship's CO₂ emissions. Three feature-oriented methods were considered: average values (AV, the current standard method), profile-driven features (PdF), and statistical pattern analysis (SPA). By developing regression models, the extracted features were then used as predictors of CO₂ emissions.

A predictive analytics comparison framework (PAC) was applied in order to robustly assess and compare the predictions of CO₂ emissions obtained with AV, PdF, and SPA. PAC contains 4 classes of regression methods that are able to cover many types of relationships between predictors and response variable. The regression methods' performance was assessed by means of the root mean squared error achieved through a double cross-validation ($RMSE_{dev}$) procedure. Each feature dictionary was considered separately. The results obtained showed the advantages of using more systematic feature-oriented methods such as PdF and SPA, and prediction performance increased in almost all the cases considered. In the short route, no method presented acceptable prediction performance because the variability is too small and the main factors contributing to CO₂ emissions are suppressed. However, in the medium and long routes, PdF and SPA achieved satisfactory predictions of CO₂ emissions. For the medium route, PdF combined with an elastic net model provided the best prediction performance; whereas for the long route, SPA combined with Lasso was the best option. The developed models were also investigated with the aim of identifying the most important predictors, which confirmed the advantages of employing feature-oriented methods that extract additional features, besides the variables' averages per voyage.

Chapter 6

Analysis of profiles using partial least-squares methods for ship performance monitoring

Shipping operators are nowadays facing the challenge of monitoring ship performance based on operational data. This is triggered by the compelling air pollution program EU 2015/757 of the European Parliament, that aims from January 2018 to monitoring, reporting, and verification of all harmful emissions of ships operating in the European Economic Area. On the other hand, the continuous acquisition of operational data, which is performed on most of the modern ships, urgently calls for the application of new and opportune statistical methods able to deal with high-dimensional data. Ship operating conditions can be in fact described by sensor signals collected throughout each voyage and stored as profiles. In this paper, the latter are analyzed through multi-way partial least-squares regression of the average fuel consumption per hour over each voyage, which is chosen as scalar performance response, being proportional to harmful emissions. The proposed approach is able to monitor profiles with different length at different voyages. Nevertheless, it is capable of indicating at which instant anomalies may have occurred in ship operating conditions. The proposed approach is shown to be able to furnish clear indications for supporting prognosis of faults. By means of real data acquired from a Ro-Pax cruise ship owned by the shipping company Grimaldi Group, a different multilinear version that explicitly takes into account the three-way structure of the data is also compared with the proposed approach.

6.1 Introduction

The shipping industry is nowadays facing a new regulatory regime that aims to give public access to emissions data. The application of the EU regulation 2015/757 (European Commission 2015), which is mandatory from January 2018, urges shipping companies to set up a system for daily monitoring, reporting, and verification (MRV) of emissions for each ship. Because of the well-known difficulties in the direct measurement of emissions, the MRV program accepts also indirect measurements obtained through the ship fuel consumption. The calculation of greenhouse gas (GHG) emissions can be in fact retrieved by on the basis of the amount of fuel consumption through opportune emission factors (European Commission 2015). Then, the fuel consumption can be regarded as the main performance.

In this paper, we focus on the analysis of the fuel consumption only during the sailing mode, which is known to represent the most relevant part of the total fuel consumption. The most common method used in the naval literature to estimate the fuel consumption concerns the use of the so-called power-speed curves, because a direct proportional relationship also holds between the engine power and the fuel consumption through the power-based specific fuel oil consumption coefficient (see e.g. Corbett & Koehler (2003)). Those curves are obtained by performing a simple regression on the vessel's speed over ground (SOG) collected through dedicated trials under standard reference conditions (Bazari 2007). Unfortunately, they lead to poor predictions in practice, even if some methods that have been proposed to improve the accuracy of these curves by exploiting the information from additional variables (Bialystocki & Konovessis 2016, Perera & Mo 2016) (e.g. acceleration, trim, displacement, wind speed and direction, stabilizer fin operational time, engine operation mode).

The prediction provided by those curves have been outperformed by using more advanced regression techniques (Lepore et al. 2017) that overcome the problem of ill-conditioned data due to collinear variables in high-dimensional systems. However, only some of those techniques are likely to be used for monitoring purposes, such as partial least-squares (PLS) regression which returns also stable predictions in all the cases presented by Lepore et al. (2017) further on and is always comparable with the best method.

At the other end of the industry, digitization and automation are reshaping the ship operation. Modern ships are in fact able to continuously measure and store complex and massive operational data, which have brought management board to call for new methodologies to turn data into value.

Note that all methods appeared so far for monitoring ship performance have been usually limited only to a single representative measurement at each voyage, even though the entire signal of each variable is available. Functional data analysis (FDA) techniques (Ramsay & Silverman 2005) are explored to convert each signal acquired by the multi-sensor system at each voyage into profiles. FDA techniques are effectively applied in several subject areas (Ramsay & Silverman 2005, 2002, Ramsay et al. 2009), but have never been implemented in the maritime field.

In this paper, at each voyage, a profile is obtained from each of the sensor signals that describe ship *operating conditions*. Then, the multi-way partial least-squares (MPLS) approach (Nomikos & MacGregor 1995a, Kourti & MacGregor 1995) is used for monitoring purposes with respect to the average fuel consumption per hour over each voyage, which we denote with FCPH, chosen as scalar *response variable*. In this way, profiles of each variable at different voyages is handled as a multivariate profile. The proposed approach is also compared with a multilinear version of PLS proposed by Bro (Bro 1996, Smilde 1997), which is called three-way partial least-squares (tri-PLS), and naturally takes into account the three-way structure of the data.

Control charts based on the T^2 and the squared prediction error statistics are used to monitor anomalies in ship operating conditions, whereas a prediction error chart will monitor the FCPH.

The rest of the paper is organized as follows. Section 6.2 describes the three main steps for the implementation of the proposed approach and details the differences between MPLS and tri-PLS. Section 6.3 presents a real case study by means of ship navigation data acquired from a Ro-Pax cruise ship owned by the Italian shipping company Grimaldi Group.

6.2 The proposed approach

Note that MPLS has been introduced for monitoring batch processes (Nomikos & MacGregor 1995a), and can be applied only if the number of observations is constant for each combination of batches and variables. Then, in this paper, in order to adapt this approach to profiles, whose length may vary among different voyages, data need to be registered through functional techniques (Ramsay & Silverman 2005). This is also convenient for tri-PLS because even if, in theory, a three-dimensional array allows for different number of observations among voyages, this does not allow to consider voyages as separate observations.

Then, both MPLS and tri-PLS methods require a three-dimensional array $\bar{\mathbf{X}}$ that contains ship operational reference profiles at given domain points with the following three dimensions: the number of replications I , the number of variables J , and the number of evaluation points K . In particular, with reference to the case study mentioned in the introduction, the I replications correspond to the I voyages with the ship sailing in conditions that are considered not exceptional and are selected to build a reference data set. The J physical predictor variables, automatically acquired on-board by multi-sensor system, are assumed to describe adequately the ship operating conditions and to influence the response variable. Lastly, the K evaluation points contain the observations of each variable, at each voyage. Thus, the scalar response variable FCPH is organized into the $(I \times 1)$ vector \mathbf{y} at each voyage.

Once a voyage is completed, the MPLS model is exploited to monitor both ship operating conditions described by the physical predictor variables and FCPH. In particular, ship operating conditions are monitored by means of the control charts based on T^2 , SPE_X , and SPE_k statistics, which will be introduced in Section 6.2. The T^2 and SPE_X statistics give both a single scalar statistic at the end of each voyage, while the SPE_k statistic gives a profile that allows monitoring the physical predictor variables at each point $k = 1, \dots, K$. This allows to determine at which instant an anomaly may have occurred. However, let us explicitly remark that the calculation of the SPE_k requires that the entire signals of the predictor variables have been observed. Then, the monitoring strategy can be applied as soon as data about an entire voyage to be monitored have been collected. This is not a big deal as modern multi-sensor systems are capable to provide fast data storage.

In the case of MPLS, contribution of physical predictor variables to the monitoring statistics can be defined in order to identify the most influential ones when an out of control occurs in one of the corresponding control chart. Otherwise, when all the physical predictor variables are in control at the end of each voyage, a prediction error chart is allowed to be used for directly monitoring the response variable. When the absolute value of the prediction error is large, possible causes are plausibly to be investigated outside the set of predictor variables used to describe ship operating conditions in the model.

Section 6.2 describes the techniques that allow getting the same number of domain points for all profiles (K). Section 6.2 describes how the MPLS and the tri-PLS models are built from those data. Section 6.2 defines the statistics used for constructing the control charts used for monitoring ship performance.

Step 0

As already mentioned, voyages have different duration. Then, in order to obtain the same number K of evaluation points to build the array $\bar{\mathbf{X}}$ both for MPLS and tri-PLS, FDA techniques are required. In order to do this, we may require different repetitions of profiles to

have the same functional domain as in many FDA applications. In the case study presented in Section 6.3 the travel duration varies significantly among voyages. Therefore, a warping function needs to be chosen to allow the comparison of navigation variable measurements over different replications, i.e., different voyages. Warping functions are commonly used in FDA for the registration of functional data (Ramsay & Silverman 2005).

In this paper, the percentage of the total distance traveled at each voyage by the ship is chosen as functional domain and act as a warping function $h_i : [0, T_i] \rightarrow [0, 1]$, where T_i is the sailing time of the i -th voyage, with $i = 1, \dots, I$. At a given domain point, the ship is almost in the same position over different replications of a given route and its operating conditions are reasonably expected to be similar when no anomalies in the ship performance occurred. After profiles are obtained, each observation can be evaluated in K equally spaced points of the domain, this gives the $(I \times J \times K)$ array $\bar{\mathbf{X}}$. This means that the i -th observation, with $i = 1, \dots, I$, is a $(J \times K)$ matrix, that we denote with $\mathbf{X}^{(i)}$.

Step 1

The main difference between MPLS and tri-PLS is the way the two methods deal with $\bar{\mathbf{X}}$. In particular, MPLS unfolds the array $\bar{\mathbf{X}}$ into a large $(I \times JK)$ matrix \mathbf{X} , and then applies standard PLS using \mathbf{y} and \mathbf{X} . On the other hand, tri-PLS makes a tri-linear decomposition of the array $\bar{\mathbf{X}}$ into a set of rank-one cubes. This should improve interpretability and predictions, and reduces the number of parameters to be estimated with respect to MPLS.

MPLS

MPLS is applied by unfolding the array $\bar{\mathbf{X}}$ into a large $(I \times JK)$ matrix \mathbf{X} . This is done by setting the matrix that is extracted from $\bar{\mathbf{X}}$ when the third dimension equal to k , i.e., $\bar{\mathbf{X}}(\cdot, \cdot, k)$, as the columns $kJ + 1, kJ + 2, \dots, kJ + J$ of \mathbf{X} , for $k = 0, \dots, K - 1$. The matrix \mathbf{X} can then be decomposed via the partial least-squares method into a smaller number of R orthogonal score vectors or components $\mathbf{t}_1, \dots, \mathbf{t}_R$, which are observations of latent variables and are arranged in a $(I \times R)$ matrix \mathbf{T} . Each score vector is obtained as the linear combination of the predictor variables having maximal covariance with the response \mathbf{y} . In particular, the first PLS component $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ is such that the weight vector \mathbf{w}_1 solves the optimization problem

$$\max_{\mathbf{w}} [\text{cov}(\mathbf{t}, \mathbf{y}) | \mathbf{t} = \mathbf{X}\mathbf{w} \quad \text{and} \quad \|\mathbf{w}\| = 1]. \quad (6.1)$$

The successive PLS components $\mathbf{t}_r = \mathbf{X}\mathbf{w}_r$, with $2 < r \leq R$, are such that the weight vector \mathbf{w}_r solves the optimization problem

$$\max_{\mathbf{w}} \left[\text{cov}(\mathbf{t}, \mathbf{y}) | \mathbf{t} = \mathbf{X}_{r-1}\mathbf{w} \quad \text{and} \quad \|\mathbf{w}\| = 1, \mathbf{w}^\top \mathbf{w}_j = 0 \quad \forall j = 1, \dots, r-1 \right], \quad (6.2)$$

where \mathbf{X}_{r-1} is the matrix \mathbf{X} orthogonalized with respect to $\mathbf{t}_1, \dots, \mathbf{t}_{r-1}$ as in the following algorithm. This is obtained by means of the nonlinear iterative partial least-squares (NIPALS) algorithm (Geladi & Kowalski 1986) for a single response variable, which is initialized by setting $\mathbf{X}_0 = \mathbf{X}$ and $\mathbf{y}_0 = \mathbf{y}$; then, at each iteration $r = 1, \dots, R$,

1. calculate the weight vector $\mathbf{w}_r = \mathbf{X}_{r-1}^\top \mathbf{y}_{r-1} / (\mathbf{y}_{r-1}^\top \mathbf{y}_{r-1})$;
2. scale \mathbf{w}_r , i.e., $\mathbf{w}_r \rightarrow \mathbf{w}_r / \|\mathbf{w}_r\|$;

3. calculate the score vector as linear combination of \mathbf{X} -columns $\mathbf{t}_r = \mathbf{X}_{r-1}\mathbf{w}_r$;
4. regress \mathbf{y}_{r-1} on \mathbf{t}_r obtaining the coefficient $q_r = \mathbf{y}_{r-1}^\top \mathbf{t}_r / (\mathbf{t}_r^\top \mathbf{t}_r)$;
5. calculate the loading vector $\mathbf{p}_r = \mathbf{X}_{r-1}^\top \mathbf{t}_r / (\mathbf{t}_r^\top \mathbf{t}_r)$;
6. deflate the matrices $\mathbf{X}_r = \mathbf{X}_{r-1} - \mathbf{t}_r \mathbf{p}_r^\top$ and $\mathbf{y}_r = \mathbf{y}_{r-1} - q_r \mathbf{t}_r$ and return to step 1 to find a new component.

Then, \mathbf{X} and \mathbf{y} can be decomposed as

$$\mathbf{X} = \mathbf{TP}^\top + \mathbf{E}, \quad (6.3)$$

$$\mathbf{y} = \mathbf{Tq} + \mathbf{f}, \quad (6.4)$$

where \mathbf{P} is the $(JK \times R)$ matrix of the \mathbf{X} -loadings, whose columns are $\mathbf{p}_1, \dots, \mathbf{p}_R$, $\mathbf{q} = (q_1, \dots, q_R)^\top$ is the $(R \times 1)$ vector of \mathbf{y} -loadings, and $\mathbf{E} = \mathbf{X}_R$ and $\mathbf{f} = \mathbf{y}_R$ are the residual matrices obtained at the R -th iteration, i.e., at the end of the NIPALS algorithm. Moreover, it can be also shown that the matrix \mathbf{T} is given by Phatak & De Jong (1997), Helland (1988)

$$\mathbf{T} = \mathbf{XW}(\mathbf{P}^\top \mathbf{W})^{-1}, \quad (6.5)$$

where \mathbf{W} is the $(JK \times R)$ matrix of the \mathbf{X} -weights, whose columns are $\mathbf{w}_1, \dots, \mathbf{w}_R$. We can also write $\mathbf{T} = \mathbf{XR}$, where $\mathbf{R} = \mathbf{W}(\mathbf{P}^\top \mathbf{W})^{-1}$. This allows to write columns of \mathbf{T} directly as linear combinations of columns of \mathbf{X} (De Jong 1993).

Tri-PLS

The tri-PLS (Bro 1996) method decomposes the array $\overline{\mathbf{X}}$ into R rank-one cubes, each of them ($r = 1, \dots, R$) consists of one score vector \mathbf{t}_r , a first $(J \times 1)$ weight vector \mathbf{w}_r^J , and a second $(K \times 1)$ weight vector \mathbf{w}_r^K such that $\|\mathbf{w}_r^J\| = \|\mathbf{w}_r^K\| = 1$. These weight vectors have the same role of the weight vector \mathbf{w} in standard PLS. In fact, the problem (6.1) can be rewritten as

$$\max_{\mathbf{w}^J, \mathbf{w}^K} \left[\text{cov}(\mathbf{t}, \mathbf{y}) | t_i = \sum_{j=1}^J \sum_{k=1}^K x_{ijk} w_j^J w_k^K \quad \text{and} \quad \|\mathbf{w}^J\| = \|\mathbf{w}^K\| = 1 \right], \quad (6.6)$$

where x_{ijk} is the element (i, j, k) of the array $\overline{\mathbf{X}}$, w_j is the j -th element of the vector \mathbf{w}^J and w_k is the k -th element of the vector \mathbf{w}^K . The trilinear PLS algorithm (Bro 1996) is initialized by setting $\overline{\mathbf{X}}_0 = \overline{\mathbf{X}}$ and $\mathbf{y}_0 = \mathbf{y}$. Then, for each $r = 1, \dots, R$,

1. calculate the $(J \times K)$ matrix \mathbf{Z} whose (j, k) element is given by $z_{jk} = \sum_{i=1}^I y_i x_{ijk}$, where y_i is the i -th element of the vector \mathbf{y}_{r-1} and x_{ijk} is the element (i, j, k) of the array $\overline{\mathbf{X}}_{r-1}$;
2. perform singular value decomposition on \mathbf{Z} , then define \mathbf{w}_r^J and \mathbf{w}_r^K as the left and right singular vectors of \mathbf{Z} , respectively;
3. calculate the $(I \times 1)$ vector $\mathbf{t}_r = (t_{1,r}, \dots, t_{I,r})^\top$ such that its i -th element is $t_i = \sum_{j=1}^J \sum_{k=1}^K x_{ijk} w_j^J w_k^K$, where w_j^J is the j -th element of \mathbf{w}_{r-1}^J and w_k^K is the k -th element of \mathbf{w}_{r-1}^K ;

4. define the $(I \times r)$ matrix \mathbf{T}_r by column as $\mathbf{T}_r = (\mathbf{t}_1, \dots, \mathbf{t}_r)$;
5. calculate $\mathbf{b} = (\mathbf{T}_r^\top \mathbf{T}_r)^{-1} \mathbf{T}_r^\top \mathbf{y}$;
6. replace $\overline{\mathbf{X}}_{r-1}$ with the array $\overline{\mathbf{X}}_r$ whose each i -th observation, with $i = 1, \dots, I$, is obtained as $\mathbf{X}_r^{(i)} = \mathbf{X}_{r-1}^{(i)} - t_{i,r} \mathbf{w}_r^J (\mathbf{w}_r^K)^\top$, where $\mathbf{X}_{r-1}^{(i)}$ is the i -th observation of $\overline{\mathbf{X}}_{r-1}$, moreover replace \mathbf{y}_{r-1} with $\mathbf{y}_r = \mathbf{y}_{r-1} - \mathbf{T}_r \mathbf{b}$ and return to step 1 to find a new component.

At the end of the algorithm, the score matrix can be defined as $\mathbf{T} = \mathbf{T}_R$.

There are two main differences between MPLS and tri-PLS. On one hand, tri-PLS allows to estimate a lower number of weights than MPLS. In fact, for each component $r = 1, \dots, R$, the number of weights estimated with MPLS is JK , while for tri-NPLS it is $J + K$. On the other hand, an unfortunate aspect is that, unlike MPLS, the scores obtained with tri-PLS are not orthogonal. This makes calculating the contribution of the predictor variables to the Hotelling T^2 statistic, which is shown for MPLS in Section 6.2, more complicated.

Note also that, both for MPLS and tri-PLS, data are scaled prior to perform the analysis by subtracting from observations of each variable the sample mean, and dividing them by the sample standard deviation. The scaling procedure is the same for both methods, that is each value is scaled with respect to the sample mean and sample standard deviation of the corresponding column of the matrix obtained after unfolding the array $\overline{\mathbf{X}}$.

Step 2

In order to implement the monitoring strategy, for a new voyage, a $(J \times K)$ matrix $\mathbf{X}^{(0)}$ of observations of the predictor variables is available. This can be unfolded to a $(JK \times 1)$ vector of observations of the predictor variables, which then can be scaled to obtain $\mathbf{x}_0 = (x_{0,1}, \dots, x_{0,JK})^\top$ by dividing each value with respect to the mean and standard deviation of the corresponding column in the reference data. The corresponding observation of the response variable is stored in the scalar y_0 .

MPLS

In order to apply MPLS to a new observation, we can calculate the scores $\mathbf{t}_0 = (t_{0,1}, \dots, t_{0,R})^\top$ as

$$\mathbf{t}_0^\top = \mathbf{x}_0^\top \mathbf{R}. \quad (6.7)$$

The predictor variables of forthcoming voyages can then be monitored by the T^2 , SPE_X , and SPE_k statistics defined as follows.

The Hotelling T^2 statistic is defined as (Kourti & MacGregor 1995, 1996, Jackson 2005)

$$T^2 = \sum_{i=1}^R t_{0,i}^2 / \hat{\lambda}_i, \quad (6.8)$$

where $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_R \geq 0$ are the eigenvalues of the covariance matrix of the scores, which is orthogonal and is calculated as $(I - 1)^{-1} \mathbf{T}^\top \mathbf{T} = \text{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_R)$. The upper control limit with a significance level α for the T^2 statistic is given by Kourti & MacGregor (1995)

$$T_{\text{lim},\alpha}^2 = \frac{R(I^2 - 1)}{I(I - R)} F_\alpha(R, I - R), \quad (6.9)$$

where $F_\alpha(R, I - R)$ is the $(1 - \alpha)$ -quantile of the Fisher distribution with R and $I - R$ degrees of freedom.

The squared prediction error statistic (SPE_X) (Nomikos & MacGregor 1995a) for residuals in the predictor variable space at each voyage is given by

$$SPE_X = \mathbf{e}_0^\top \mathbf{e}_0, \quad (6.10)$$

where the $(JK \times 1)$ vector $\mathbf{e}_0 = (e_{0,1}, \dots, e_{0,JK})^\top = \mathbf{x}_0 - \mathbf{P}\mathbf{t}_0$ contains the corresponding \mathbf{X} -residuals. Moreover, the SPE_X statistic can be specialized to a single instantaneous evaluation point k (i.e., a given percentage of distance travelled) as

$$SPE_k = \sum_{c=(k-1)J+1}^{kJ} e_{0,c}^2. \quad (6.11)$$

The SPE_k statistic represents the perpendicular distance of the instantaneous ship operating condition measurements from the reduced predictor variable space obtained based on the reference data, and then is able to clearly detect problems at a specific point k .

The upper control limit at a significance level α for the SPE_X statistic (6.10) is given by Nomikos & MacGregor (1995b)

$$SPE_{\text{lim},\alpha} = \theta_1 \left[\frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} + 1 \right]^{\frac{1}{h_0}}, \quad (6.12)$$

where $\theta_1 = \text{trace}(\mathbf{V})$, $\theta_2 = \text{trace}(\mathbf{V}^2)$, and $\theta_3 = \text{trace}(\mathbf{V}^3)$, with $\mathbf{V} = (I - 1)^{-1} \mathbf{E} \mathbf{E}^\top$, while $h_0 = 1 - 2\theta_1 \theta_3 / 3\theta_2^2$. The upper control limit of the SPE_k statistic is obtained in the same way as for SPE_X , where θ_1 , θ_2 , θ_3 , and \mathbf{V} are calculated by replacing the matrix \mathbf{E} with the $(I \times J)$ matrix \mathbf{E}_k , whose columns are the columns from $(k - 1)J + 1$ to kJ of \mathbf{E} .

When, at a given voyage a monitoring statistic lies outside the corresponding control limit, detailed information can be obtained about plausible causes of anomalies by exploring the contribution of the j -th individual variable to the T^2 statistic (6.8). Since scores are orthogonal, such contributions can be defined as

$$\text{Contribution}_{T^2,j} = \sum_{k=0}^{K-1} \sum_{i=1}^R \frac{t_{0,i}}{\hat{\lambda}_i} x_{0,j+kJ} r_{j+kJ,i}, \quad (6.13)$$

where $r_{a,b}$ is the (a, b) entry of the matrix \mathbf{R} . The contribution is obtained by considering that the T^2 statistic can be decomposed as

$$T^2 = \sum_{i=1}^R \frac{t_{0,i}^2}{\hat{\lambda}_i} = \sum_{i=1}^R \frac{t_{0,i}}{\hat{\lambda}_i} \sum_{p=1}^{JK} x_{0,p} r_{p,i} = \sum_{p=1}^{JK} \left(\sum_{i=1}^R \frac{t_{0,i}}{\hat{\lambda}_i} x_{0,p} r_{p,i} \right). \quad (6.14)$$

Then, terms in the first sum can be grouped by each physical predictor variable, because the columns $j + kJ$, with $k = 0, \dots, K - 1$, give the values of the j -th physical predictor variable at each k . From those terms we get

$$T^2 = \sum_{j=1}^J \left(\sum_{k=0}^{K-1} \sum_{i=1}^R \frac{t_{0,i}}{\hat{\lambda}_i} x_{0,j+kJ} r_{j+kJ,i} \right). \quad (6.15)$$

This allows to interpret the T^2 statistic as the sum of contributions of the physical predictor variables, defined in Equation (6.13).

Accordingly, one can examine the contribution of the j -th individual variable to the SPE_X statistic through

$$\text{Contribution}_{SPE_X, j} = \sum_{k=0}^{K-1} e_{0, j+kJ}^2. \quad (6.16)$$

If a forthcoming voyage shows no anomalies, i.e., the monitoring statistics do not exceed control limits, the scalar performance indicator can be monitored by means of approximate intervals for the prediction error of the FCPH $y_0 - \mathbf{t}_0^\top \mathbf{q}$, whose limits can be calculated as (Nomikos & MacGregor 1995b)

$$\pm t_{I-R-1, \alpha/2} \hat{\sigma} \sqrt{1 + \mathbf{t}_0^\top (\mathbf{T}^\top \mathbf{T})^{-1} \mathbf{t}_0}, \quad (6.17)$$

where $\hat{\sigma} = \mathbf{f}^\top \mathbf{f} / (I - 1)$, and $t_{I-R-1, \alpha/2}$ is the $(1 - \alpha/2)$ -quantile of a Student distribution with $I - R - 1$ degrees of freedom.

Tri-PLS

Given a new observation $\mathbf{X}^{(0)}$ and the corresponding unfolded vector \mathbf{x}_0 , it is possible to calculate the tri-PLS scores $\mathbf{t}_0 = (t_{0,1}, \dots, t_{0,R})^\top$ as (Smilde 1997)

$$\mathbf{t}_0^\top = \mathbf{x}_0^\top \mathbf{V}, \quad (6.18)$$

where

$$\mathbf{V} = [\mathbf{w}_1 | (\mathbf{I} - \mathbf{w}_1 \mathbf{w}_1^\top) \mathbf{w}_2 | \dots | (\mathbf{I} - \mathbf{w}_1 \mathbf{w}_1^\top) (\mathbf{I} - \mathbf{w}_2 \mathbf{w}_2^\top) \dots (\mathbf{I} - \mathbf{w}_{R-1} \mathbf{w}_{R-1}^\top) \mathbf{w}_R], \quad (6.19)$$

with $\mathbf{w}_r = \mathbf{w}_r^J \otimes \mathbf{w}_r^K$ indicating the $(JK \times 1)$ vector obtained as the Kronecker product between \mathbf{w}_r^J and \mathbf{w}_r^K . We can compare the Hotelling T^2 statistic obtained with MPLS as described in Section 6.2 with the one defined for tri-PLS

$$T^2 = \frac{1}{I-1} \mathbf{t}_0^\top (\mathbf{T}^\top \mathbf{T})^{-1} \mathbf{t}_0, \quad (6.20)$$

where \mathbf{T} is the matrix of scores obtained with the tri-PLS algorithm. The upper control limit is given by Equation (6.9) as for MPLS.

We also compare the squared prediction error statistic (SPE_X) for residuals in the predictor variable space at each voyage, obtained for MPLS as described in Section 6.2 with the one defined for tri-PLS as

$$SPE_X = \mathbf{e}_0^\top \mathbf{e}_0, \quad (6.21)$$

where the vector $\mathbf{e}_0^\top = (e_{0,1}, \dots, e_{0,JK}) = \mathbf{x}_0^\top - \mathbf{t}_0^\top \mathbf{W}^\top$, with $\mathbf{W} = [\mathbf{w}_1^K \otimes \mathbf{w}_1^J | \dots | \mathbf{w}_R^K \otimes \mathbf{w}_R^J]$. The upper limit for this statistic is given by (6.12) as for MPLS, where the residual matrix used to calculate $\theta_1, \theta_2, \theta_3$ is $\mathbf{E} = \mathbf{X} - \mathbf{T} \mathbf{W}^\top$.

Finally, for voyages that show no anomalies among the predictor variables, the prediction error given by MPLS can be compared with the one given by tri-PLS as $y_0 - \mathbf{t}_0^\top \mathbf{b}$, where \mathbf{b} is the coefficient obtained from the regression of \mathbf{y} on \mathbf{T} in the tri-PLS algorithm.

6.3 Real case study

The proposed approach has been applied to the real operational data mentioned in the introduction in order to illustrate how the proposed approach can also furnish clear indications for supporting prognosis of faults. The name of the ship, route and voyage dates are omitted for confidentiality reasons. The multi-sensor system continuously acquires data for each variable, which are stored with five minute frequency in the raw data set used to illustrate the MPLS approach. A route has been analyzed such that every observation corresponds to a single voyage of the ship. As anticipated in the introduction, only data collected during the sailing mode have been considered. In particular, this operational mode has been automatically detected as reported by Zaman et al. (2017)

In particular, a detailed description of the ship navigation data is addressed to Section 6.3, while the results obtained through both MPLS and tri-PLS are compared in Section 6.3. Section 6.4 draws conclusions and future directions.

Ship navigation data

For each voyage, $J = 7$ variables stored as profiles have been considered to describe the ship operating conditions and to be used for the prediction of FCPH, which is the scalar response:

1. SOG [kn], V ;
2. acceleration over ground [kn/s], A ;
3. power difference between port and starboard propeller shafts [kW], ΔP ;
4. power difference between port and starboard shaft generators [kW], ΔSG ;
5. longitudinal wind [kn], W_L ;
6. side wind [kn], W_S ;
7. distance from the nominal route [NM], $Dist$.

The first variable is the SOG, which is known as the most important predictor of the FCPH (Bialystocki & Konovessis 2016, Bocchetti et al. 2015, Erto et al. 2015, Lepore et al. 2017). Moreover, FDA recommends to explore instantaneous information not only about variables, but also about their derivatives, as they often give useful additional information. Then, the SOG derivative (i.e., the acceleration over ground), can be used as additional predictor. Note that the acceleration is obtained as the derivative of the functional SOG with respect to the time at first. Then, the functional observations of acceleration are warped by the percentage of sailed distance $[0, 1]$ as for the other variable profiles.

The power difference between port and starboard propeller shafts is considered in order to discover any anomaly in the main engines for propulsion, whereas the power difference between port and starboard shaft generators can guide the identification of the navigation modes (viz., combinator and constant).

The longitudinal and the side wind are obtained from the anemometer sensor and have the potential of accounting for the influence of weather conditions on the ship performance.

The last predictor variable is the distance of the ship from the nominal route, which is the route recommended by the ship captain in normal conditions.

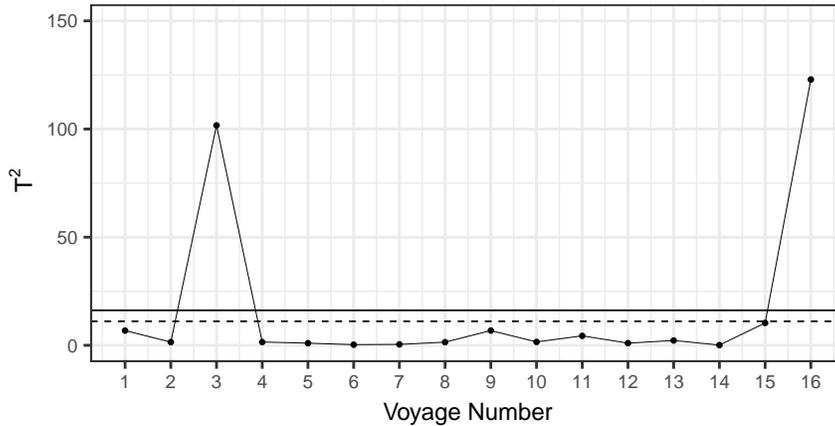


Figure 6.1: Hotelling T^2 control chart obtained with MPLS for monitoring voyages, with 95% and 99% upper control limits (dashed and solid line, respectively).

Analysis and interpretation

Functional data are obtained via 100 B-spline basis functions with equally spaced knots and a roughness penalty on the second derivative of 10^{-10} , which has shown to produce good results for all the functional variables. In theory, a different value for both the number of basis functions and the roughness penalty should be provided for each functional variable. One of the most common criteria used to choose the parameters is the minimisation of the generalized cross-validation (GCV) (Ramsay & Silverman 2005). However, since GCV values often change slowly near the minimizing value, a fairly wide range of the roughness values may give the same GCV in practice, as in this case.

Each functional variable has been evaluated in $K = 100$ equally spaced domain points to get the three-dimensional array $\bar{\mathbf{X}}$. The MPLS and tri-PLS approaches have been then applied to a set of $I = 52$ reference voyages. A single run of 10-fold cross-validation procedure based on the PRESS statistic (Hastie et al. 2009) has been carried out for both MPLS and tri-PLS. Note also that the data partition into the ten segments used for the 10-fold cross-validation is the same for both approaches. The number of latent variables selected for MPLS by cross-validation is $R = 4$. The coefficient of determination is equal to 0.82 and confirms the model is able to adequately predict the FCPH at each voyage. The tri-PLS approach has also been applied to the same set of voyages and the number of latent variables selected through cross-validation is 2.

Ship performance has been then monitored on 16 successive voyages. By using MPLS, Figure 6.1 shows the T^2 statistic for the monitoring voyages. Unusual operating conditions are highlighted by the control chart for voyages 3 and 16. The SPE_X statistic in Figure 6.2 points out the same voyages out of control. The performance of these two charts is compared with the results obtained below via tri-PLS. Figure 6.3 shows the Hotelling T^2 control chart obtained with tri-PLS for the same 16 monitoring voyages. The results are very similar in practice to those obtained for MPLS in Figure 6.1. Some differences in the control limits are due to the fact that tri-PLS chose a higher number of components. As in Figure 6.2, the SPE_X control chart obtained with tri-PLS in Figure 6.4 shows that all the voyages but 3 and 16 are in control. Therefore, in this case we can conclude that, for monitoring purposes,

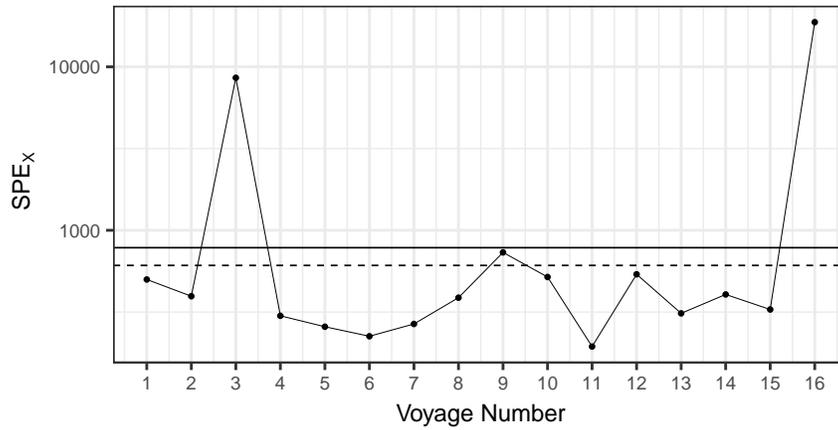


Figure 6.2: SPE_X control chart obtained with MPLS for monitoring voyages, with 95% and 99% upper control limits (dashed and solid line, respectively)

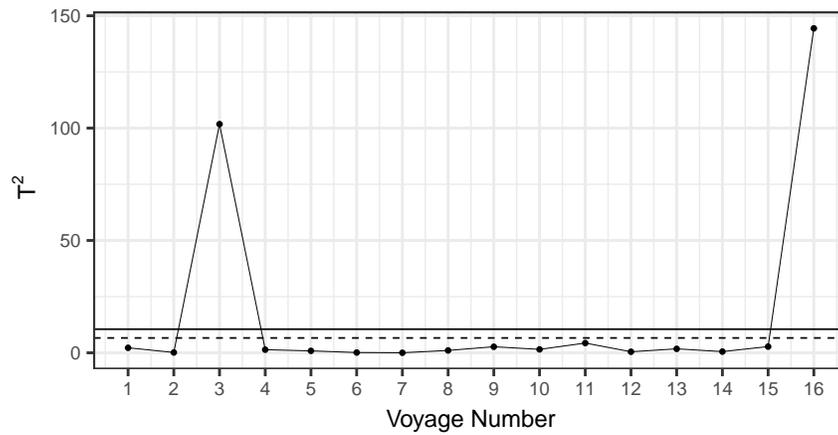


Figure 6.3: Hotelling T^2 control chart obtained with tri-PLS for monitoring voyages, with 95% and 99% upper control limits (dashed and solid line, respectively).

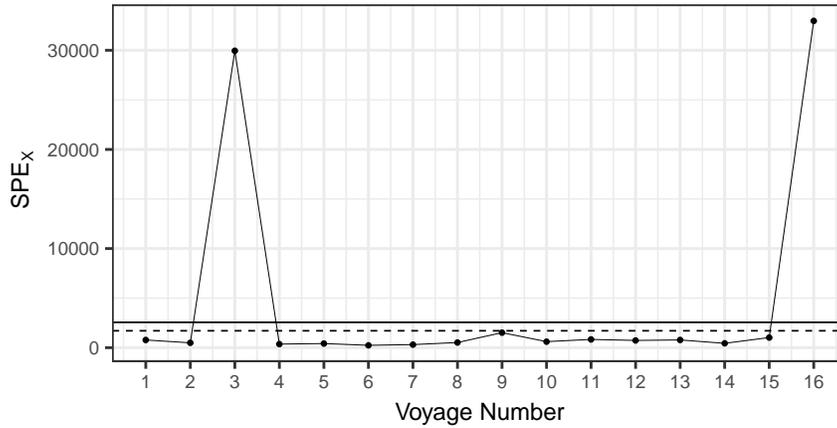


Figure 6.4: SPE_X control chart obtained with tri-PLS for monitoring voyages, with 95% and 99% upper control limits (dashed and solid line, respectively.)

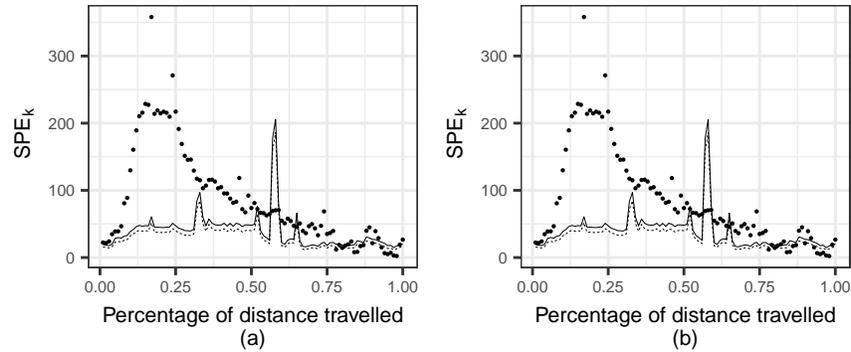


Figure 6.5: SPE_k control chart for one in control voyage 4 (a) and one out of control voyage 3 (b), with 95% and 99% upper control limits (dashed and solid line, respectively).

both the considered PLS methods furnish similar results and conclusions.

The voyages identified as anomalous can be further investigated. In fact, the SPE_k control chart obtained via MPLS allows examining individually each single voyage, as shown in Figure 6.5, which illustrates the use and interpretation of the SPE_k statistic.

In particular, Figure 6.5 (a) shows the SPE_k control chart for the voyage 4, which is in control, i.e., does not show anomalies in the SPE_k statistic at any instant k . Whereas, Figure 6.5 (b) shows the SPE_k control chart for the voyage 3, which has been pointed out to be out of control in the SPE_X control chart (Figure 6.2). For this voyage, most of the points of the SPE_k statistic fall above the 99% upper control limit, and especially in the first part of the voyage, the statistic has a very large value. The anomalous voyage 3 can be checked against the reference model to determine the reason for their difference. This can be investigated by using the contribution plots, which are shown in Figure 6.6. The contributions to the T^2 statistic (Figure 6.6 (a)) show that the main variable responsible for

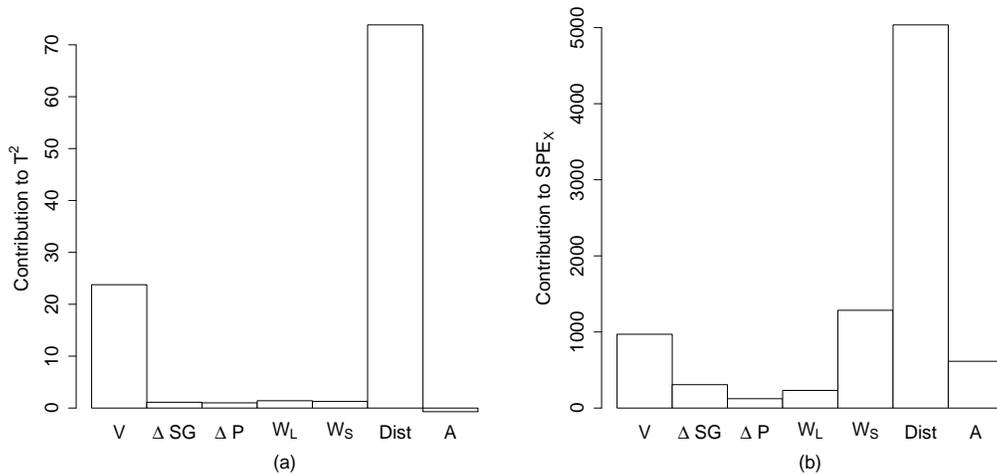


Figure 6.6: Contribution of the variables to the T^2 statistic (a) and to the SPE_X statistic for the voyage 3.

the out of control is the distance from the mean route, even though the SOG has also a large contribution. The distance from the mean route is the most influencing variable also for the out of control of the SPE_X statistic (Figure 6.6 (b)). Other less influencing variables are the SOG and the side wind. The contribution plots allow to return to the physical variable space and to identify the physical predictor variables to be investigated.

Figure 6.7 shows the observations of these functional variables against the reference ones. In this case, the ship has sailed with a different route than the nominal one (Figure 6.7 (a)). Moreover, the SOG is particularly smaller than that of the reference voyages, especially in the first part of the voyage (Figure 6.7 (b)). This is in accordance with the behaviour of the SPE_k graph drawn in Figure 6.5 (b). The side wind (Figure 6.7 (c)) does not show very anomalous behaviour, apart from a larger-than-average values in the middle of the voyage.

The voyage 16 is also out of control in both T^2 and SPE_X statistics, as shown in Figure ???. The SPE_k control chart can then be used for evaluating in which parts of the voyage anomalous conditions may have occurred. In Figure 6.8 the SPE_k statistic is shown to fall above the upper 99% control limit over most of the voyage, but in the central part, between domain points 0.3 and 0.6, achieves very large values. This plausibly indicates that the main anomaly has occurred in the middle of the voyage.

Contribution plots shown in Figure 6.9 can be used to identify the possible causes. Note that in voyage 3 the contribution to the two statistics indicated the same variables as responsible for the out of control. This could mislead to the conclusion that only one control chart is sufficient to implement a complete multivariate monitoring strategy. This is not true for latent variable models, for which it is necessary to look into both the reduced latent variable space—by means of the T^2 statistic—and the residual space—by means of the SPE_X statistic. This is confirmed for the voyage 16 and in fact Figure 6.9 (b) reveals the acceleration variable as one of the biggest responsible for the out of control, whereas Figure 6.9 (a) is not able to detect any acceleration problem.

We can summarize the information of both contribution plots, which indicate the SOG,

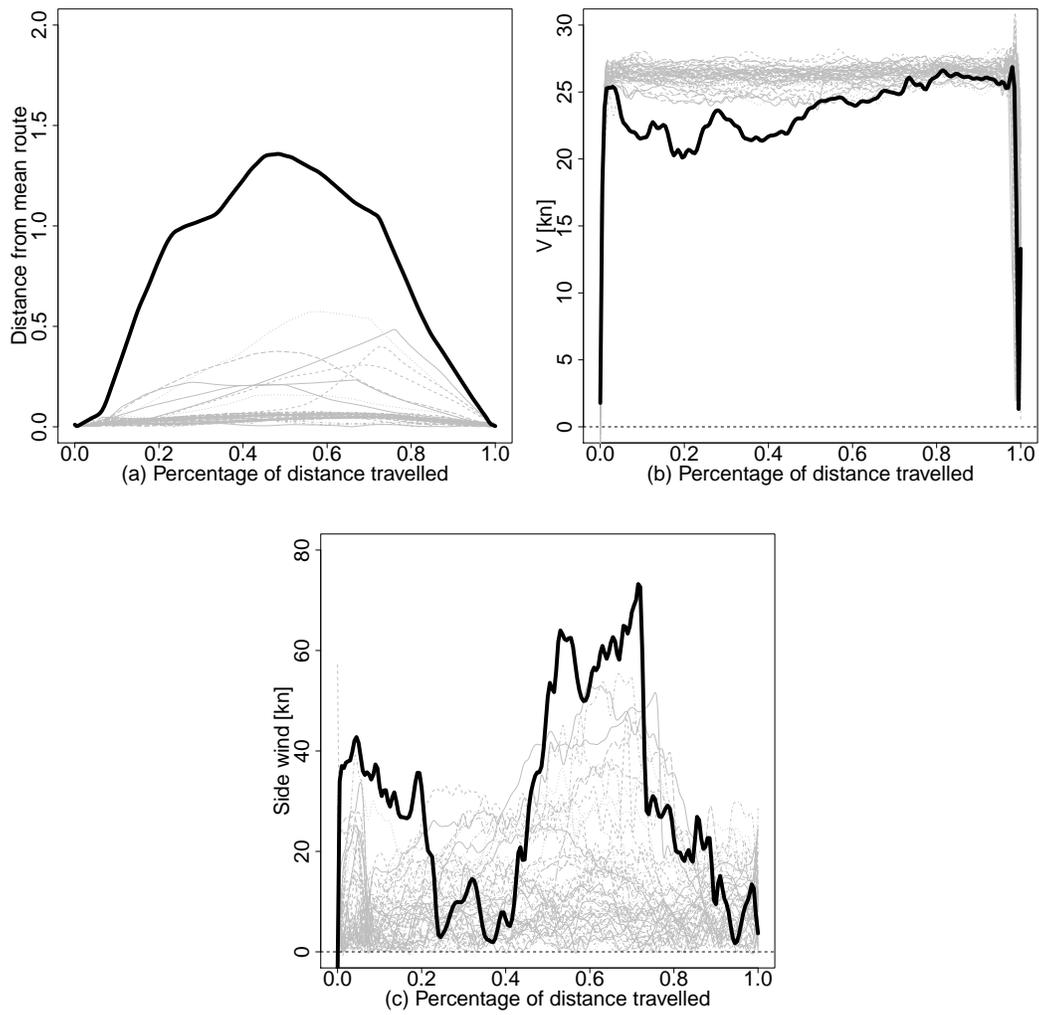


Figure 6.7: Functional observation of distance from mean route (a), SOG (b), and side wind (c) variables for voyage 3 against reference voyages. Note that in each plot the bold line refers to the voyage 3, while the grey lines are the observations of the corresponding variable for the 52 reference voyages.

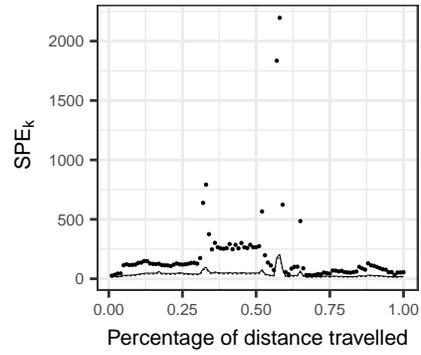


Figure 6.8: SPE_k control chart for voyage 16, with 95% and 99% upper control limits (dashed and solid line, respectively).

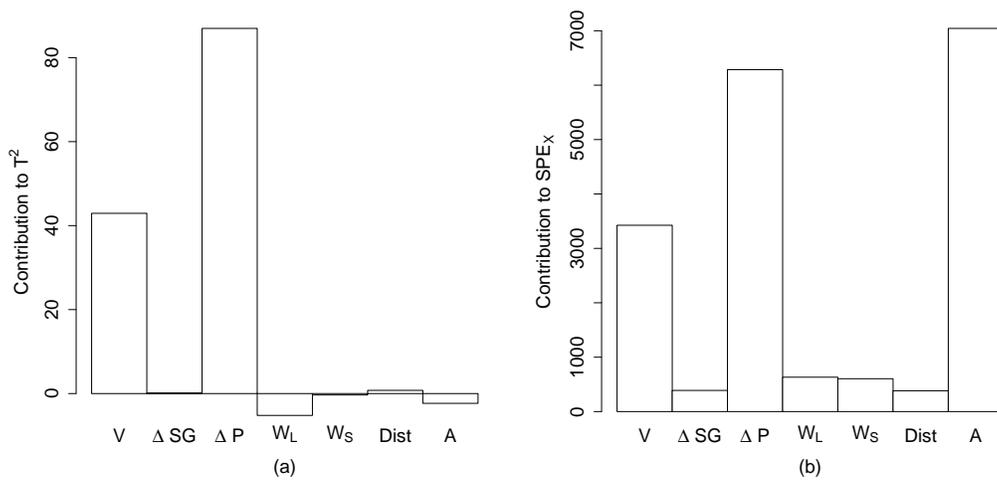


Figure 6.9: Contribution of the variables to the T^2 statistic (a) and to the SPE_x statistic for the voyage 16.

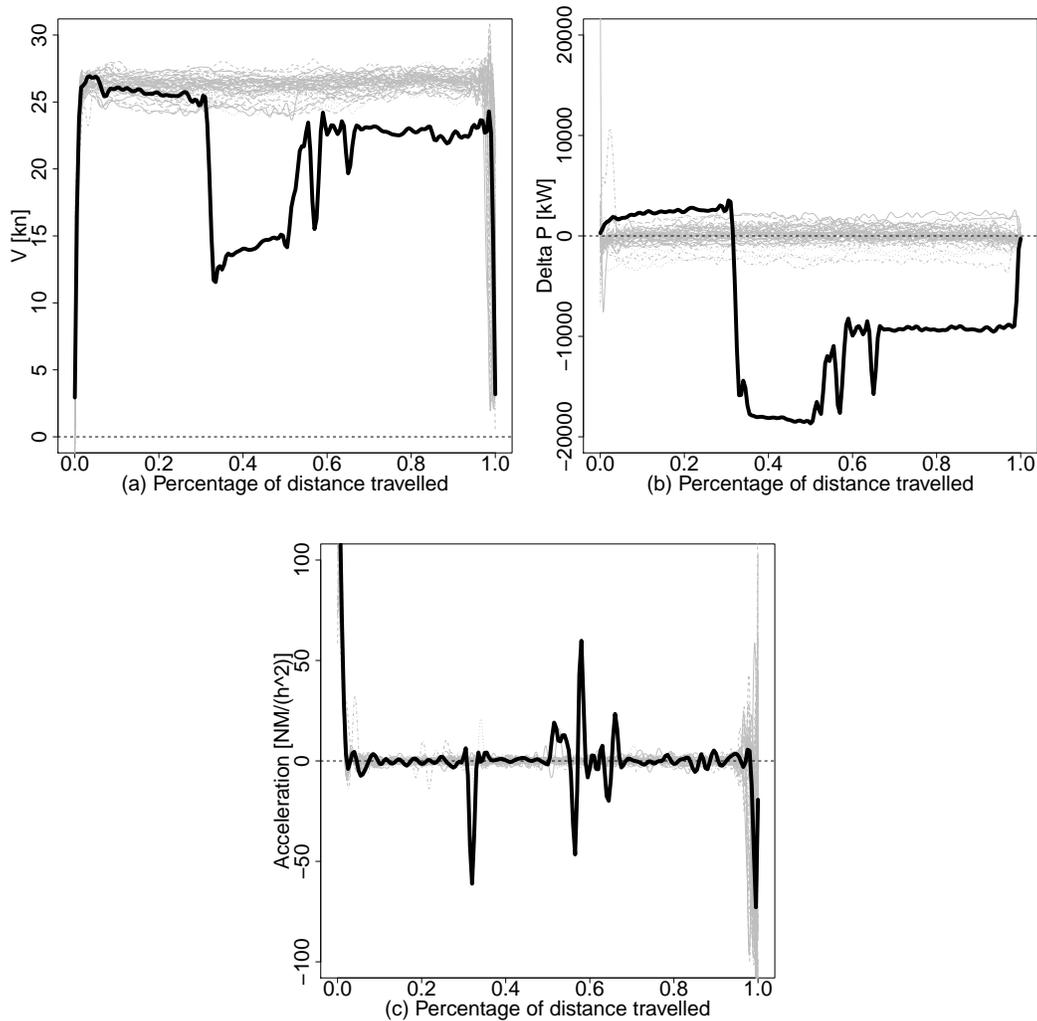


Figure 6.10: Functional observation of SOG (a), power difference between port and starboard propeller shafts variables (b), and acceleration variables for voyage 16 against reference voyages. Note that in each plot the bold line refers to the voyage 16, while the grey lines are the observations of the corresponding variable for the 52 reference voyages.

the power difference between port and starboard propeller shafts, and the acceleration as the variables to investigate. Figure 6.10 shows the plots of these variables. Accordingly with the SPE_k plot in Figure 6.8, it is to note that the voyage is characterized by a first part, which corresponds to domain point values between 0 and about 0.3, where the profiles are in accordance with reference data. Whereas, at a domain point equal to about 0.3, the SOG suddenly decreases. This corresponds to a deceleration which is pointed out. At the same time, the graph of the power difference between port and starboard propeller shafts indicates a big departure from zero. This was due to a malfunctioning in one engine on board. After

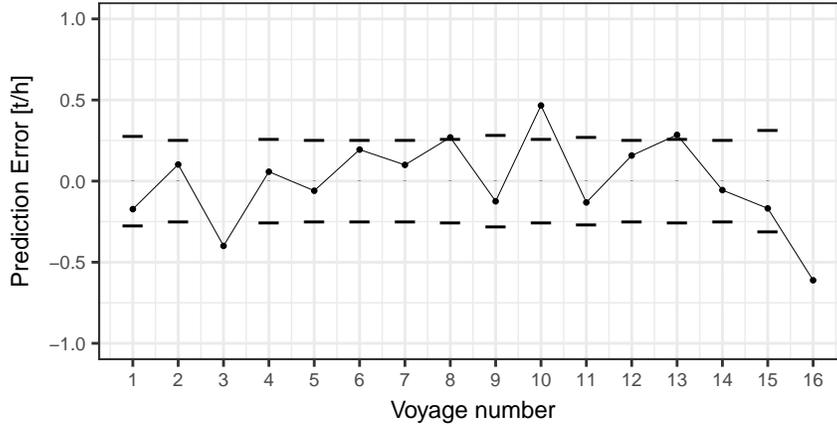


Figure 6.11: Prediction error control chart.

some time, starting from domain point 0.5, the power difference between port and starboard propeller shafts decreases with local changes and this turned into a higher SOG and large acceleration values. However, for the rest of the voyage the profiles still result to be quite different from the reference data. This is also in accordance with the SPE_k plot in Figure 6.8.

For those voyages where the monitoring statistics of the predictor variable space, i.e., T^2 , SPE_X , and SPE_k , are out of control, the prediction of the response variable through the MPLS model cannot be considered. When instead voyages do not show problems in the predictor variables, then the FCPH can be monitored through the prediction error chart. Figure 6.11 illustrates the prediction error chart for the monitoring voyages considered in the proposed case study. In this chart, the prediction error limits for voyages 3 and 16 are not shown, because they have shown anomalies in the predictor variable space. Note that voyages that are in control in the predictor variable space, for which the prediction error falls outside the limits given by Eq. 6.17, require further investigation on those variables that have not been considered in the MPLS model.

Finally, the predictive ability is compared between MPLS and tri-PLS. Since the reference data have been used as training set to select the number of latent variables via cross-validation and fit the model, the monitoring data can be considered as the new test set on which the model selection can be performed (Hastie et al. 2009). The sum of squared prediction error on the voyages that are in control in the predictor variable space is $0.556 t^2/h^2$ for MPLS and $0.865 t^2/h^2$ for tri-PLS. This clearly shows that for the considered real case study the MPLS outperforms tri-PLS in FCPH prediction.

6.4 Conclusion

To respond to the compelling air pollution programs, shipping companies are nowadays setting-up on their fleets modern sensor systems that stream massive amounts of operational data to remote servers. Motivated by this context, the statistical process monitoring approach based on multi-way partial least-squares method has been combined with techniques popular in the functional data analysis context in order to allow monitoring profiles with a different

length. This is in fact the case of operating conditions collected over different voyages that, in general, have different travel duration or distance traveled. All profiles then needed to be warped through the percentage of traveled distance with respect to the nominal route. In such a way, variable profiles can be used to build a multi-way partial least-squares model both for monitoring of ship operating conditions and for predicting the average fuel consumption per hour. The proposed approach has also been compared to the trilinear PLS, that takes into account the three-way structure in the data. A real case study is presented to illustrate that the proposed procedure is able to furnish adequate predictions and to indicate if and when anomalies occur. The squared prediction error statistic evaluated at a single domain point gives clear indications in this regard. This would have not been feasible through statistical models built using a single variable observation per each voyage.

The MPLS approach allows shipping operators to monitor complex multi-sensor data and to explore the possibility of comparing ship performance under different operating conditions.

The results showed that while trilinear PLS offered parsimony in the number of parameters to be estimated, the proposed approach has a higher prediction ability. Moreover, MPLS allows to obtain orthogonal scores and to calculate contribution to the Hotelling T^2 statistic.

Future research should be addressed to build a model able to deal with incomplete data and thus, to perform real-time monitoring before the completion of a voyage. In fact, the proposed approach requires that the entire voyage to be monitored is completed to allow obtaining the score observations needed to calculate the monitoring statistics and to predict the response variable. A suitable completion method would then allow to perform real-time monitoring and prediction.

Chapter 7

Monitoring ship operating conditions and CO₂ emissions using control charts based on scalar-on-function regression

To respond to the compelling air pollution programs, shipping companies are nowadays setting-up on their fleets modern multi-sensor systems that stream massive amounts of observational data, which can be considered as varying over a continuous domain. Motivated by this context, a novel procedure is proposed that extends classical multivariate techniques to the monitoring of multivariate functional data and a scalar quality characteristic related to them. The proposed procedure is shown to be also applicable in real time and is illustrated by means of a real-case study in the maritime field on the continuous monitoring of operating conditions (i.e., the multivariate functional data) and total CO₂ emissions (i.e., the scalar quality characteristic) at each voyage of a cruise ship. The real-time monitoring is particularly helpful for promptly supporting managerial decision making by indicating if and when an anomaly occurs during the navigation.

7.1 Introduction

In many statistical process control (SPC) applications, the quality characteristic to be monitored is influenced by one or more explanatory variables (referred to also as *covariates*). The regression control chart idea, classically addressed by Mandel (1969), where a scalar quality characteristic is affected by a single scalar variable, is nowadays exacerbated by the availability of massive amounts of data stored from multiple sources. This increases the complexity and dimension of the information and calls for extensions of the classical statistical methods toward new mathematical settings. In this perspective, Nomikos & MacGregor (1995*a,b*) and Kourti & MacGregor (1996) have introduced methods in SPC of batch processes to address the problem of dimensionality reduction and monitor one or more quality characteristics on the basis of several covariates observed over a discrete time domain. In their works, the dimensionality reduction is achieved by projection methods of a multivariate domain, such as principal component analysis (PCA), and indeed allows to jointly monitor, in addition to the quality characteristic under study, the covariates themselves. These multivariate methods have the potential to cope with technological problems and contexts where the quality characteristic or the covariates are described by smooth functions

over some continuous domain (sometimes referred to also as profile (Woodall et al. 2004)) and gave raise, in the last decade, to the new SPC field known as *profile monitoring* (Noorossana et al. 2012, Colosimo & Pacella 2007, 2010). The increasing need of handling those type of data (Happ & Greven 2018, Chen & Jiang 2017) have naturally unleashed cross-fertilization of profile monitoring with functional data analysis (FDA) (Ramsay & Silverman 2005, Wang et al. 2016). The possibility of using derivative information in FDA gives many advantages in dealing with complex objects, mainly due to its nonparametric nature. Nevertheless, it allows retrieving and extending techniques from the multivariate settings, e.g., regression models, PCA.

FDA techniques can be exploited to fill the gap in the SPC literature on methods for the joint monitoring of multivariate functional data observed over multi-dimensional domains (Happ & Greven 2018) and quality characteristics related to them. In this work, the quality characteristic is supposed to be a scalar and covariates to be real-valued functions with one-dimensional domain.

In what follows, the regression control chart is extended to the functional case by considering the scalar-on-function regression (Reiss et al. 2017), i.e., a functional linear model with scalar response and functional covariates. In particular, we develop the idea introduced by Chiou et al. (2016), who performs a multivariate functional principal component analysis (MFPCA) (Chen & Jiang 2017) on the functional covariates and uses the retained principal components to model the relationship with the scalar response. In addition, we discuss the optimal choice of the functional principal components to retain into the model, with the aim of considering also the variability in the covariates that is useful for the prediction of the scalar response, which is an issue raised also in the multivariate context (Jolliffe 2002) and usually overlooked in the classical PCA. Moreover, MFPCA allows the extension of profile monitoring techniques based on the Hotelling T^2 and squared prediction error (SPE) control charts to the joint monitoring of the multivariate functional covariates. As in the multivariate case, contribution plots shall be defined accordingly to help diagnosing variables, among the functional covariates, that are possibly responsible for an *out-of-control* (OC) condition detected by the T^2 or SPE statistics.

The proposed monitoring strategy can be recapped in the following three main steps:

- (i) Phase I: estimating a scalar-on-function regression model based on an *in-control* (IC) reference data set that is supposed to contain all the structural information about how the variable measurements deviate from their average trajectories under normal operation (also referred to as *training* or *Phase I* sample);
- (ii) Phase II: monitoring of new observations of the functional covariates, by means of functional T^2 and SPE control charts, and of the scalar response, via regression control chart, i.e., testing whether the new observation behaviour is consistent with that of the Phase I sample or signals an OC condition;
- (iii) diagnosing faults when an OC condition is detected i.e., highlighting the most influencing variable(s) by means of contribution plots to T^2 or SPE statistics.

According with the SPC literature (Woodall et al. 2004, Montgomery 2007), Step (i) will be hereinafter referred to as Phase I and step (ii) as Phase II. Furthermore, the proposed method is suitably generalized to be used for the real-time monitoring (e.g., up to any intermediate time domain point) of the functional covariates and scalar response. The proposed monitoring strategy is motivated and illustrated by means of a real-case study from the maritime field

in monitoring CO₂ emissions during the navigation phase of a roll-on/roll-off passenger (Ro-Pax) cruise ship, whose data are courtesy of the owner Grimaldi Group.

The paper is organized as follows: Section 7.2 describes the motivating example related to the problem of CO₂ emission in the maritime field; Section 7.3 sets up the notation and recall the main aspects of the scalar-on-function regression methodology, introduces the proposed functional control charts and the regression control chart; Section 7.4 presents the real-case study; Section 7.5 illustrates the proposed functional real-time monitoring strategy, and Section 7.6 draws conclusions.

7.2 A Motivating Example

In the last years, the problem of monitoring CO₂ emissions in the maritime transportation field has become of paramount importance in view of the climate change and global warming issues. The Marine Environment Protection Committee of the International Maritime Organization has given raise at each continent level to extensive air pollution programs (European Commission 2015, IMO 2012*a,b,d*, 2014, Smith et al. 2015) that require monitoring and verification of CO₂ emissions.

To respond to this compelling regulatory regime, shipping companies are nowadays setting-up modern multi-sensor systems on their fleets that allow massive amounts of observational data to be automatically streamed and stored to a remote server, bypassing human intervention. However, monitoring of the measured emissions still represents an open challenge for both shipping operators and energy policy makers. Several additional factors can in fact affect vessel performance, e.g., ship type, draught, speed, acceleration, encounter angle, wind regime, sea state (Bialystocki & Konovessis 2016), which are, in general, also function of time.

The maritime field constitutes a new challenging area for FDA and related SPC methods. The problems addressed are, on the one hand, to build models that allow prediction of ship CO₂ emissions based on observational data describing ship operating conditions, and, on the other, to monitor operating conditions for detecting anomalies and diagnosing faults.

Maritime engineering literature is mainly devoted to physical deterministic relationships under standard conditions and dedicated speed-trial test data and have strong limitations when applied to real data, which are typically more complex, larger in size, and collected from various sources. Few attempts to circumvent these issues can, however, be found in the following works. Perera & Mo (2016) drew empirical relationships between ship resistance and speed through data visualization methods. Petersen et al. (2012) investigated artificial neural networks and Gaussian Process approaches for statistical modeling of fuel efficiency. Lu et al. (2015) developed a semi-empirical ship operational performance predictive model to estimate the ship's added resistance considering specific additional variables. Bocchetti et al. (2015) proposed a statistical approach founded on multiple linear regression which allows for both pointwise and interval predictions of the fuel consumption at given operating conditions.

Statistical approaches have bend modern multivariate analytics to the maritime context only in the very last years (see e.g., Lepore et al. (2017) for a thorough comparison). However, the majority of the approaches that have already appeared in the maritime literature do not exploit the potential for continuously supporting managerial decision-making through the monitoring of the entire voyage profiles acquired on board.

7.3 Methodology

The scalar-on-function regression model is illustrated in Section 7.3, while the Phase I model estimation is described in Section 7.3. The Phase II monitoring procedure and fault diagnosis are introduced in Sections 7.3 and 7.3, respectively.

Scalar-on-Function Regression Model

We consider the Hilbert space \mathbb{H} of P -dimensional vectors whose components are functions in the space $L^2(\mathcal{T})$, with compact domain $\mathcal{T} \subset \mathbb{R}$. Functions $f, g \in \mathbb{H}$ can be written as $f = (f_1, \dots, f_P)$ and $g = (g_1, \dots, g_P)$, where $f_p, g_p \in L^2(\mathcal{T})$. In this setting, we can define the inner product of \mathbb{H} as $\langle f, g \rangle_{\mathbb{H}} = \sum_{p=1}^P \langle f_p, g_p \rangle$, where $\langle f_p, g_p \rangle = \int_{\mathcal{T}} f_p(t)g_p(t)dt$ is the inner product of $L^2(\mathcal{T})$, and the induced norm of \mathbb{H} as $\|f\|_{\mathbb{H}} = \langle f, f \rangle_{\mathbb{H}}^{1/2}$.

Let us denote with $\tilde{\mathbf{X}} = (\tilde{X}_1, \dots, \tilde{X}_P)$ a random element that takes values in \mathbb{H} , i.e., $\tilde{X}_1, \dots, \tilde{X}_P$ are random elements that take values in $L^2(\mathcal{T})$, which are hereinafter referred to as *functional covariates*. Moreover, let $\tilde{\mathbf{X}}$ have mean function $\boldsymbol{\mu}^X = (\mu_1^X, \dots, \mu_P^X)$, with $\mu_p^X(t) = E(\tilde{X}_p(t))$ for every $t \in \mathcal{T}$, variance function $\mathbf{v}^X = (v_1^X, \dots, v_P^X)$, where $v_p^X(t) = \text{Var}(\tilde{X}_p(t))$, and correlation function $\mathbf{C} = \{C_{p_1 p_2}\}_{p_1, p_2=1, \dots, P}$, with $C_{p_1 p_2}(t_1, t_2) = \text{Corr}(\tilde{X}_{p_1}(t_1), \tilde{X}_{p_2}(t_2)) = \text{Cov}(\tilde{X}_{p_1}(t_1), \tilde{X}_{p_2}(t_2))v_{p_1}(t_1)^{-1/2}v_{p_2}(t_2)^{-1/2}$. To deal with infinite dimensionality of the data, $\tilde{\mathbf{X}}$ is decomposed through multivariate functional principal component analysis (MFPCA). However, as is known, this method is not scale-invariant, thus $\tilde{\mathbf{X}}$ are suitably scaled through the normalization approach proposed by Chiou, Chen & Yang (2014). The normalized functional covariates are denoted by $\mathbf{X} = (X_1, \dots, X_P)$, and $X_p(t)$, $p = 1, \dots, P$, are obtained as $v_p(t)^{-1/2}(\tilde{X}_p(t) - \mu_p^X(t))$. Trivially note that \mathbf{X} has zero mean and covariance function that coincides with \mathbf{C} .

Denote by y the scalar response variable, let $\{(\tilde{\mathbf{X}}_i, y_i)\}_{i=1, \dots, n}$ be a random sample from $(\tilde{\mathbf{X}}, y)$, with $\tilde{\mathbf{X}}_i = (\tilde{X}_{i1}, \dots, \tilde{X}_{iP})$. The conditional distribution of y_i given the corresponding observation of the normalized functional covariates \mathbf{X}_i is modelled by means of the following scalar-on-function regression

$$y_i = \beta_0 + \langle \mathbf{X}_i, \boldsymbol{\beta} \rangle_{\mathbb{H}} + \varepsilon_i = \beta_0 + \sum_{p=1}^P \int_{\mathcal{T}} X_{ip}(t)\beta_p(t)dt + \varepsilon_i, \quad i = 1, \dots, n, \quad (7.1)$$

where $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P) \in \mathbb{H}$ are the coefficient to be estimated, and $\varepsilon_1, \dots, \varepsilon_n$ are the error terms, which are assumed to be independent identically distributed normal random variables with mean zero and variance σ^2 . Moreover, they are assumed to be uncorrelated with the functional covariates, i.e. $E(\varepsilon_i X_p(t)) = 0$ for each $i = 1, \dots, n$, $p = 1, \dots, P$, and $t \in \mathcal{T}$.

Phase I Model Estimation

Instead of using a random sample $\{(\mathbf{X}_i, y_i)\}_{i=1, \dots, n}$, SPC literature is more concerned to use in Phase I a reference data set that can be assumed representative of the normal behavior of the functional covariates and of the relation of the latter with the scalar response. Then, the coefficients β_0 and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P)$ in Equation (7.1) can be estimated by solving the

following least-squares problem

$$\min_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{H}} \sum_{i=1}^n (y_i - \beta_0 - \langle \mathbf{X}_i, \boldsymbol{\beta} \rangle_{\mathbb{H}})^2. \quad (7.2)$$

As is known, this problem is not well-posed since the solution has to be found among all the possible elements of the infinite-dimensional Hilbert space \mathbb{H} on the basis of a finite sample. However, as in Chiou et al. (2016), the problem can be approached by considering the Karhunen-Loève expansion of \mathbf{X}

$$\mathbf{X}(t) = \sum_{m=1}^{\infty} \xi_m \boldsymbol{\psi}_m(t), \quad (7.3)$$

where the multivariate functional principal components $\{\boldsymbol{\psi}_m = (\psi_{m1}, \dots, \psi_{mP})\}_{m \in \mathbb{N}}$, with $\boldsymbol{\psi}_m \in \mathbb{H}$, form an orthonormal basis of \mathbb{H} , i.e.

$$\langle \boldsymbol{\psi}_{m_1}, \boldsymbol{\psi}_{m_2} \rangle_{\mathbb{H}} = \sum_{p=1}^P \langle \psi_{m_1 p}, \psi_{m_2 p} \rangle = \begin{cases} 1 & \text{if } m_1 = m_2 \\ 0 & \text{if } m_1 \neq m_2 \end{cases}. \quad (7.4)$$

The latter represent the eigenfunctions of the covariance operator $\Gamma : \mathbb{H} \rightarrow \mathbb{H}$ defined by $\Gamma \boldsymbol{x} = \mathbb{E}[\langle \boldsymbol{x}, \mathbf{X} \rangle_{\mathbb{H}} \mathbf{X}]$, i.e. they are the solutions of $\Gamma \boldsymbol{\psi}_m = \lambda_m \boldsymbol{\psi}_m$, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$. In Equation (7.3), the multivariate functional principal component scores, or scores, $\xi_m = \langle \mathbf{X}, \boldsymbol{\psi}_m \rangle_{\mathbb{H}} = \sum_{p=1}^P \langle X_p, \psi_{mp} \rangle$ are random coefficients with $\mathbb{E}(\xi_m) = 0$, $\mathbb{E}(\xi_m^2) = \lambda_m$ and $\mathbb{E}(\xi_{m_1} \xi_{m_2}) = 0$ when $m_1 \neq m_2$.

The coefficient $\boldsymbol{\beta}$ in the model in Equation (7.1) can be expressed by using the same eigenbasis of \mathbb{H}

$$\boldsymbol{\beta}(t) = \sum_{m=1}^{\infty} b_m \boldsymbol{\psi}_m(t). \quad (7.5)$$

In this way, by substituting Equation (7.3) and (7.5) into Equation (7.1) we get

$$y_i = \beta_0 + \sum_{m=1}^{\infty} \langle \xi_{im} \boldsymbol{\psi}_m, b_m \boldsymbol{\psi}_m \rangle_{\mathbb{H}} + \varepsilon_i = \beta_0 + \sum_{m=1}^{\infty} \xi_{im} b_m + \varepsilon_i, \quad i = 1, \dots, n, \quad (7.6)$$

where $\xi_{im} = \langle \mathbf{X}_i, \boldsymbol{\psi}_m \rangle_{\mathbb{H}}$ are the scores of the i -th observation \mathbf{X}_i . Since the scores are orthogonal, the coefficients b_m can be estimated separately because they only depend on the corresponding ξ_m . However, we would not be able to estimate infinite parameters and get $\hat{\boldsymbol{\beta}}$ that minimizes Equation (7.2) because of the finite number of available data. Therefore, we consider an M -dimensional approximation of $\mathbf{X}(t)$ in Equation (7.3)

$$\mathbf{X}_M(t) = \sum_{m \in \mathcal{M}} \xi_m \boldsymbol{\psi}_m(t), \quad (7.7)$$

where $\mathcal{M} = \{m_1, \dots, m_M\} \subset \mathbb{N}$ is a set of M distinct natural numbers, indicating which principal components to retain in the scalar-on-function regression model.

The choice of \mathcal{M} is usually carried out by maximizing the proportion of the total variability explained by the principal components. According to this criterion, the optimal choice is to retain the first M components, i.e., $\mathcal{M} = \{1, \dots, M\}$ (Chiou, Zhang, Chen & Chang 2014). However, this is not the only possible choice. The variable selection problem for

principal component regression is well known in the multivariate setting and discussed in Jolliffe (2002). In fact, if one is interested in prediction of the scalar response, there are some components that may have small predictive ability, for which ideally the coefficient b_m is zero. Therefore, retaining those components in the model would not be beneficial for the estimation of b_m . On the other hand, we are still interested in retaining components with large variances, for which the corresponding estimates of b_m are more stable, and discarding components with low variance.

A parsimonious choice may be discarding all components whose variance is less than a threshold and result not significant for the regression on the scalar response. For this purpose, an error statistic, e.g., the prediction sum of squares (PRESS) statistic (Montgomery et al. 2012), calculated by cross-validation, can be considered. In this paper, the PRESS statistic is obtained via leave-one-out cross-validation as

$$PRESS = \sum_{i=1}^n (y_i - \hat{y}_{[i]})^2, \quad (7.8)$$

where $\hat{y}_{[i]}$ is the prediction of y_i based on the scalar-on-function regression model with the i -th observation removed from the reference data set. The idea is then to select only those components that achieve a PRESS reduction larger than a threshold. The practical illustration of this procedure for selecting \mathcal{M} can be found in the real-case study addressed in Section 7.4. By considering the approximation in Equation (7.7) and taking into account Equation (7.6), we can write the model in Equation (7.1) as

$$y_i = \beta_0 + \sum_{m \in \mathcal{M}} \xi_{im} b_m + \varepsilon_i^M, \quad i = 1, \dots, n, \quad (7.9)$$

where $\varepsilon_i^M = \sum_{m \in \mathbb{N} \setminus \mathcal{M}} \xi_{im} b_m + \varepsilon_i$ is as close to ε_i as ξ_{im} has low variance or b_m is close to zero.

To get the final least squares estimate of β_0 and β in Equation (7.5) based on a set of n observations $(\tilde{\mathbf{X}}_i, y_i)$, we first estimate the mean function as $\hat{\boldsymbol{\mu}}^X(t) = \sum_{i=1}^n \tilde{\mathbf{X}}_i(t)/n$ and the variance function as $\hat{\boldsymbol{v}}^X(t) = \sum_{i=1}^n (\tilde{\mathbf{X}}_i(t) - \hat{\boldsymbol{\mu}}^X(t))^2 / (n-1)$, then standardize $\tilde{\mathbf{X}}_i$ and obtain $\mathbf{X}_i = (X_{i1}, \dots, X_{iP})$, where $X_{ip}(t) = \hat{v}_p(t)^{-1/2} (\tilde{X}_{ip}(t) - \hat{\mu}_p^X(t))$.

The estimates of eigenvalues $\hat{\lambda}_m$ and eigenfunctions $\hat{\boldsymbol{\psi}}_m$ of the covariance operator Γ can be obtained by applying MFPCA on the observed data, for example by using the principal analysis by conditional expectation algorithm (Happ & Greven 2018), or, alternatively, through the spectral decomposition of the discrete version of the estimate of the correlation function (Chiou et al. 2016). In any case, the principal component scores can be eventually estimated as $\hat{\xi}_{im} = \langle \mathbf{X}_i, \hat{\boldsymbol{\psi}}_m \rangle_{\mathbb{H}}$. Note that, on the basis of a finite sample of size n , the maximum number of multivariate functional principal components that can be estimated is $n-1$, i.e., $\hat{\lambda}_m = 0$ for $m \geq n$.

Based on \mathcal{M} , the intercept can be estimated as $\hat{\beta}_0 = \sum_{i=1}^n y_i/n$ (since the scores have null means) and the coefficients b_m can be estimated separately as

$$\hat{b}_m = \frac{\sum_{i=1}^n y_i \hat{\xi}_{im}}{\sum_{i=1}^n \hat{\xi}_{im}^2}. \quad (7.10)$$

Accordingly, the estimate of β can be obtained as

$$\hat{\beta}(t) = \sum_{m \in \mathcal{M}} \hat{b}_m \hat{\boldsymbol{\psi}}_m(t), \quad (7.11)$$

and the prediction of y_i results to be

$$\hat{y}_i = \hat{\beta}_0 + \sum_{m \in \mathcal{M}} \hat{\xi}_{im} \hat{b}_m, \quad (7.12)$$

Finally, we can also get an estimate of the variance σ^2 of the error in the model in Equation (7.1)

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - M - 1}. \quad (7.13)$$

Phase II Monitoring of Multivariate Functional Covariates and Scalar Response

The information that allows Phase II monitoring is assumed to be incorporated in the multivariate functional principal components estimated on the IC reference data (Phase I). Suppose that, on the basis of the Phase I sample $\{(\mathbf{X}_i, y_i)\}_{i=1, \dots, n}$, the scalar-on-function regression model has been estimated together with all the parameters described in Section 7.3. Moreover, suppose that a new observation $(\mathbf{X}^{new}, y^{new})$ is available from (\mathbf{X}, y) , where y^{new} is the new (i.e., future) observation of the response variable and $\mathbf{X}^{new} = (X_1^{new}, \dots, X_P^{new})$ is the corresponding new standardized observation of $\tilde{\mathbf{X}}$. The new scores $\{\hat{\xi}_m^{new}\}_{m \in \mathcal{M}}$ can be calculated as

$$\hat{\xi}_m^{new} = \langle \mathbf{X}^{new}, \hat{\psi}_m \rangle_{\mathbb{H}}, \quad m \in \mathcal{M}. \quad (7.14)$$

Note that the inner product in Equation (7.14) assumes \mathbf{X}^{new} to be completely observed over the domain \mathcal{T} to calculate the statistics used for monitoring the operating conditions defined as in Equation (7.15), (7.17), and (7.18). The first two, namely the Hotelling T^2 and SPE statistics, define two functional control charts for monitoring the multivariate functional covariates. The third statistic allows monitoring the scalar error term pertaining to the response variable and will be hereinafter referred to as *response prediction error*.

Hotelling T^2 control chart The Hotelling T^2 statistic monitors the components retained to estimate the scalar-on-function regression model

$$T^2 = \sum_{m \in \mathcal{M}} \frac{(\hat{\xi}_m^{new})^2}{\hat{\lambda}_m}. \quad (7.15)$$

That is the part of variability in the functional covariates which is informative for the prediction of the scalar response. The distribution of the T^2 statistic depends on the distribution of the scores in \mathcal{M} , which in general is not known. An upper control limit can be set as the empirical $(1 - \alpha_{T^2})$ quantile of the T^2 statistic values obtained for the Phase I sample.

Squared prediction error control chart The SPE statistic looks at the norm of the residual function obtained by approximating \mathbf{X}^{new} with $\hat{\mathbf{X}}_M^{new} = (\hat{X}_{M1}^{new}, \dots, \hat{X}_{MP}^{new})$

$$\hat{\mathbf{X}}_M^{new}(t) = \sum_{m \in \mathcal{M}} \hat{\xi}_m^{new} \hat{\psi}_m(t). \quad (7.16)$$

That is the part of variability in the functional covariates not considered in the T^2 statistic, i.e., related to those components that have little relevance in the prediction of the scalar

response. By noting that $\mathbf{X}^{new}(t) - \hat{\mathbf{X}}_M^{new}(t) = \sum_{m \in \{1, \dots, n-1\} \setminus \mathcal{M}} \hat{\xi}_m^{new} \hat{\psi}_m(t)$, we can write

$$\begin{aligned}
 SPE &= \|\mathbf{X}^{new} - \hat{\mathbf{X}}_M^{new}\|_{\mathbb{H}}^2 = \\
 &= \left\langle \sum_{m_1 \in \{1, \dots, n-1\} \setminus \mathcal{M}} \hat{\xi}_{m_1}^{new} \hat{\psi}_{m_1}, \sum_{m_2 \in \{1, \dots, n-1\} \setminus \mathcal{M}} \hat{\xi}_{m_2}^{new} \hat{\psi}_{m_2} \right\rangle_{\mathbb{H}} = \\
 &= \sum_{m_1 \in \{1, \dots, n-1\} \setminus \mathcal{M}} \sum_{m_2 \in \{1, \dots, n-1\} \setminus \mathcal{M}} \hat{\xi}_{m_1}^{new} \hat{\xi}_{m_2}^{new} \langle \hat{\psi}_{m_1}, \hat{\psi}_{m_2} \rangle_{\mathbb{H}} = \\
 &= \sum_{m \in \{1, \dots, n-1\} \setminus \mathcal{M}} (\hat{\xi}_m^{new})^2. \quad (7.17)
 \end{aligned}$$

As for the Hotelling T^2 statistic, an upper control limit can be set as the empirical $(1 - \alpha_{SPE})$ quantile of the SPE statistic values obtained for the Phase I sample.

Response prediction error control chart Beside monitoring of functional covariates, the scalar response can also be monitored itself through the *response prediction error* given by

$$y^{new} - \hat{y}^{new} = y^{new} - \hat{\beta}_0 - \sum_{m \in \mathcal{M}} \hat{\xi}_m^{new} \hat{b}_m. \quad (7.18)$$

Since the experimental errors are assumed to have independent identical normal distribution, the lower $-L_{\alpha_y}$ and upper L_{α_y} control limits for the response prediction error can be obtained by setting

$$\begin{aligned}
 L_{\alpha_y} &= t_{n-M-1, 1-\alpha_y/2} \left[\hat{\sigma}^2 \left(1 + \frac{1}{n-1} \sum_{m \in \mathcal{M}} \frac{(\hat{\xi}_m^{new})^2}{\hat{\lambda}_m} \right) \right]^{1/2} = \\
 &= t_{n-M-1, 1-\alpha_y/2} \left[\hat{\sigma}^2 \left(1 + \frac{T^2}{n-1} \right) \right]^{1/2} \quad (7.19)
 \end{aligned}$$

where $t_{n-M-1, 1-\alpha_y/2}$ is the $(1 - \alpha_y/2)$ quantile of the Student distribution with $n - M - 1$ degrees of freedom. Note that the limits depend on the value of the T^2 statistic. A higher value in T^2 determines wider prediction error limits. If the distribution of the experimental errors is not normal and more generally does not belong to a scale and location family, the limits cannot be standardized to be equal. Nevertheless, they can be estimated non-parametrically only asymptotically, because when the sample size grows they tend to be of constant amplitude whatever the new observation of functional covariates. This control chart can be regarded as the natural extension of the regression control chart known in the SPC literature firstly introduced by Mandel (1969) to the case of multivariate functional covariates by means of the scalar-on-function regression model of Equation (7.1).

Since the simultaneous use of three control charts boils down in testing three hypotheses for each observation, the control limits have to be selected to control the (type-I) family-wise error rate (FWER) for a significance level α . In what follows, we denote by α_{T^2} , α_{SPE} , and α_y the significance levels to be separately used in the T^2 , SPE , and response prediction error control charts, respectively. In all those cases when these three control charts can be

assumed as independent the Šidák correction (Šidák 1967) gives an exact FWER of α by choosing α_{T^2} , α_{SPE} , and α_y such that

$$(1 - \alpha_{T^2})(1 - \alpha_{SPE})(1 - \alpha_y) = 1 - \alpha. \quad (7.20)$$

and is conservative if they are positively dependent. However, as is known the Šidák correction cannot be used if tests are suspected to be negatively dependent. In this latter case, the alternative is the classical Bonferroni correction that can be utilized to guarantee that the type-I FWER is not larger than α , by choosing α_{T^2} , α_{SPE} , and α_y such that

$$\alpha_{T^2} + \alpha_{SPE} + \alpha_y = \alpha. \quad (7.21)$$

However, as is known, this correction is more conservative than the previous one, and results in a lower power. Whatever multiple correction one wants to use, a possible choice is to assign the same correction to the three control charts, then

$$\alpha_{T^2} = \alpha_{SPE} = \alpha_y = 1/3. \quad (7.22)$$

A suitable alternative is to split equally the FWER into the control level of the functional covariate control charts (T^2 and SPE) and scalar response prediction error control chart

$$\alpha_{T^2} = \alpha_{SPE} = \alpha/4, \quad \alpha_y = \alpha/2. \quad (7.23)$$

Fault Diagnosis via Contribution Plots

The behavior of a new observation is assessed by comparing the T^2 , SPE and response prediction error statistics with respect to the control limits built in Phase I. If at least one statistic is out of the control limits, then an OC alarm is issued. Unusual behaviors can be explored by analyzing the single contribution of each variable to trigger the OC as follows.

As proposed in Kourti & MacGregor (1996), the overall contribution of each functional variable to the Hotelling statistic T^2 can be defined by observing that

$$T^2 = \sum_{m \in \mathcal{M}} \frac{\hat{\xi}_m^{new}}{\hat{\lambda}_m} \hat{\xi}_m^{new} = \sum_{m \in \mathcal{M}} \frac{\hat{\xi}_m^{new}}{\hat{\lambda}_m} \langle \mathbf{X}^{new}, \hat{\psi}_m \rangle_{\mathbb{H}} = \sum_{p=1}^P \sum_{m \in \mathcal{M}} \frac{\hat{\xi}_m^{new}}{\hat{\lambda}_m} \langle X_p^{new}, \hat{\psi}_{mp} \rangle. \quad (7.24)$$

Then, we can write

$$CONT_p^{T^2} = \sum_{m \in \mathcal{M}} \frac{\hat{\xi}_m^{new}}{\hat{\lambda}_m} \langle X_p^{new}, \hat{\psi}_{mp} \rangle, \quad p = 1, \dots, P. \quad (7.25)$$

The contribution of each functional variable to the SPE statistic, rewritten as

$$SPE = \sum_{p=1}^P \|X_p^{new} - X_{Mp}^{new}\|^2, \quad (7.26)$$

can be analogously defined as

$$CONT_p^{SPE} = \|X_p^{new} - X_{Mp}^{new}\|^2, \quad p = 1, \dots, P. \quad (7.27)$$

Note that, even if both T^2 and SPE statistics are non negative, $CONT_p^{T^2}$ can be negative for some variable. In general, the contributions have not the same distribution for all the

variables. A proper upper limit for each variable contribution to T^2 and SPE statistics has to be set to support the identification of anomalous variables. A plausible choice of the upper limit is to estimate it from its empirical distribution based on the reference data set. A multiple test correction should then be used to control the type-I FWER. By using the Bonferroni correction, which is the simplest choice, the upper control limits can be set as the $(1 - \alpha_{T^2}/P)$ and $(1 - \alpha_{SPE}/P)$ quantiles of the empirical distribution for each contribution.

On the other hand, the response prediction error does not benefit from a decomposition into interpretable contributions. Then, when an OC is issued by the response prediction error control chart, possible causes have to be investigated outside the set of variables included in the model as functional covariates.

7.4 A Real-Case Study

Data collected from a Ro-Pax cruise ship owned by the Italian shipping company Grimaldi Group are used to illustrate the proposed method. Functional data from the multi-sensor system installed on board are used for constructing the trajectories of the different ship operating conditions (i.e., the functional covariates) and monitor CO₂ emissions, (i.e., the response variable). In view of the recent regulations discussed in Section 7.2, monitoring of CO₂ emissions is of great interest in the maritime sector to timely plan energy efficiency improvement operations and react to anomalies.

Section 7.4 describes the variables chosen as functional covariates and scalar response. Section 7.4 illustrates the preprocessing step required to obtain functional covariates from the acquired signals. Section 7.4 shows the implementation details for estimation of the scalar-on-function regression model (Phase I) and illustrates a scenario in which the proposed monitoring strategy is applied to Phase II monitoring. Section 7.4 compares the latter with simpler method based on scalar observations per each voyage.

Variable Description

The considered Ro-Pax ship has two engine sets. Each engine set has two main engines for propulsion with a variable pitch propeller and a shaft generator for electric power supply. The name of the ship, route, and voyage dates are omitted for confidentiality reasons.

Table 7.1 reports the $P = 9$ variables used as functional covariates in the scalar-on-function regression model in Equation (7.1). The scalar response variable, y , is the total CO₂ emissions, measured in tonne (t). Each observation refers to the *navigation phase* of each ship voyage at given route and direction. More precisely, the navigation phase starts with the *finished with engine order* (when the ship leaves the departure port) and ends with the *stand by engine order* (when the ship enters the arrival port).

The *cumulative sailing time* variable, measured in hour (h) is the cumulative voyage navigation time. The *speed over ground* (SOG) variable, measured in knot (kn), is the ratio between the sailed distance over ground, i.e., the distance travelled by the vessel during the navigation phase, hereinafter denoted by d , and the cumulative sailing time. The former is measured in nautic mile (NM) and calculated from latitude and longitude data acquired by the by the GPS sensor through the Haversine formula. The *acceleration* variable is obtained as the first derivative of SOG with respect to the sailing time. The *power difference between port and starboard propeller shafts* is included for discovering anomalies or malfunctioning in the main engines, e.g., when one of the engines is turned off. The *Distance from the nominal route* variable, measured in nautic mile (NM), is calculated as the distance, at each domain

Table 7.1: Functional covariates used in the scalar-on-function regression model.

Variable number	Variable name	Symbol	Unit of measurement
1	Speed Over Ground (SOG)	V	kn
2	Acceleration	A	NM/h^2
3	Power difference between port and starboard propeller shafts	ΔP	kW
4	Distance from the nominal route	$Dist$	NM
5	Longitudinal wind component	W_L	kn
6	Transverse wind component	W_T	kn
7	Air Temperature, mean of four engines	T	$^{\circ}C$
8	Cumulative sailing time	H	h
9	Trim	$Trim$	m

point, of the actual GPS position of the vessel from the position indicated in the nominal route. The wind component variables are calculated on the basis of the wind speed W , measured in kn , and direction relative to the ship Ψ , measured in radiant, acquired by the anemometer sensor. The *longitudinal wind component* variable is obtained as $W_L = W \cos \Psi$. The *transverse wind component* variable is obtained as $W_T = |W \sin \Psi|$. The *air temperature* variable is the average of the temperatures measured from the sensors installed on each of the four main engines. The *Trim* variable is obtained through the inclinometer sensor measurements. Additional information about the variables can be found in Bocchetti et al. (2015), Erto et al. (2015), Lepore et al. (2017), and Reis et al. (2020).

Preprocessing and Registration

In the proposed real-case study, functional data are obtained from profiles collected during the navigation at five-minute frequency by the multi-sensor system on-board. The first step to be carried out is to get smooth observations $\tilde{\mathbf{X}}_i = (\tilde{X}_{i1}, \dots, \tilde{X}_{iP})$ of the functional covariates $\tilde{\mathbf{X}}$ at each voyage $i = 1, \dots, n$. For each $i = 1, \dots, n$ and $p = 1, \dots, P$, \tilde{X}_{ip} can be obtained from the discrete data x_{ipn} , $n = 1, \dots, N_i$, using a cubic B-spline basis with equally spaced knots

$$\tilde{X}_{ip}(t) = \sum_{q=1}^Q c_{iqp} \phi_q(t), \quad i = 1, \dots, n, \quad p = 1, \dots, P, \quad t \in \mathcal{T}, \quad (7.28)$$

where ϕ_1, \dots, ϕ_Q are the B-spline basis functions and c_{iqp} are the basis coefficients. Functional data have been obtained by smoothing data with regularization, using the R package `fda` (Ramsay et al. 2018). Since the number of basis functions should be large enough to ensure that the regularization is controlled by the choice of the smoothing parameter, 100 bases with equally spaced knots and a roughness penalty on the integrated squared second derivative have been chosen. For each functional variable and each observation, the smoothing parameter are chosen separately by minimizing the generalized cross-validation criterion (Ramsay & Silverman 2005).

Even if time is naturally prone to be chosen as functional domain, total travelling time could vary from voyage to voyage. Thus, a reasonable choice is to use the fraction of distance travelled over the voyage as the common domain $\mathcal{T} = [0, 1]$ of the functional data (Abramowicz et al. 2018). This choice can be regarded as a landmark registration (Ramsay

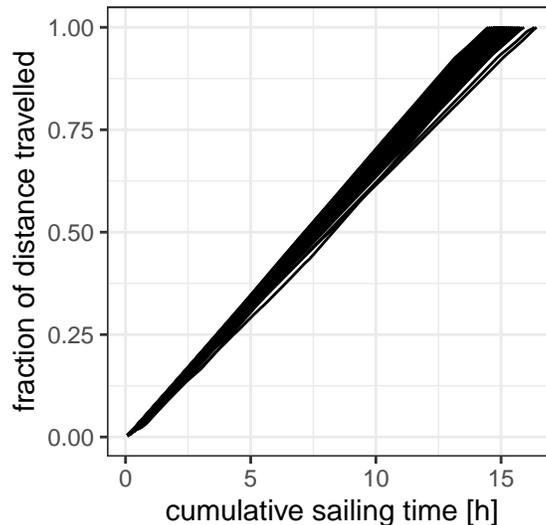


Figure 7.1: Warping functions mapping at each voyage the cumulative sailing time to the common domain $\mathcal{T} = [0, 1]$, which is the fraction of distance travelled over the voyage.

et al. 2009) of the functional data set from the function-specific temporal domain to the common domain $[0, 1]$ with the group of affine transformations with positive slope as the group of the warping functions and voyage starting and ending points as landmarks.

Model Estimation and Perspective Monitoring

The reference data set has $n = 139$ observations and is used to estimate the control limits for the perspective (Phase II) monitoring of 30 consecutive voyages. The estimation of multivariate functional principal components and corresponding scores are obtained through the R package `fda` (Ramsay et al. 2018). As explained in Section 7.3, the choice of \mathcal{M} , i.e., the set of components to retain in the model, is carried out by considering both the variability of the covariates explained by the principal components, reported in Figure 7.2a, and PRESS statistic calculated by Equation (7.8), reported in Figure 7.2b, as a function of the first m retained principal components in Equation (7.9). As previously stated in Section 7.3, we get $\mathcal{M} = \{1, 2, 5, 6, 12\}$ as the set of principal components that achieve the higher reductions of the PRESS statistic and percentages of variance explained larger than the threshold value (0.01) (Figure 7.2a). The normality assumption for the errors is supported by the Shapiro-Wilk test (p -value = 0.11).

To give the same importance to functional covariates and scalar response, the functional control charts are built by choosing the Bonferroni correction as $\alpha_{T^2} = \alpha_{SPE} = \alpha/4$ and $\alpha_y = \alpha/2$ as proposed in Equation (7.23), with $\alpha = 0.05$. The upper control limits of the Hotelling T^2 and SPE control charts are calculated as the Phase I empirical $(1 - \alpha_{T^2}/2)$ quantiles of the corresponding statistic. The limits for the response prediction error control chart have been calculated using Equation (7.19). Figure 7.3 shows the three control charts proposed in Section 7.3 used for the perspective monitoring of the upcoming voyages.

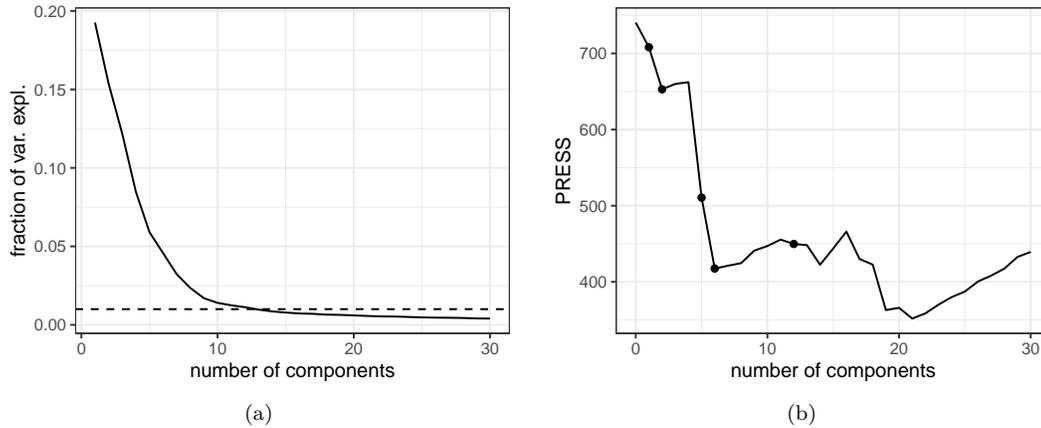


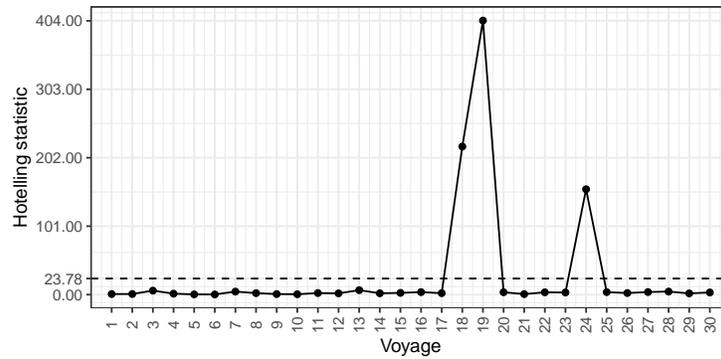
Figure 7.2: (a) Fraction of the variance of the functional covariates explained by the multivariate functional principal components (solid line) and threshold (dashed line) equal to 0.01. (b) PRESS statistic calculated by Equation (7.8) as a function of the first m retained principal components; trivially, when $m = 0$, the PRESS is obtained on the basis of the predictions calculated as the sample mean of the response variable; the points are the PRESS values corresponding to the multivariate functional principal components retained in the model.

Figures 7.4a and 7.4b report the boxplots of the functional covariates contribution to the Hotelling T^2 and SPE statistics, respectively, to visually highlight contributions are not identically distributed. Then, as discussed in Section 7.3, different limits for each variable are set for contributions to both T^2 and SPE by using the Bonferroni-like correction as the empirical $(1 - \alpha/(4P))$ -quantiles of the corresponding contribution obtained based on the Phase I reference sample.

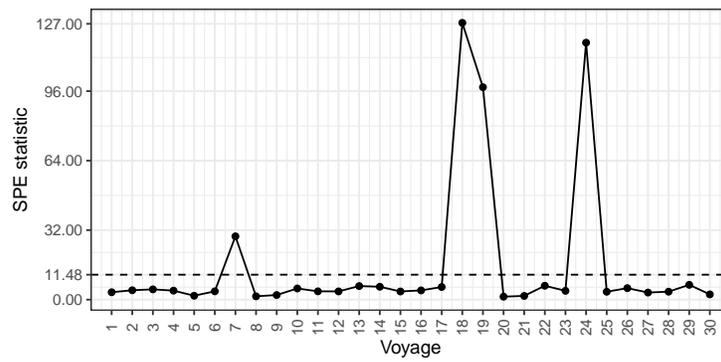
OC points signalled by T^2 or SPE control charts (Figures 7.3a and 7.3b) are investigated by means of the corresponding contribution plots and the most paradigmatic cases are illustrated and discussed below in chronological order.

Voyage 7

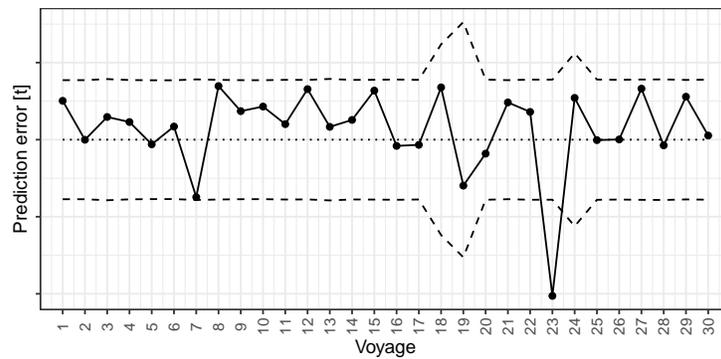
Voyage 7 is the first one to be signalled, by the SPE statistic, as OC. Contributions to the SPE statistic of the functional covariates are shown in Figure 7.5. From that figure, the variables to appear the main responsible for the OC condition are acceleration (A) and SOG (V), while contributions of cumulative sailing time (H) and mean air temperature of the engines (T), moderately exceed their limits. The corresponding trajectories are then explored in Figure 7.6. From the SOG profile (Figure 7.6a) it is clear that the ship was sailing at a lower speed for a short initial fraction of the voyage, which assuredly affected the sailing time (Figure 7.6b). A SOG profile higher than the average after the slowdown allowed the ship to complete the voyage without delay at the final destination. Accordingly, the acceleration (Figure 7.6c) shows SOG variations, and the mean air temperature of the engines (Figure 7.6d) reflects the same behaviour of the SOG.



(a)



(b)



(c)

Figure 7.3: (a) Hotelling T^2 , (b) SPE , and (c) response prediction error control charts used for monitoring the Phase II voyages. In each control chart, points joint by a line indicate monitoring statistic values at each voyage, while dashed lines indicate control chart limits as in Equation (7.23), with $\alpha = 0.05$.

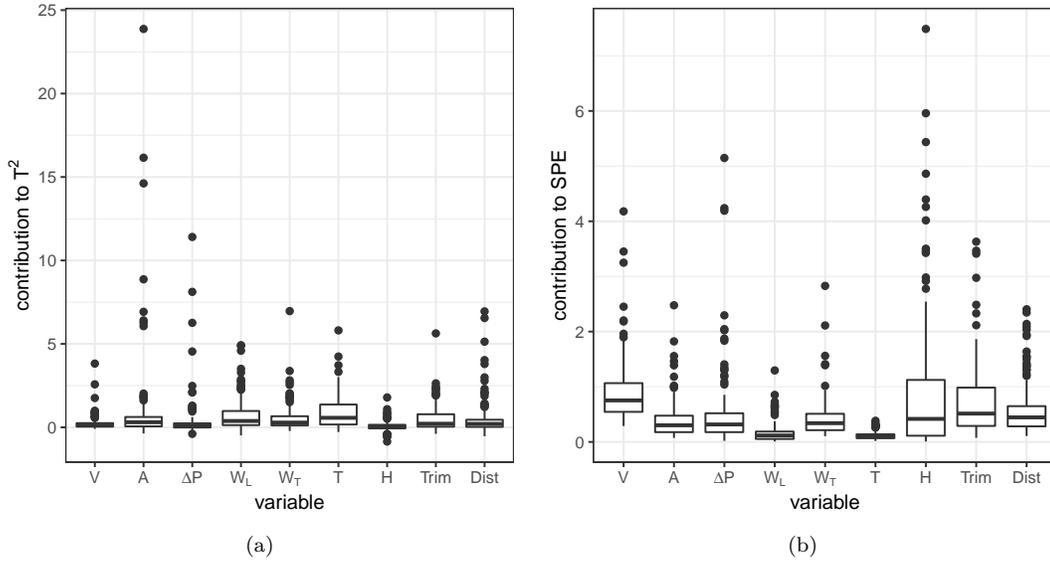


Figure 7.4: Box plots of the contributions of the functional covariates to (a) the Hotelling T^2 statistic and (b) the SPE statistic for the reference voyages.

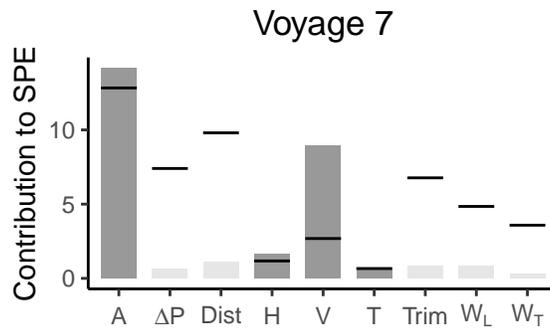


Figure 7.5: Contribution of the functional covariates to the SPE statistic for voyage 7. The bars are the contributions of the variables, with the darker ones indicating values exceeding the limit, while the black dashes are the limits calculated on the basis of the reference voyages.

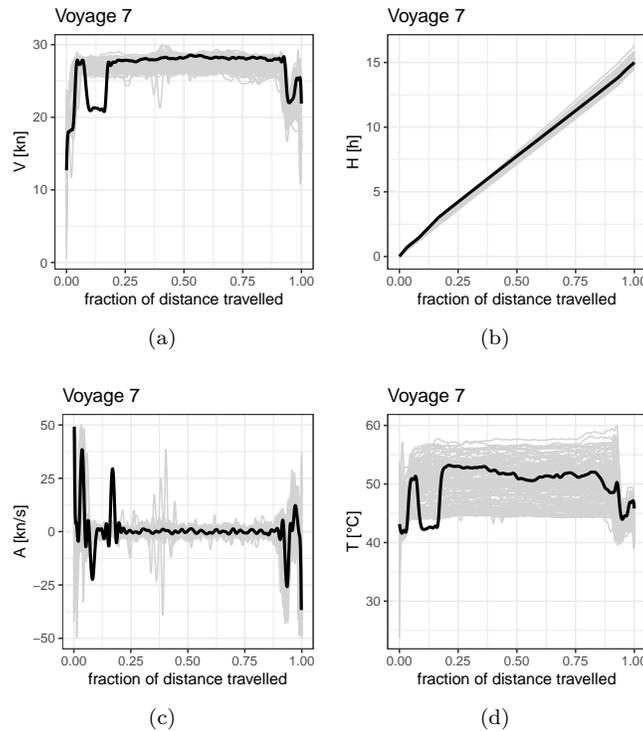


Figure 7.6: Observations of the critical (i.e., indicated as responsible of OC by contributions plots in Figure 7.5) functional covariates for Phase II monitoring of the voyage 7, viz. (a) SOG (V), (b) cumulative sailing time (H), (c) acceleration (A), (d) power difference between port and starboard propeller shafts (ΔP), and (e) longitudinal wind component (W_L). In each plot, the black line indicates the current observation, while the reference functional observations are plotted in grey.

Voyage 18

Voyage 18 is signalled as OC by both T^2 and SPE functional control charts. Note that to higher values of the Hotelling T^2 statistic correspond wider intervals in the response prediction error control chart. In Figure 7.7a, the contribution plot to the Hotelling T^2 statistic signals the SOG (V) and power difference between port and starboard propeller shafts (ΔP) as anomalous. Whereas, in Figure 7.7b, contribution plot to the SPE statistic signals longitudinal wind component (W_L), cumulative sailing time (H), and acceleration (A) variables. The trajectories of the functional covariates can be then exploited, as shown in Figure 7.8. As in the case of the voyage 7, this voyage is characterized by an atypical SOG profile (Figure 7.8a), with a lower average value throughout the entire voyage and atypical alternation of accelerations and decelerations (Figure 7.8c). This affected the sailing time (Figure 7.8b), which shows a strong delay on arrival. By looking at the profile for the power difference between port and starboard propeller shafts (ΔP) in Figure 7.8d, the ship is noticed to have had one of the main engines turned off for most of the voyage duration. This is also exacerbated by a very high longitudinal wind component profile (Figure 7.8e).

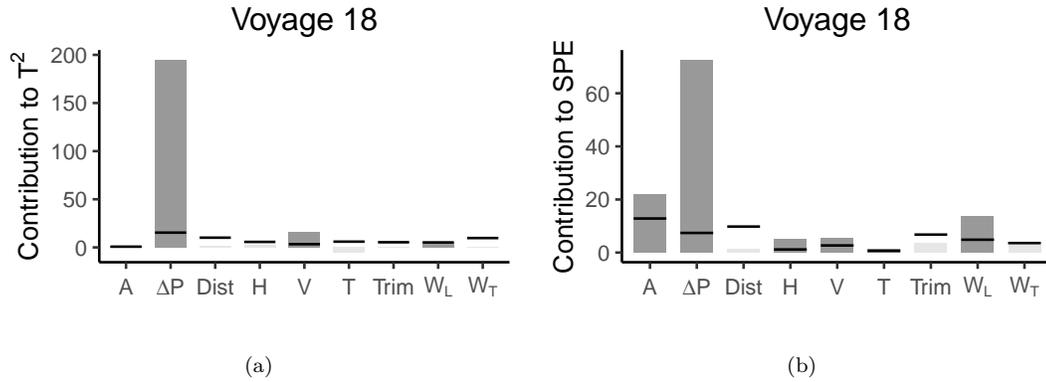


Figure 7.7: Contribution of the functional covariates to the Hotelling T^2 statistic (a) and SPE statistic (b) for voyage 18. The bars are the contributions of the variables, with the darker ones indicating values exceeding the limit, while the black dashes are the limits calculated on the basis of the reference voyages.

Voyage 19

As in the case of voyage 18, voyage 19 is signalled as OC by both T^2 and SPE functional control charts, and the higher-than-average value of the Hotelling T^2 statistic results in wider control limits for the response prediction control charts. The contribution plots to the T^2 , reported in Figure 7.9a, indicates the power difference between port and starboard propeller shafts (ΔP), cumulative sailing time (H), speed over ground (V), and acceleration (A) as anomalous. In addition to the latter variables, the contribution to the SPE statistic in Figure 7.9b also reports the wind components (W_L and W_T) slightly above their control limit. The plots of those functional covariates are explored in Figure 7.10 and evidence a SOG variable profile (Figure 7.10a) unusually low during the first part of the voyage and reflects also in atypical pattern of the acceleration profile (Figure 7.10c). Accordingly, the sailing time (Figure 7.10b) is higher than usual in the middle of the voyage to complete the voyage with no delay on arrival. The power difference between port and starboard propeller shafts (ΔP) (Figure 7.10d) helped to diagnose malfunctioning of one of the main engines in the initial part of the voyage.

Voyage 23

Voyage 23 has IC covariates (i.e., T^2 and SPE statistics do not exceed their control limits in Figure 7.3) but the response prediction error control chart indicates that the total CO_2 emissions are lower than the predicted value (Figure 7.3c). In this case, plausible fault causes are to be investigated outside the covariates included in Table 1.

Voyage 24

Voyage 24 is signalled by both T^2 and SPE functional control charts (Figure 7.3) as OC. Note that, as for voyage 18 and 19, the higher the value of the Hotelling T^2 statistic, the wider the corresponding interval of the response prediction error control chart. In Figure 7.11a, the contribution to the Hotelling T^2 statistic signals SOG (V), power difference between port and

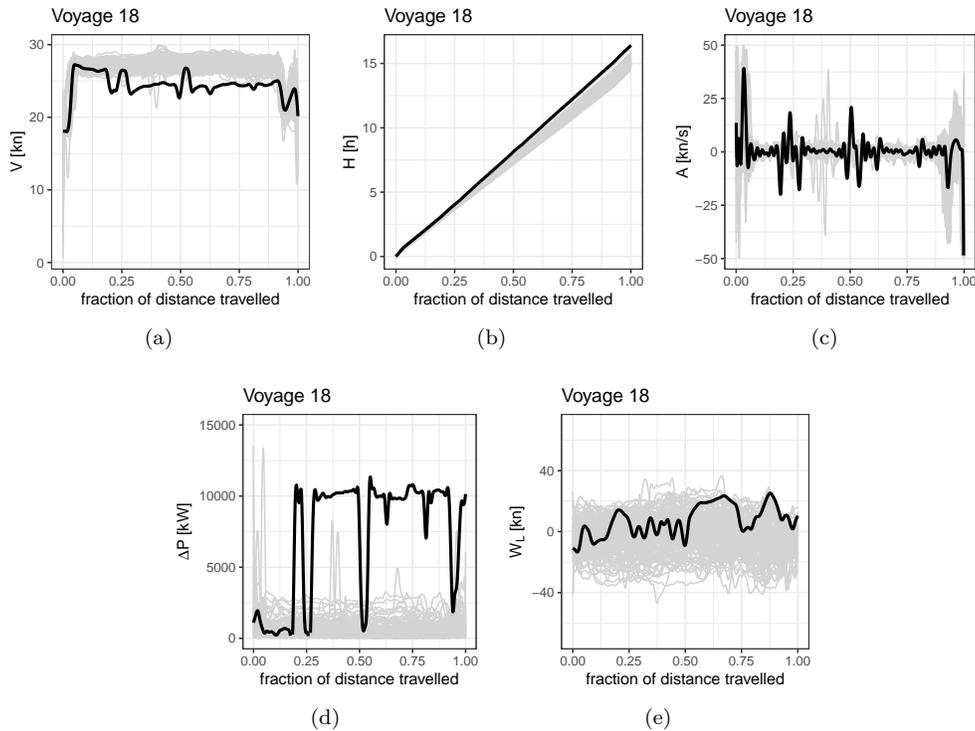


Figure 7.8: Observations of the critical (i.e., indicated as responsible of OC by contributions plots in Figure 7.7) functional covariates for Phase II monitoring of the voyage 18, viz. (a) SOG (V), (b) cumulative sailing time (H), (c) acceleration (A), (d) power difference between port and starboard propeller shafts (ΔP), and (e) longitudinal wind component (W_L). In each plot, the black line indicates the current observation, while the reference functional observations are plotted in grey.

starboard propeller shafts (ΔP), and acceleration (A) as anomalous variables. Additionally, in Figure 7.11b, the contribution to the SPE statistic also indicates the longitudinal wind component (W_L), and confirms V and ΔP . Plots of the functional covariates, shown in Figure 7.12, evidences unusually low SOG profile (Figure 7.12a) during the very beginning and final part of the voyage. Surprisingly, the acceleration variable is not indicated as anomalous, since the variations in SOG are not so pronounced. The power difference between port and starboard propeller shafts (ΔP) (Figure 7.10d) supported diagnosis of an anomaly in one of the main engines in correspondence of the part of the voyage with low SOG. Also in this case the longitudinal wind profile W_L shows larger than usual values at the beginning of the voyage.

Comparison with scalar method

To assess the performance of the proposed methodology and to point out the advantage of using a functional approach, we perform a comparison with a simpler model, where a single scalar measurement is considered for each covariate and each voyage. In particular,

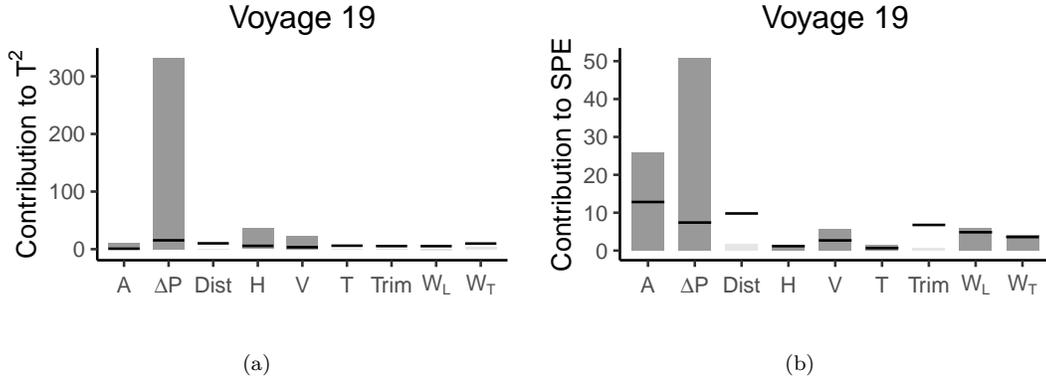


Figure 7.9: Contribution of the functional covariates to the Hotelling T^2 statistic (a) and SPE statistic (b) for voyage 19. The bars are the contributions of the variables, with the darker ones indicating values exceeding the limit, while the black dashes are the limits calculated on the basis of the reference voyages.

the proposed scalar model is as follows

$$y_i = \gamma_0 + \sum_{p=1}^P Z_{ip}\gamma_p + \varepsilon_i, \quad i = 1, \dots, n, \quad (7.29)$$

where $Z_{ip} = \int_0^1 X_{ip}(t)dt$ is the sample mean of the functional covariate X_p at voyage i , $\gamma_0, \gamma_1, \dots, \gamma_P$ are the regression coefficients, and ε_i is the error term. The multivariate monitoring of the scalar covariates can be performed by calculating the Hotelling T^2 statistic (Montgomery 2007), obtained as

$$T^2 = (\mathbf{Z} - \hat{\boldsymbol{\mu}}_Z)^\top \hat{\boldsymbol{\Sigma}}_Z^{-1} (\mathbf{Z} - \hat{\boldsymbol{\mu}}_Z), \quad (7.30)$$

where $\mathbf{Z} = (Z_1, \dots, Z_p)$, while $\hat{\boldsymbol{\mu}}_Z$ and $\hat{\boldsymbol{\Sigma}}_Z$ are the estimates of the mean and covariance matrix of \mathbf{Z} , respectively. The upper control limit of the control chart is calculated non-parametrically, as it is done for the functional control charts. Given a new observation $(\mathbf{Z}^{new}, y^{new})$, the scalar response variable can be monitored by means of the response prediction error control chart, by defining $y^{new} - \hat{y}^{new} = y^{new} - \hat{\gamma}_0 - \sum_{p=1}^P \mathbf{Z}_p^{new} \hat{\gamma}_p$, where $\hat{\gamma}_0, \hat{\gamma}_1, \dots, \hat{\gamma}_P$ are the least squares estimators of the regression coefficients. The control limits of the response prediction error control chart are the same as in Equation (7.19), where T^2 is calculated as in Equation (7.30).

Figures 7.13a and 7.13b show the scalar control charts for the Phase II voyages. Note that voyage 7 is not detected as anomalous in this case, as it is done in the SPE functional control chart. This gives a practical evidence of the advantages of the proposed functional monitoring strategy that is able to highlight anomalies that are present only in a small part of the functional domain (Figure 7.6). In order to compare the predictive performance of the two models, we use the mean squared error (MSE) of prediction calculated for Phase II voyages, using the proposed scalar-on-function model (7.1) and the scalar model (7.29). Note that in the MSE calculation for both models, we did not include the anomalous voyage 23 signalled as OC by the response prediction error control chart in Figure 7.3c. Unsurprisingly,

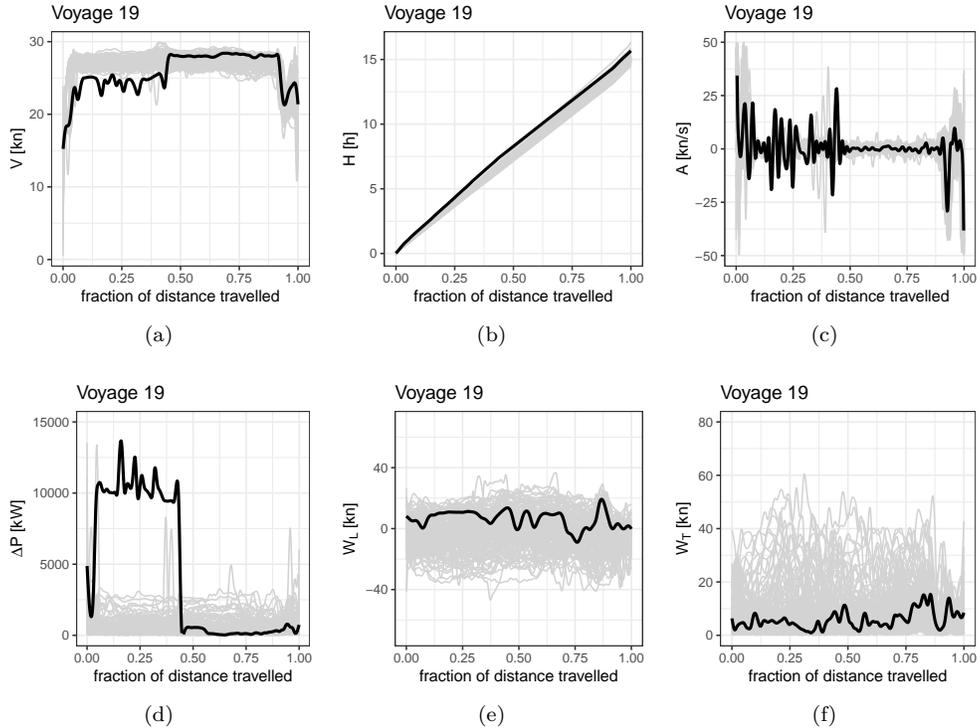


Figure 7.10: Observations of the critical (i.e., indicated as responsible of OC by contributions plots in Figure 7.9) functional covariates for Phase II monitoring of the voyage 19, viz. (a) SOG, (b) cumulative sailing time (H), (c) acceleration, (d) power difference between port and starboard propeller shafts (ΔP), (e) longitudinal wind (W_L), and (f) transverse wind (W_T). In each plot, the black line indicates the current observation, while the reference functional observations are plotted in grey.

the scalar-on-function model achieves an $MSE = 100.2$, that is comparatively much lower than the scalar model $MSE = 122.5$.

7.5 Functional real-time control charts

Up to this point, the proposed monitoring strategy presented in Section 7.3 assumed all functional covariates fully observed in their compact domain \mathcal{T} . By relaxing this assumption, we want to enlarge its use for functional real-time monitoring up to any intermediate domain point. In order to do that, a real-time warping procedure needs to be defined, and, without loss of generality, it will be presented in what follows with reference to the proposed real-case study. Hereinafter, the real-time version of the functional control charts are referred to as *functional* real-time control charts, where the term *functional* is used to emphasize that, instead of monitoring the instantaneous value of the covariates and response only, they exploit the new voyage information cumulated up to the current domain point.

Accordingly, let us denote with t^* the instant at which we want to apply the monitoring

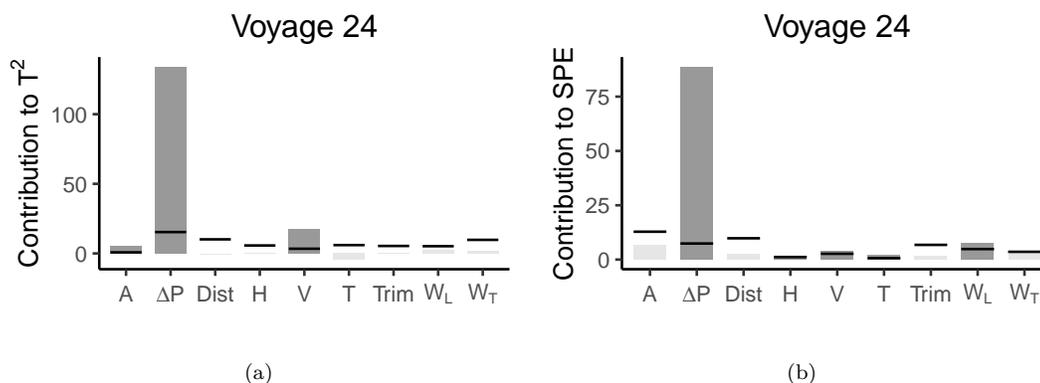


Figure 7.11: Contribution of the functional covariates to the Hotelling T^2 statistic (a) and SPE statistic (b) for voyage 24. The bars are the contributions of the variables, with the darker ones indicating values exceeding the limit, while the black dashes are the limits calculated on the basis of the reference voyages.

procedure and with $k^* \in \mathcal{T} = [0, 1]$ the corresponding fraction of distance travelled. The response variable to be monitored, denoted by y^* , is suitably chosen in this setting as the total CO_2 emissions up to t^* (and k^*). Trivially, when $k^* = 1$ (i.e., at the end of the voyage), y^* coincides with y^{new} . At any instant t^* , we use as real-time warping function the map $f: [0, t^*] \rightarrow [0, k^*]$ that associates to each $t \in [0, t^*]$ the corresponding fraction of distance travelled as $k(t) = d(t)/d$, where $d(t)$ is the distance travelled until $t \leq t^*$ and d is the total distance travelled at the end of the voyage. Unfortunately, as d is not known yet at t^* , the following steps are required to calculate $k(t)$ and thus $k^* = k(t^*)$:

1. consider the current GPS position P^* of the ship, given by its longitude and latitude;
2. identify the point \bar{P}^* on the nominal route as that with minimal distance from the current position of the ship at the considered instant t^* (Figure 7.14);
3. calculate the fraction of distance travelled at t^* as $k^* = d^*/d \in [0, 1]$, where d^* is the length of the nominal route from departure port to \bar{P}^* and d is the length of the whole nominal route.

The reference data set at t^* can be obtained by truncating the reference observations of covariates at k^* , such that the new functional domain is $[0, k^*]$. The reference data set so obtained can be used to repeat the Phase I, described in Section 7.3, to calculate the monitoring statistics at k^* and the limits of the corresponding control charts. In particular, note that for every k^* the limits of the T^2 and SPE control charts do not change in Phase II from voyage to voyage. Therefore, they need to be calculated at the end of Phase I only once for k^* values varying on an appropriate discretization of the domain $\mathcal{T} = [0, 1]$. In particular, in this real-case study we consider 50 equally spaced values between 0.15 and 1, where 0.15 is arbitrarily chosen as burn-in domain portion. Contrarily, the limits of the prediction error control chart depend also on the current covariate profiles (i.e., the profiles observed at k^*), and must be calculated in real time for each voyage in Phase II. In view of

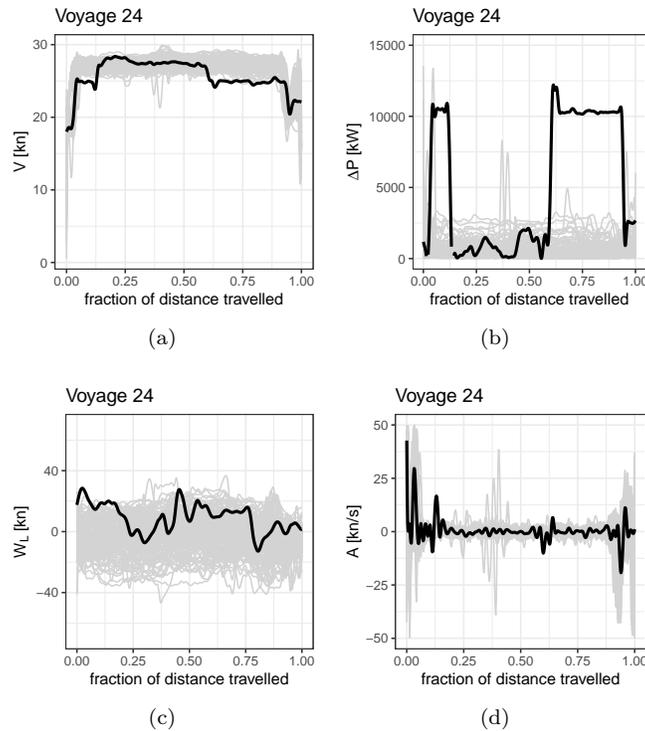
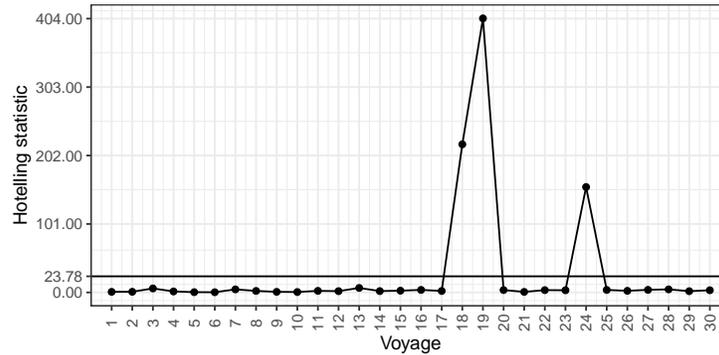


Figure 7.12: Observations of the critical (i.e., indicated as responsible of OC by contributions plots in Figure 7.11) functional covariates for Phase II monitoring of the voyage 24, viz. (a) SOG (V), (b) power difference between port and starboard propeller shafts (ΔP), (c) longitudinal wind component (W_L), and (d) acceleration (A). In each plot, the black line indicates the current observation, while the reference functional observations are plotted in grey.

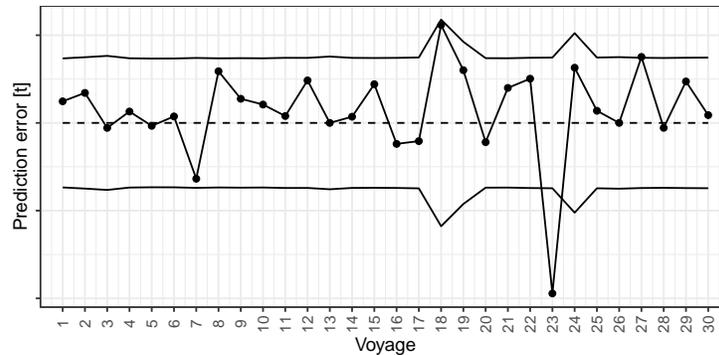
this, most of the computational effort is required only at the end of Phase I and the real-time monitoring is fast in practice.

As an example, in Figure 7.15 we report the functional real-time control charts built for those voyages that we already signalled as OC by at least one of the control charts in Figure 7.3 at the end the voyage (i.e., $k^* = 1$), viz. voyages 7, 18, 19, 23 and 24. By those voyages, we aim to demonstrate the straightforward real-time use of the proposed monitoring strategy in tracking dynamic scenarios that may occur during navigation and in supporting more prompt diagnosis of faults. Moreover, in what follows we compare the behaviour of the functional real-time control charts of Figure 7.15 with profiles of critical covariates indicated as responsible for the OC signal by contributions plots reported in Section 7.4.

Specifically, in Figure 7.15a, the functional real-time monitoring of voyage 7— signalled as OC only by the SPE control chart of Figure 7.3—exhibits the real-time profile of the SPE statistic out of the control limit on the entire domain. In addition, the functional real-time T^2 control chart in Figure 7.15a is able to indicate also anomalous T^2 values at the beginning of the voyage ($k^* \leq 0.25$). This is coherent with Figure 7.6, where critical



(a)



(b)

Figure 7.13: (a) Scalar control chart for the covariates based on the Hotelling statistic (7.30) and (b) scalar control chart for the response prediction error used for monitoring the Phase II voyages. In each control chart, points joint by a line indicate monitoring statistic values at each voyage, while solid lines indicate control chart limits.

covariates exhibit anomalous behaviour at the beginning of the voyage.

Similarly, T^2 and SPE profiles of voyages 18, 19, and 24—signalled as OC by both T^2 and SPE control charts of Figure 7.3—plot predominantly OC, although with different patterns (Figures 7.15b, 7.15c, and 7.15e). For voyage 18, the functional real-time T^2 and SPE control charts reported in Figure 7.15b signal OC except for the very beginning of the voyage. This is coherent with Figure 7.8, where most of the the observation of critical covariates lie far from the reference data set from values of k^* approximately larger than 0.20. Similarly, for voyage 19, T^2 and SPE real-time profiles plot out of the control limits. Note that, while the SPE statistic is always much larger than the control limit, the T^2 statistic shows a decreasing trend. This is in accordance with the critical covariates reported in Figure 7.10, where all covariates plot close to the reference dataset only in the second half of the voyage. The straightforward interpretation of the functional real-time control charts applies also for voyage 24. Indeed, Figure 7.15e shows the real-time SPE statistic much larger than its control limit on the entire domain, whereas the T^2 goes very close to its control limit in the middle part of the voyage This behaviour iagrees, as in the cases before,

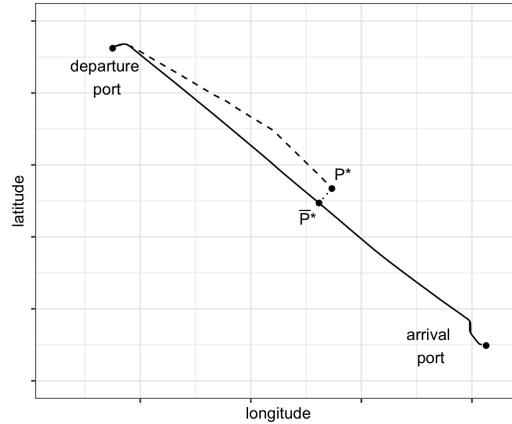


Figure 7.14: Graphical example showing how $k^* = d^*/d$ is determined for a new voyage. The dashed curve represents the route travelled by the ship up to the current GPS position, which is the point labeled as P^* . The solid curve represents the nominal route and its point nearest to P^* is labeled as \bar{P}^* . d^* is the length of the portion of the solid curve from the departure port to \bar{P}^* . d is the total length of the nominal route.

with covariate profiles (Figure 7.12). Differently from voyage 7, the real-time prediction of the cumulative CO₂ emissions is consistent with the data even though functional covariates are predominantly in OC state. This is displayed in the third column panels of voyages 18, 19 and 24 in Figure 7.15, where the actual response plots within the corresponding prediction interval ($\alpha_y = \alpha/2 = 0.025$). Instead, note that voyage 23—which has T^2 and SPE statistics within the control limits during the whole voyage (see Figures 7.3 and 7.15d)—has actual cumulative CO₂ emissions below the lower limit throughout the voyage. This supports practitioners to plausibly investigate anomaly causes not related to the covariates included in the model (i.e., utilized to characterize ship operational conditions).

As theoretically expected, we can observe that control limits of the response prediction error control chart become wider as k^* increases, because the response (total CO₂ emissions) is chosen as cumulative up to t^* .

In conclusion, even if the considered real-case study has most of the functional covariates with smooth behaviours (e.g., they are generally constant or linear over the entire voyages), the proposed functional real-time control charts are able to detect anomalies more promptly ($k^* \leq 1$) than the monitoring performed only at the end of each voyage ($k^* = 1$) examined in Section 7.4. Those advantages are expected to definitely increase in more dynamic scenarios.

7.6 Conclusion

The need of handling complex data from modern ship multi-sensor systems have naturally called for the implementations of new statistical methodologies that extend the multivariate monitoring techniques to the case of multivariate functional data. In this work, a joint real-

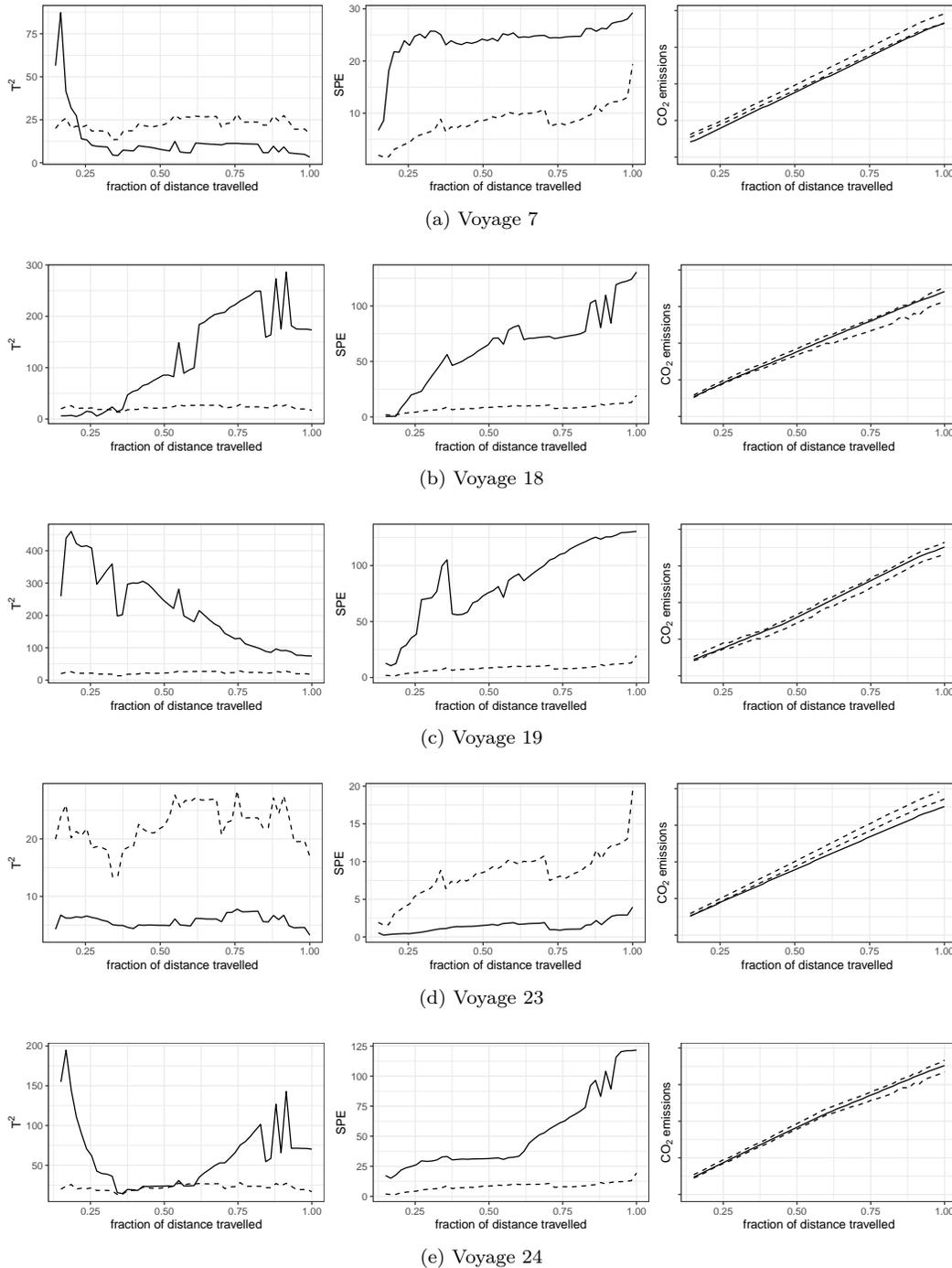


Figure 7.15: Real-time monitoring of voyages 7 (a), 18 (b), 19 (c), 23 (d) and 24 (e). For each voyage, the T^2 , SPE , and response prediction error control charts are plotted. In each plot, solid lines indicate the profiles of the monitoring statistics during the voyage, and dashed lines indicate the corresponding control limits as in Equation (7.23), with $\alpha = 0.05$.

time monitoring procedure for functional covariates and a scalar response related to them, is proposed, which, to the best of the authors' knowledge, is new in both statistical and maritime field. A suitable landmark registration is proposed to warp all functional observations into the same domain. Signals acquired from different kind of sources and with different units of measurement are shown to be easily integrated into a multivariate functional regression model. Besides, the joint use of the Hotelling T^2 and squared prediction error functional control charts, estimated by means of multivariate functional principal component analysis, is shown to be able to effectively monitor the ship operating conditions of the upcoming voyages and to highlight unusual behaviour with respect to a reference-good data set of past voyages. In case of an out-of-control signal, the corresponding contribution plots are demonstrated to be powerful tools for supporting diagnosis of faults. The optimal choice of the functional principal components to retain has been also discussed with the aim of considering the variability in the covariates that is beneficial for the response prediction performance. Then, by means of response prediction error control chart, the proposed procedure is demonstrated to be able to monitor ship CO₂ emissions and to plausibly indicate if an anomaly occurs in the scalar-on-function linear model, i.e., outside the covariates that characterize ship operational conditions acquired on board and included as covariates. To allow the joint use of the three control charts, control limits have been opportunely corrected so that the type-I family-wise error rate achieves at most a fixed significance level. The problem of multiple comparison is addressed to plot the limits of the contribution plots, in a fully real-time scenario, which is itself an issue rarely discussed in the mainstream literature. The proposed monitoring procedure is in fact shown to be applicable also for functional real-time monitoring up to any intermediate time domain point of both functional covariates and scalar response. This can greatly help shipping practitioners to support real-time managerial decision making by promptly indicating anomalies during the navigation phase. The advantages of the proposed functional real-time monitoring procedure are expected to escalate in those scenarios where the functional data reveal very dynamic behaviours.

Chapter 8

The R package `funcharts`

The goal of `funcharts` is to provide control charts for functional data. The methodology is described in Chapter 7.

8.1 Installation

You can install the development version from GitHub with:

```
# install.packages("devtools")
devtools::install_github("unina-sfere/funcharts")
```

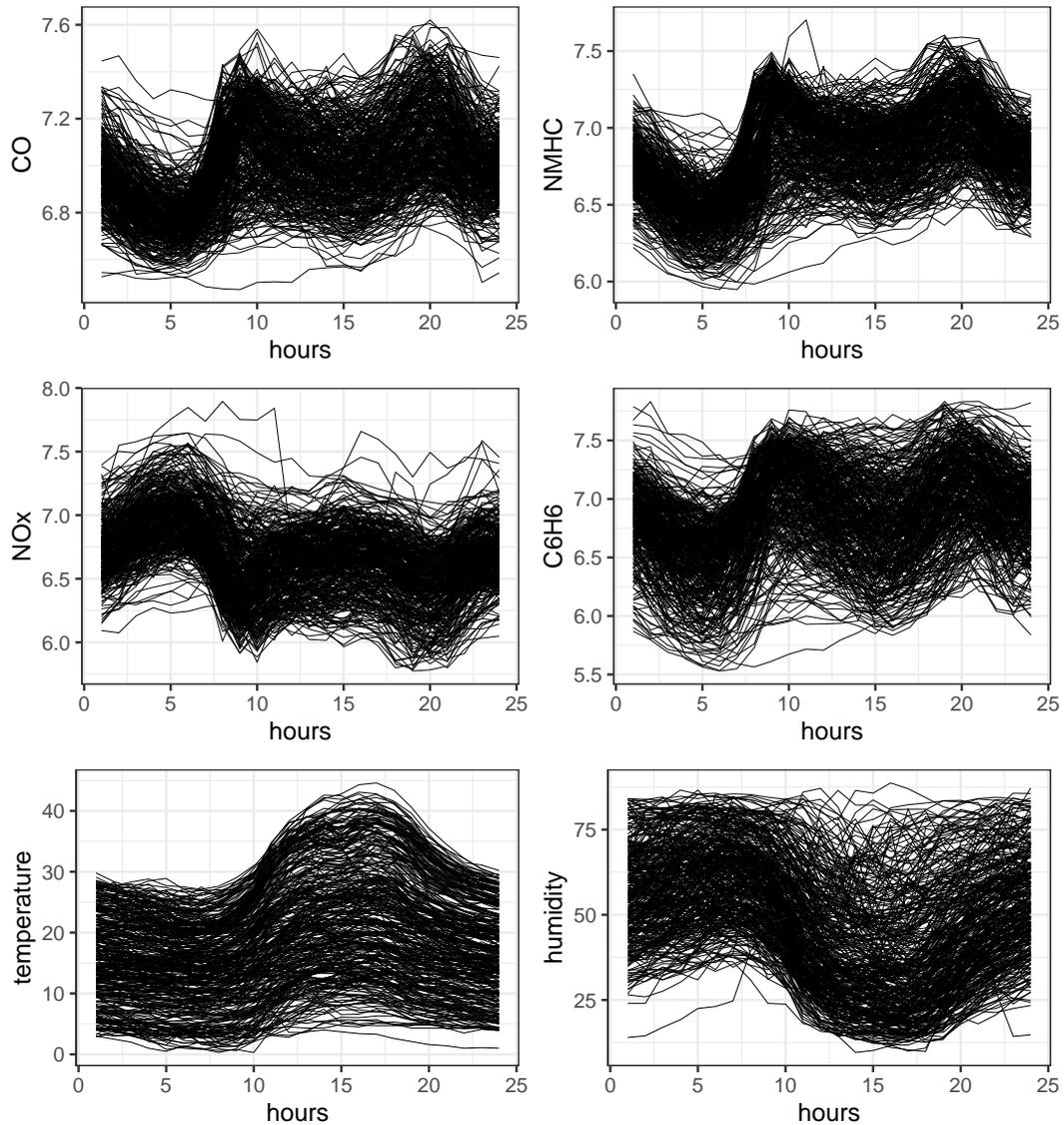
8.2 Usage of `funcharts`

Let us show how the `funcharts` package works through an example with the dataset `air` from the package `FRegSigCom` used in the paper of Qi & Luo (2019). First of all, let us arrange the data in an appropriate format, we need a dataframe with the following columns:

- an `id_variable` column, identifying the observation in the functional dataset;
- a `domain_var` column, identifying the variable giving the domain point;
- then we need one column for each functional variable.

Qi & Luo (2019) propose a function-on-function regression model of the `N02` functional variable on all the other functional variables available in the dataset. In order to show the package, we consider a scalar-on-function regression model, where we take the mean of `N02` at each observation as scalar response and all other functions as functional covariates. Since for each functional variable data are arranged in a matrix format, we can easily obtain the long format using `tidyverse` through the function `pivot_longer` or `gather`. Let us also plot the functions (not smoothed) to see the functional variables.

```
library(FRegSigCom)
library(tidyverse)
library(funcharts)
library(gridExtra)
library(grid)
df <- lapply(seq_along(air), function(ii) air[[ii]] %>%
  t %>%
  as.data.frame %>%
  setNames(1:ncol(.)) %>%
  add_column(hours = 1:24) %>%
  pivot_longer(- hours, names_to = "rep",
  values_to = names(air[ii])) %>%
  mutate(rep = as.numeric(rep))
  ) %>%
  bind_cols %>%
  dplyr::select(hours, rep, names(air)) %>%
  mutate(rep = factor(rep)) %>%
  group_by(rep) %>%
  mutate(NO2 = mean(NO2)) %>%
  ungroup
fun_covariates <- c("CO", "NMHC", "NOx", "C6H6", "temperature", "humidity")
lapply(fun_covariates,
  function(var) df %>%
    ggplot +
    geom_line(aes_string("hours", var, group = "rep"), size = .1) +
    theme_bw()
  ) %>%
  do.call(grid.arrange, .)
```



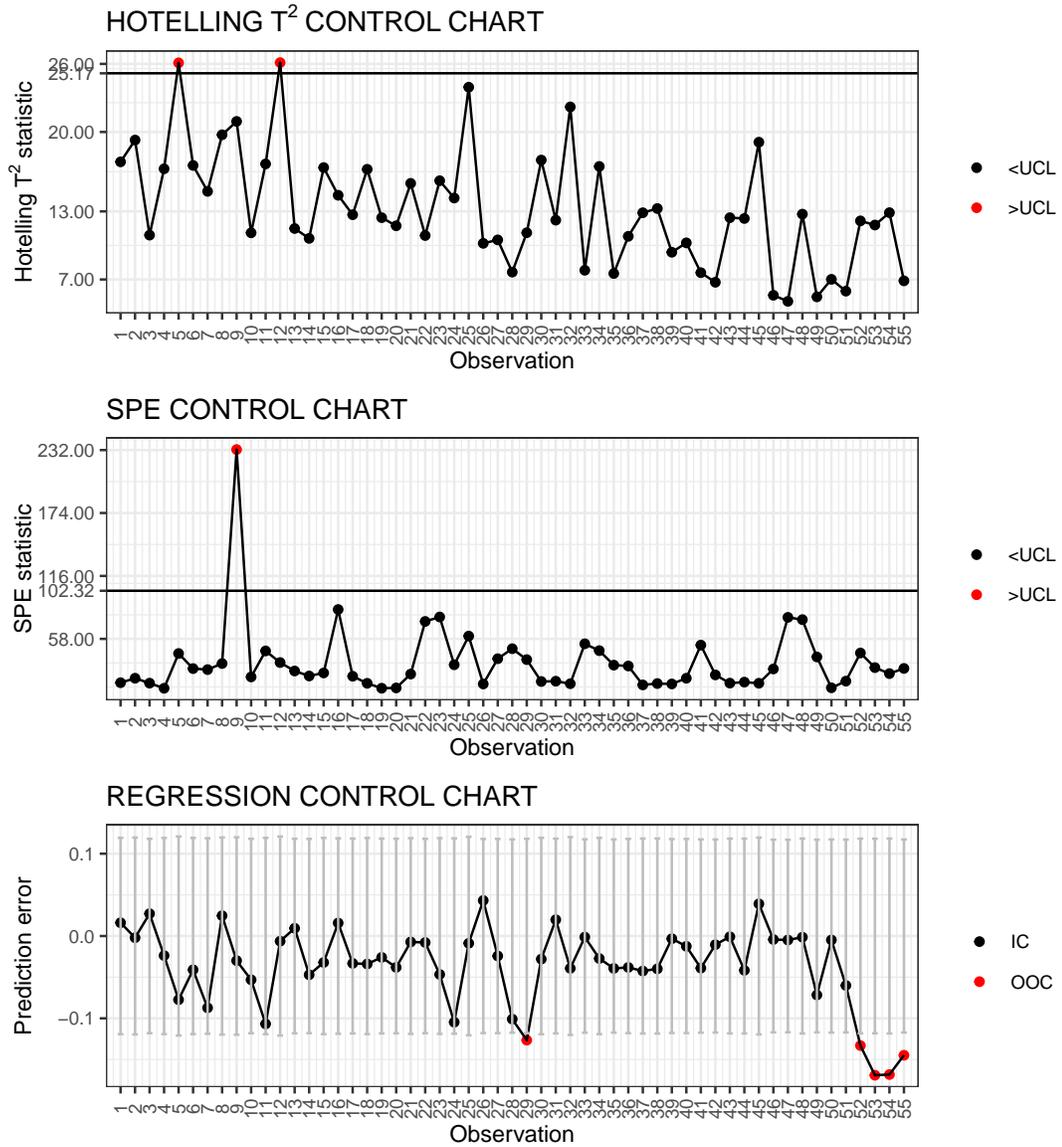
Now we split observations into a phase I training/calibration set and a phase II monitoring set, then we can use the funchart package and perform the analysis using the function `do_analysis`.

```
obs_list <- list(list(cal = factor(1:300), mon = factor(301:355)))

results <- do_analysis(
  dt = df,
  domain = c(1, 24),
  n_basis = 40,
  domain_var = "hours",
  id_variable = "rep",
```

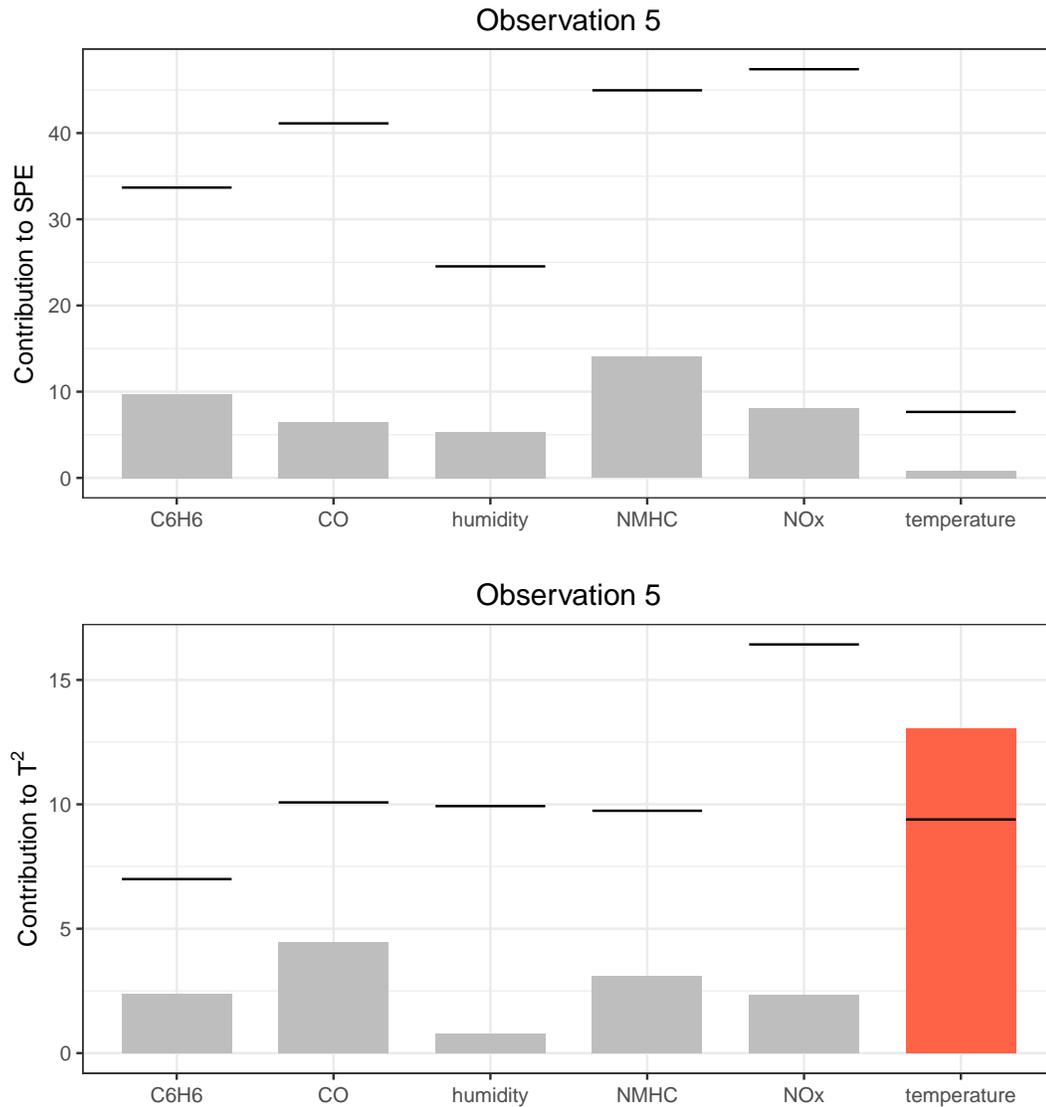
```
id_group = NULL,
variables = fun_covariates,
ncores = detectCores(logical = FALSE) - 1,
loglambda_search = seq(from = - 10, to = 0, by = 1),
scalar_response = "NO2",
identify_outlier = FALSE,
confidence.interval = FALSE,
obs_list = obs_list,
selection = "PRESS",
variance_explained = 0.8,
alpha = 0.05,
min_var_explained = 0.01
)

gtable_rbind(
  ggplotGrob(results$monitoring_list$plot$plot.list$`1`$plot.hot),
  ggplotGrob(results$monitoring_list$plot$plot.list$`1`$plot.spe),
  ggplotGrob(results$monitoring_list$plot$plot.list$`1`$plot.y),
  size = "first") %>%
  grid.draw
```



From the control charts it is possible to see phase II observations 5, 7, and 12 (they are actually days 305, 307 and 312) out of control. Let us see contributions for observation 5

```
do_cont_plot(result_phase1_list = results$result_phase1_list,
             monitoring_list = results$monitoring_list,
             id_observation = 5) %>%
lapply(function(x) x[[1]] %>% ggplotGrob) %>%
do.call(gtable_rbind, .) %>%
grid.draw
```



For the 5-th observation in the phase II dataset, we have the `temperature` variable causing the out of control for the Hotelling control chart.

```

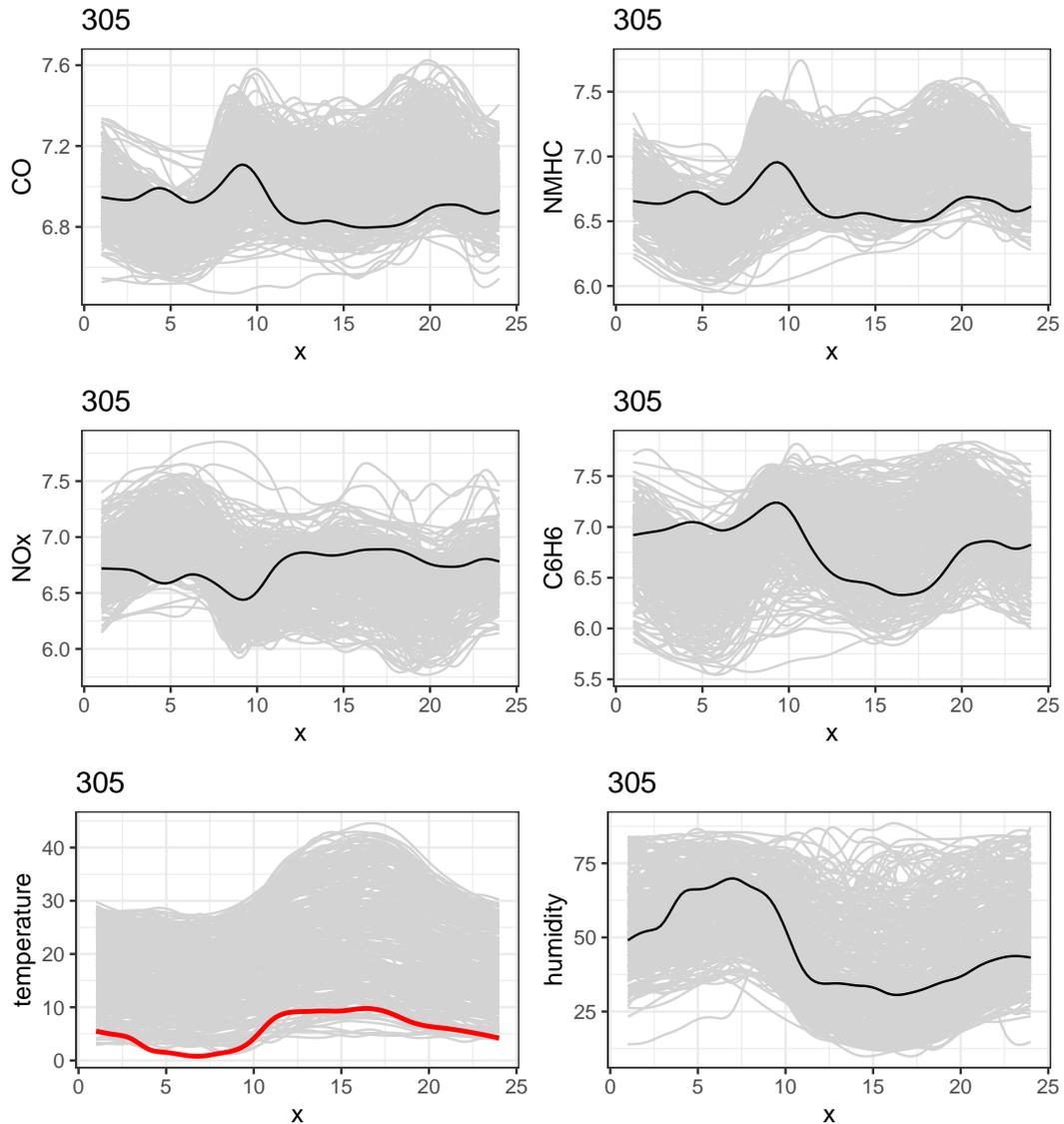
fault_vars <- "temperature"
fun_covariates %>%
  lapply(function(var)
    plotfunmon(
      fobj_segm_cal_mon = results$fobj_segm_cal_mon,
      group = 1,
      variable = var,
      obs = 5,
      col = if (var %in% fault_vars) "red" else "black",
      lwd = if (var %in% fault_vars) 1 else .5
    )
  )

```

```

)) %>%
do.call(grid.arrange, .)

```

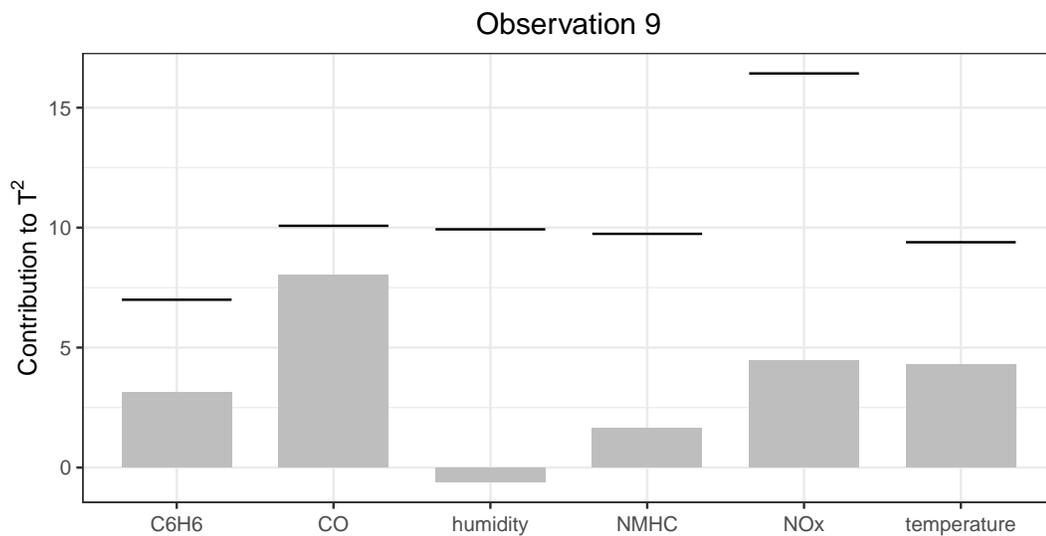
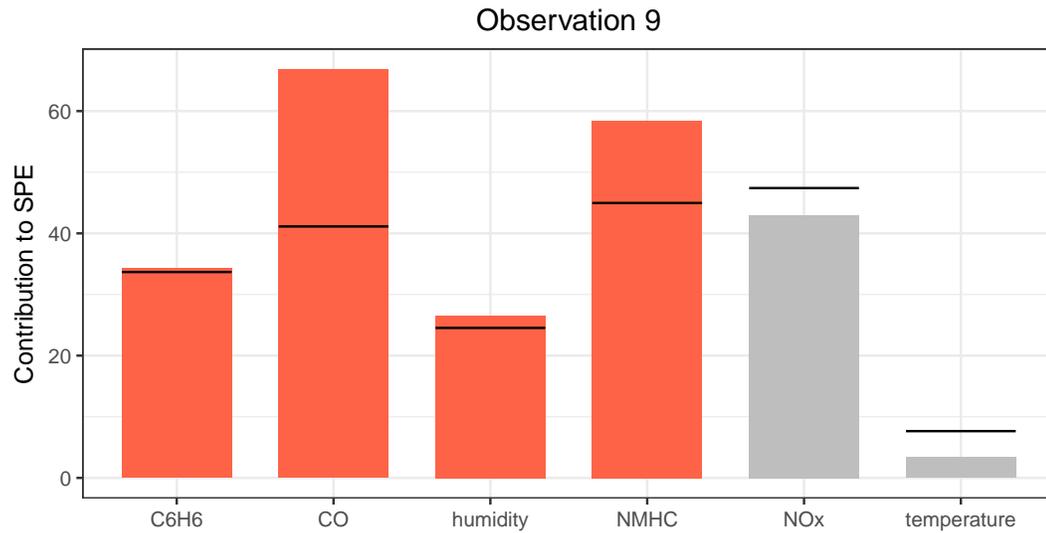


Let us see contributions for observation 9:

```

do_cont_plot(result_phase1_list = results$result_phase1_list,
             monitoring_list = results$monitoring_list,
             id_observation = 9) %>%
lapply(function(x) x[[1]] %>% ggplotGrob) %>%
do.call(gtable_rbind, .) %>%
grid.draw

```



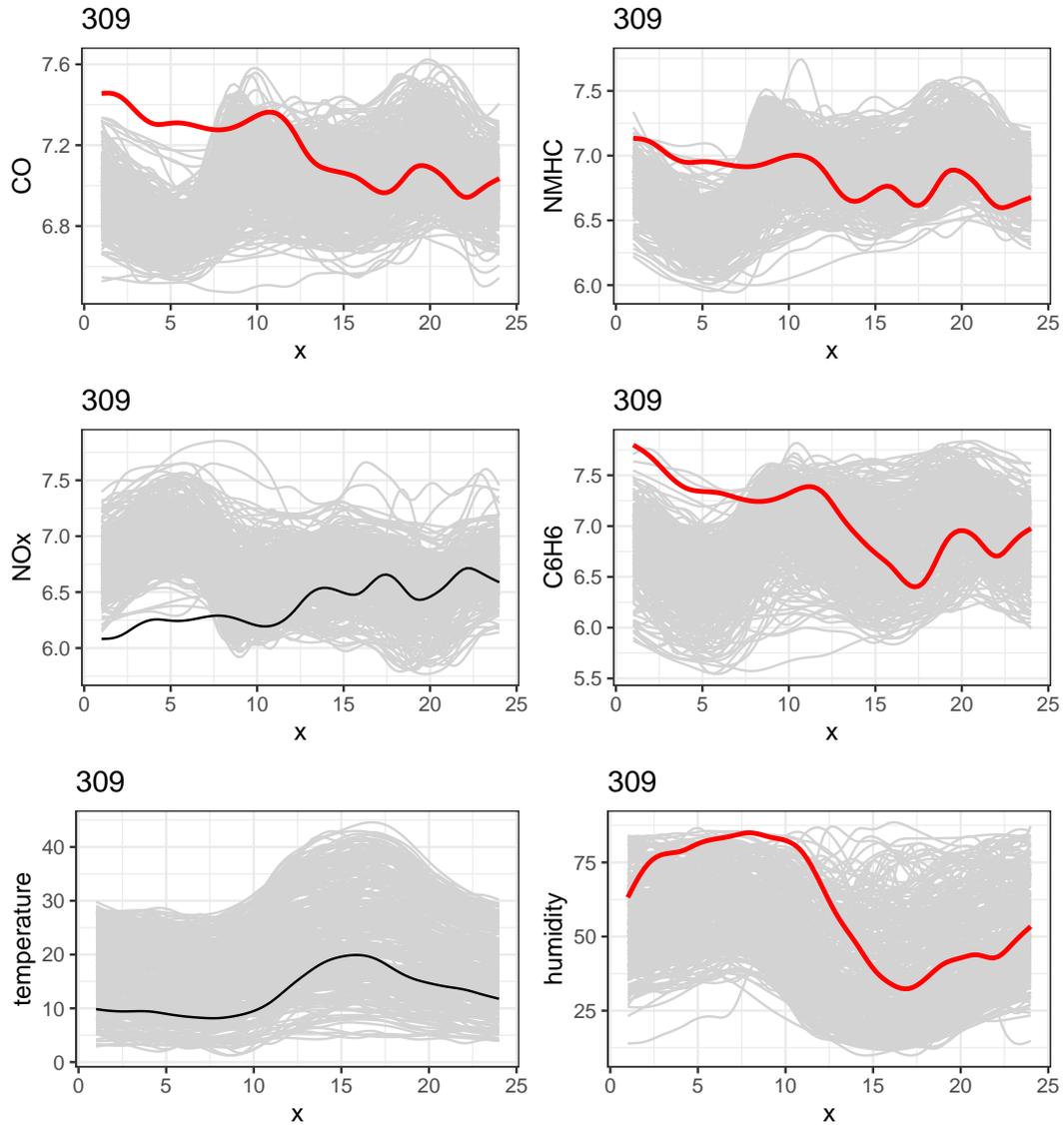
Many variables contribute to SPE

```

fault_vars <- c("C6H6", "CO", "humidity", "NMHC")
fun_covariates %>%
  lapply(function(var)
    plotfunmon(
      fobj_segm_cal_mon = results$fobj_segm_cal_mon,
      group = 1,
      variable = var,
      obs = 9,
      col = if (var %in% fault_vars) "red" else "black",
      lwd = if (var %in% fault_vars) 1 else .5
    ) %>%

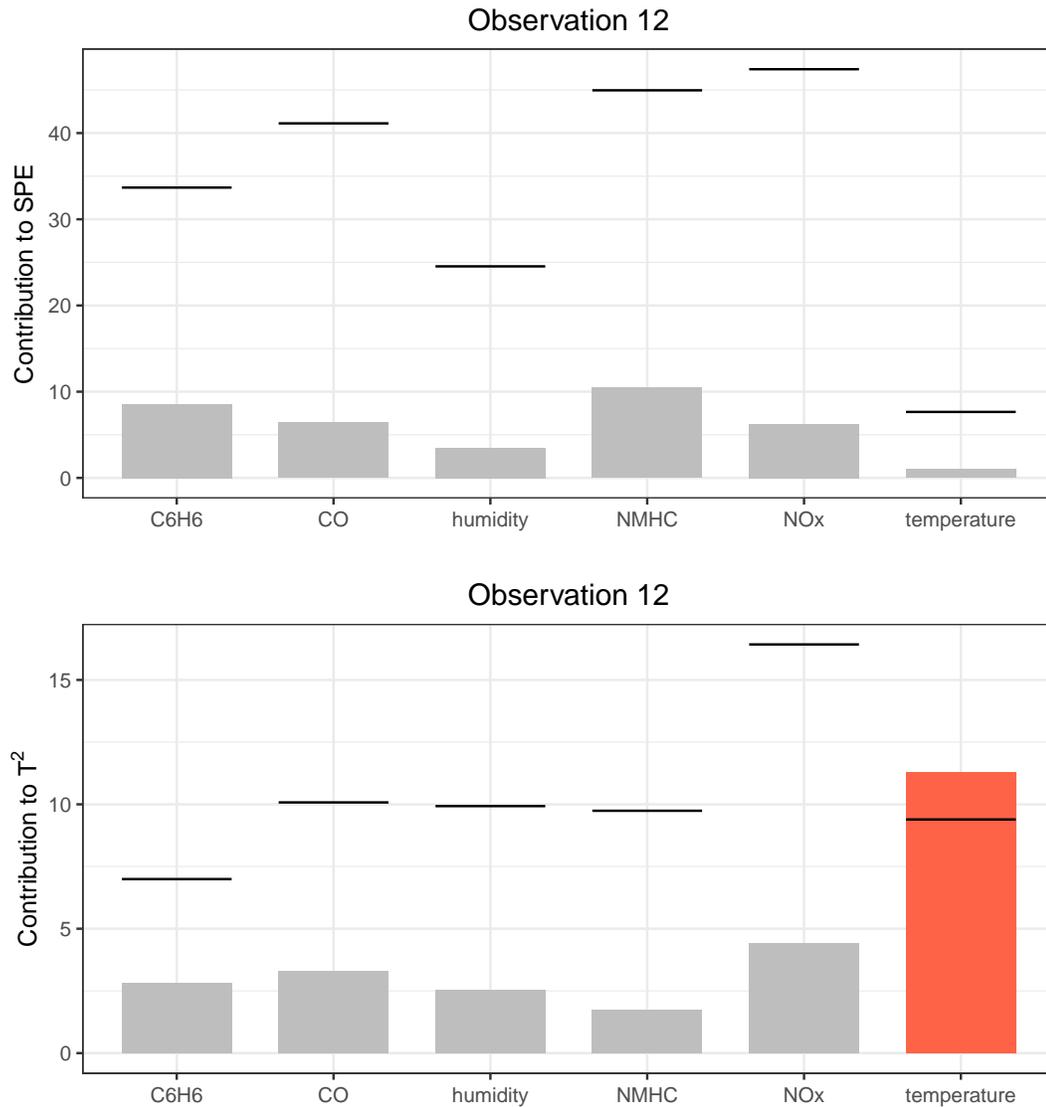
```

```
do.call(grid.arrange, .)
```



Let us see contributions for observation 12:

```
do_cont_plot(result_phase1_list = results$result_phase1_list,
             monitoring_list = results$monitoring_list,
             id_observation = 12) %>%
  lapply(function(x) x[[1]] %>% ggplotGrob) %>%
  do.call(gtable_rbind, .) %>%
  grid.draw
```



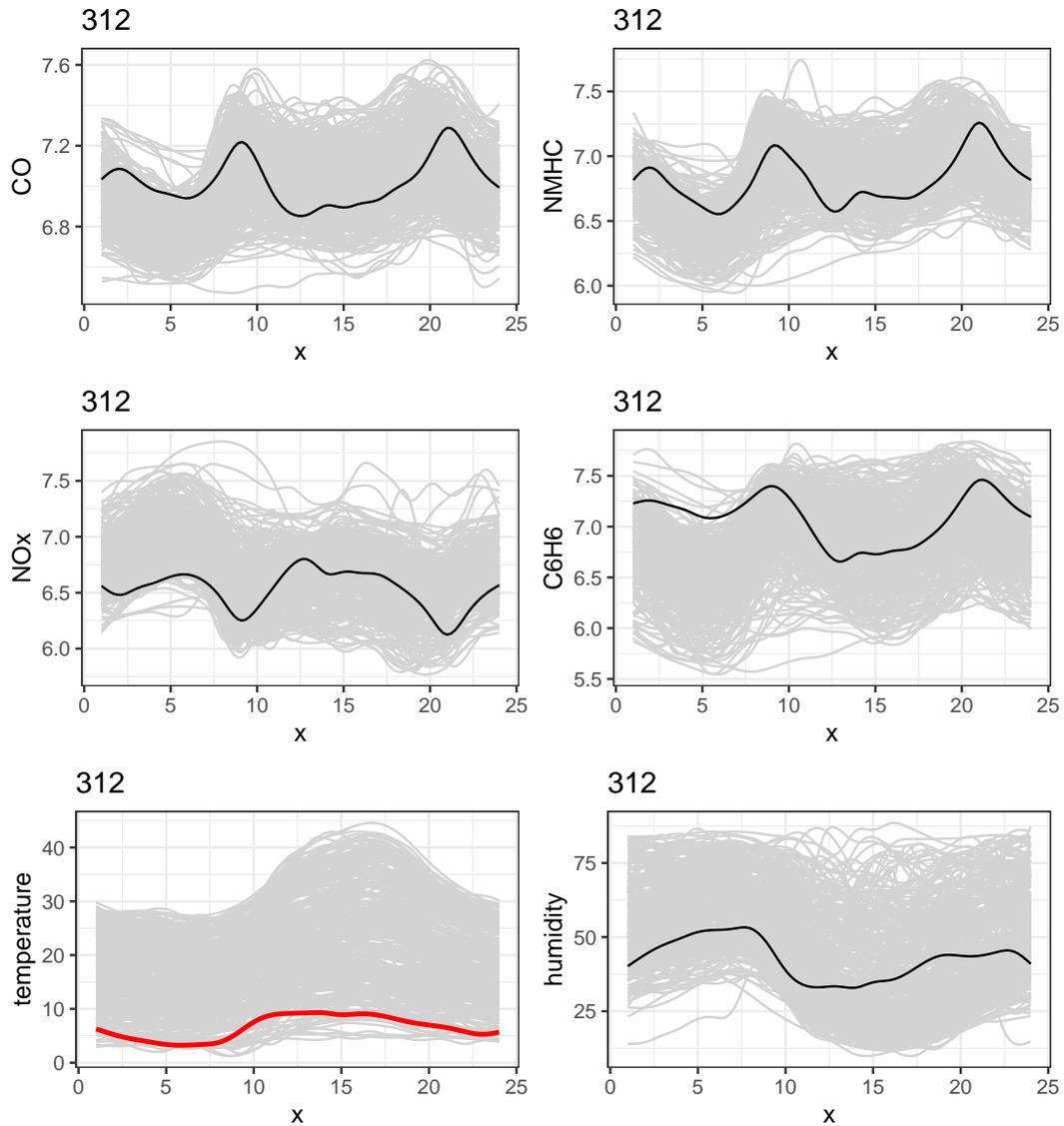
Again we have temperature signaled:

```

fault_vars <- "temperature"
fun_covariates %>%
  lapply(function(var)
    plotfunmon(
      fobj_segm_cal_mon = results$fobj_segm_cal_mon,
      group = 1,
      variable = var,
      obs = 12,
      col = if (var %in% fault_vars) "red" else "black",
      lwd = if (var %in% fault_vars) 1 else .5
    ) %>%

```

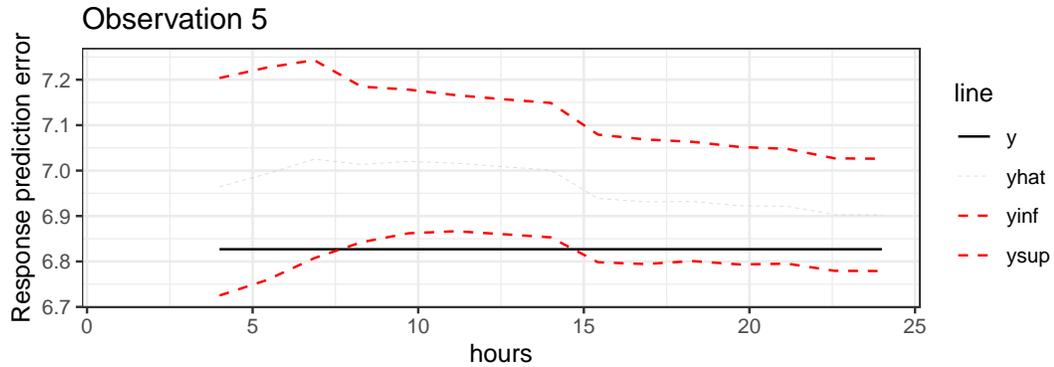
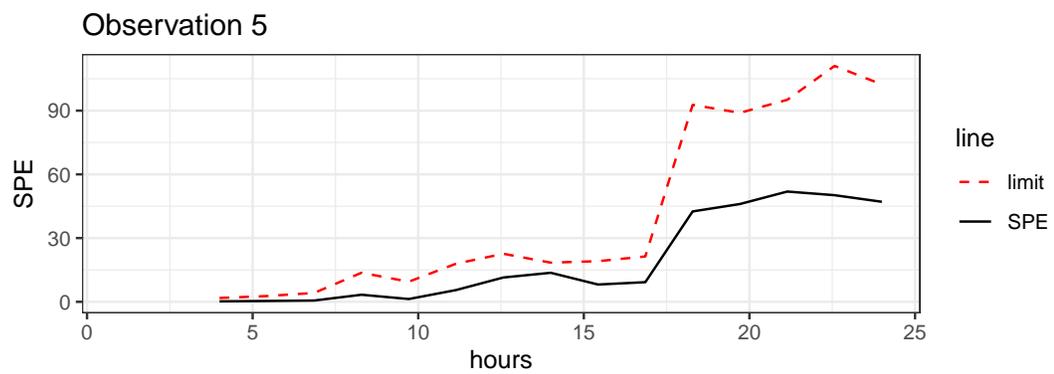
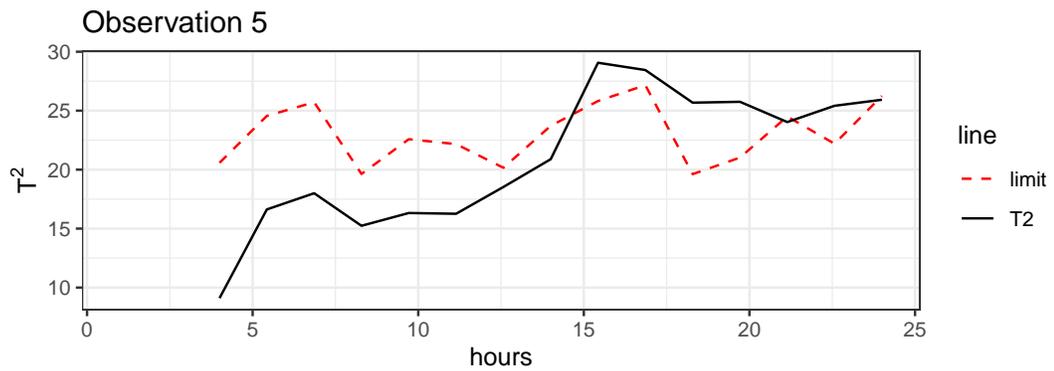
```
do.call(grid.arrange, .)
```



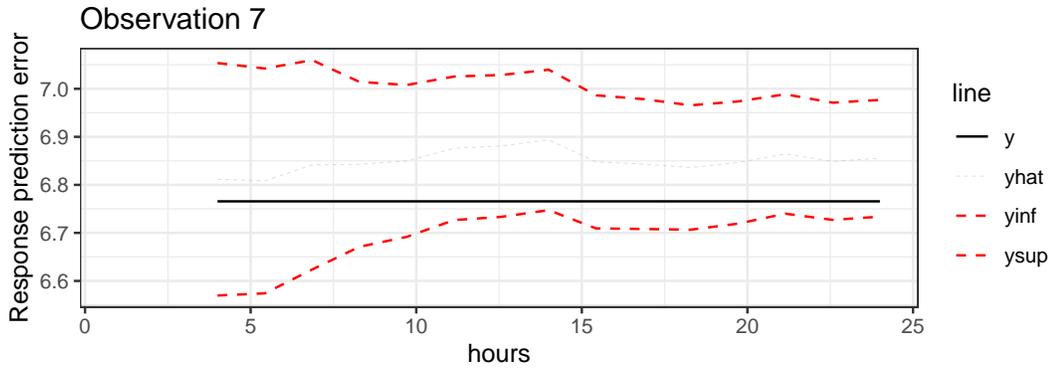
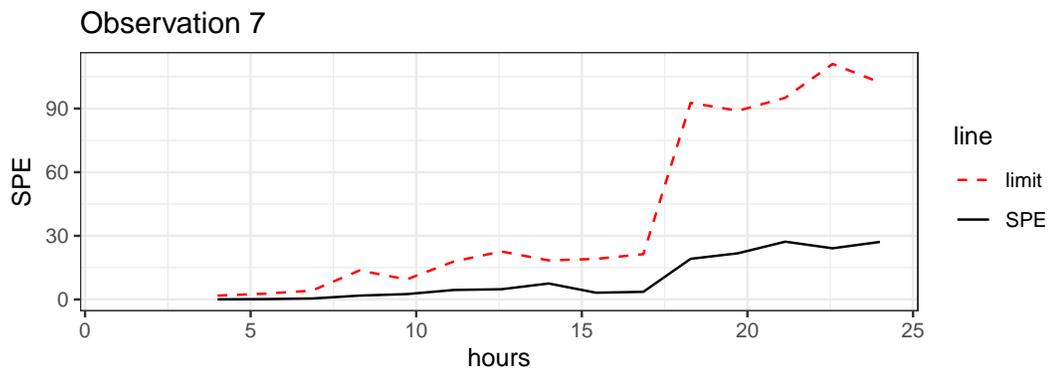
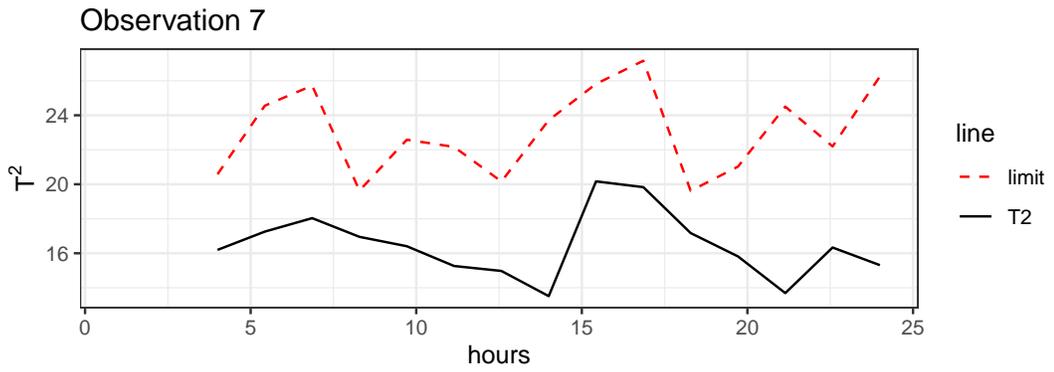
Now we show the real-time monitoring results for the same voyages.

```
## Real time
out_realttime <- do_real_time(
  kk = seq(from = 4, to = 24, length.out = 15),
  dt = df,
  domain = c(1, 24),
  n_basis = 40,
  domain_var = "hours",
  id_variable = "rep",
```

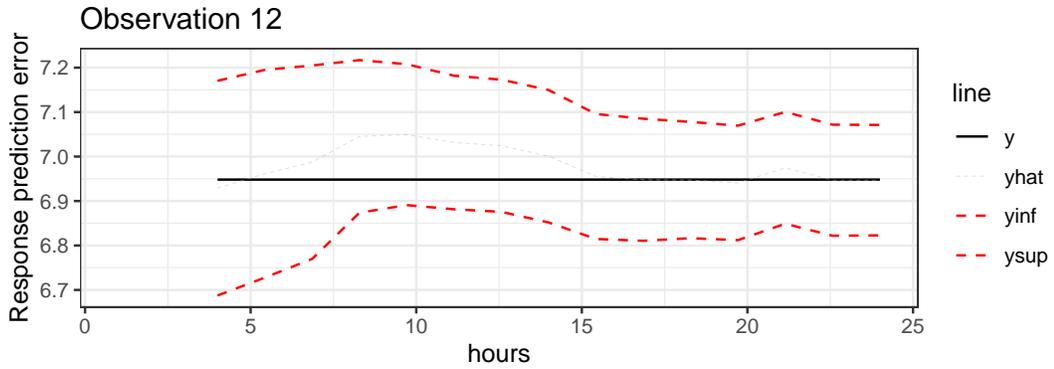
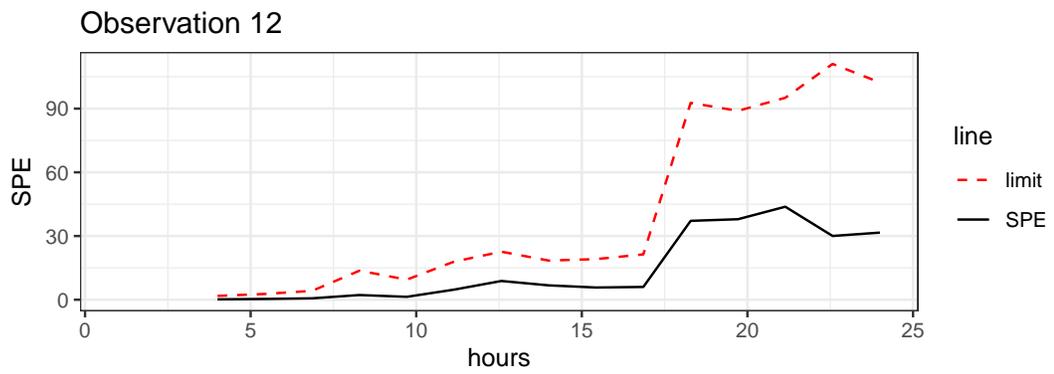
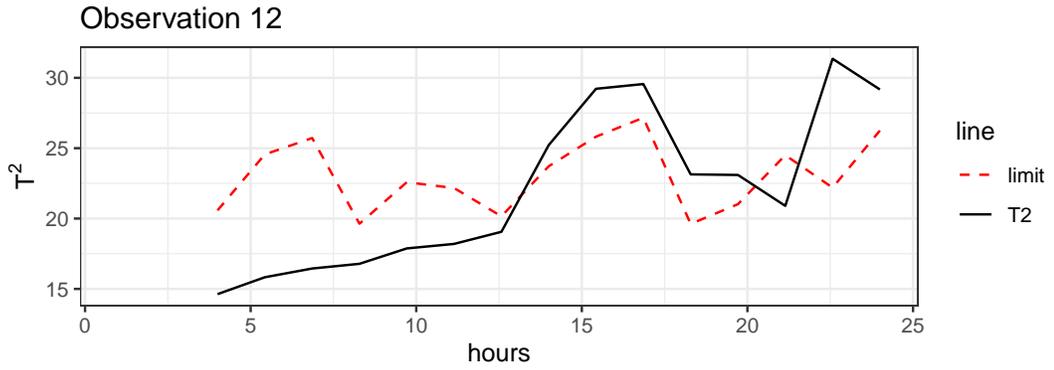
```
id_group = NULL,
variables = fun_covariates,
scalar_response = "NO2",
lambda = NULL,
loglambda_search = seq(from = - 10, to = 0, by = 1),
data_mon = NULL,
identify_outlier = FALSE,
confidence.interval = FALSE,
obs_list = obs_list,
selection = "PRESS",
variance_explained = 0.8,
alpha = 0.05,
n_fold = 5,
min_var_explained = 0.01,
ncores = detectCores()
)
grid.draw(plotrealtime(5, out_realtime$`1`))
```



```
grid.draw(plotrealtime(7, out_realtime$`1`))
```



```
grid.draw(plotrealtime(12, out_realtime$`1`))
```



Conclusion

In this thesis, methods for industrial process monitoring based on modern multivariate regression techniques and functional data analysis have been proposed. The proposed methodologies are shown to achieve accurate predictions of the CO₂ emissions or fuel consumption and, at the same time, to be capable of interpretable indications about the main causes of anomalies, possibly in real-time. This latter aspect is the main advantage over the existing methods, especially in the case of profile monitoring in the functional data analysis context. The proposed methodologies have the aim to provide results that are interpretable for the practitioner and are able to support the technical management in the decision-making process, which is an advantage over the black-box artificial intelligence alternatives. Finally, open-source software is provided that not only is available to practitioners in the maritime transportation sector, but it also generalizes to any application in the industrial process monitoring setting where a scalar quantity depends on a set of functional variables.

Regarding the future directions of the research, there are several ways in which the work proposed in this thesis can be improved, they are listed below.

Monitoring and prediction of a functional response variable

The focus of the proposed thesis is on a scalar response variable that is summarized into a single statistic, usually the total amount of CO₂ emissions at the end of a voyage. We think that it is valuable to work directly with the single value of the scalar response variable at the end of each voyage, because the specific application of this thesis is the prediction and monitoring of the total amount of CO₂ emissions of a ship due to navigation at each voyage. Then, especially in the case of retrospective monitoring, the main interest of the practitioner is in the cumulative quantity and on whether this quantity was particularly high. However, being able to extract for example the information of the effects of various factors on the emissions at different stages of a voyage is of great interest as well. This can be achieved through function-on-function regression, which is not considered in this thesis.

Functional time series

All the models proposed in this thesis assume that observations of the response variable are statistically independent conditionally on the covariates, both in the scalar and in the functional case. We believe that, conditionally on the covariates, the main temporal dependence structure is the one observed within each single voyage, which is modeled well when treating data as profiles. However, functional time series techniques may further allow to consider the temporal dependence across different voyages, if any, and improve the prediction and monitoring performance.

Functional generalized additive models

Most of the regression models proposed in this thesis assume there exists a linear relationship between the functional covariates and the scalar response. Indeed there exist richer alternatives. For the multivariate case, one possibility is to use generalized additive models (Wood 2017), where the relationship between covariates and the expected value of a response variable can be described through smooth functions. This also holds for functional regression models, for which several extensions are possible, such as functional generalized additive models (Scheipl et al. 2015, 2016), where the effect of a functional covariate over the response variable is a generic smooth functional effect instead of a linear one.

Completion methods for functional data

Another important research question in functional data analysis that, surprisingly, raised the attention only recently is how to do statistical inference in real time for functional variables defined over temporal domains. While in Chapter 7 we have provided a method for real-time monitoring, which gives indications about anomalies up to the current instant, it may be interesting to consider the case where, conditional on the currently observed part, one wants to have a prediction about the remaining part of a functional observation and predict if this future unobserved part will be anomalous. Some recent works that describe the state-of-the-art for this future research line are shown below.

One of the first works in this context is the paper from Yao et al. (2005), which estimates functional data from sparse observations, however it assumes functional data are sparsely observed, while in our application navigation data are available with high frequency. For dense data, Kraus (2015) proposed a method to complete the missing part of partially observed functional data from the available observation in the more general case where the functional domain is not necessarily time and the missing part is any interval contained in the functional domain. However the proposed method does not guarantee continuity at the boundary point between observed and unobserved portions, which is an important desired aspect. Goldberg et al. (2014) proposed a method to predict the continuation of a function, which is suitable to the application proposed in this thesis. It guarantees continuation of curve forecasting, however it is strictly related on splines. In a very recent work, Kneip & Liebl (2020) propose a reconstruction operator that also aims to recover the missing parts of a function given the observed parts. This new operator belongs to a new, very large class of functional operators which includes the classical regression operators as a special case and guarantees continuity at the boundary point.

Bibliography

- Abramowicz, K., Häger, C. K., Pini, A., Schelin, L., Sjöstedt de Luna, S. & Vantini, S. (2018), ‘Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament’, *Scandinavian Journal of Statistics* **45**(4), 1036–1061.
- 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.
- Andersen, C. M. & Bro, R. (2010), ‘Variable selection in regression—a tutorial’, *Journal of Chemometrics* **24**(11-12), 728–737.
- Bazari, Z. (2007), ‘Ship energy performance benchmarking/rating: methodology and application’, *Journal of Marine Engineering & Technology* **6**(1), 11–18.
- Bertram, V. (2011), *Practical ship hydrodynamics*, Elsevier.
- Beşikçi, E. B., Arslan, O., Turan, O. & Ölçer, A. (2016), ‘An artificial neural network based decision support system for energy efficient ship operations’, *Computers & Operations Research* **66**, 393–401.
- Bialystocki, N. & Konovessis, D. (2016), ‘On the estimation of ship’s fuel consumption and speed curve: a statistical approach’, *Journal of Ocean Engineering and Science* **1**(2), 157–166.
- Bocchetti, D., Lepore, A., Palumbo, B. & Vitiello, L. (2013), A statistical control of the ship fuel consumption, in ‘RINA, Royal Institution of Naval Architects - Design and Operation of Passenger Ships’.
- Bocchetti, D., Lepore, A., Palumbo, B. & Vitiello, L. (2015), ‘A statistical approach to ship fuel consumption monitoring’, *Journal of Ship Research* **59**(3), 162–171.
- Box, G. & Narasimhan, S. (2010), ‘Rethinking statistics for quality control’, *Quality Engineering* **22**(2), 60–72.
- Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and regression trees*, CRC Press.
- Bro, R. (1996), ‘Multiway calibration. multilinear pls’, *Journal of chemometrics* **10**(1), 47–61.
- Buhaug, Ø., Corbett, J. J., Endresen, Ø., Eyring, V., Faber, J., Hanayama, S., Lee, D. S., Lee, D., Lindstad, H., Markowska, A. et al. (2009), ‘Second IMO GHG Study 2009’, *International Maritime Organization (IMO) London, UK* .

- Burnham, A. J., MacGregor, J. F. & Viveros, R. (1999), 'Latent variable multivariate regression modeling', *Chemometrics and Intelligent Laboratory Systems* **48**(2), 167–180.
- Canu, S., Grandvalet, Y., Guigue, V. & Rakotomamonjy, A. (2005), 'SVM and Kernel Methods Matlab Toolbox', Perception Systemes et Information, INSA de Rouen, Rouen, France.
- Cao, D.-S., Xu, Q.-S., Liang, Y.-Z., Zhang, L.-X. & Li, H.-D. (2010), 'The boosting: A new idea of building models', *Chemometrics and Intelligent Laboratory Systems* **100**(1), 1–11.
- Chen, H. (1989), A new approach to ship speed performance monitoring and prediction, in 'Ship Technology and Research (STAR) symposium'.
- Chen, L.-H. & Jiang, C.-R. (2017), 'Multi-dimensional functional principal component analysis', *Statistics and Computing* **27**(5), 1181–1192.
- Chiou, J.-M., Chen, Y.-T. & Yang, Y.-F. (2014), 'Multivariate functional principal component analysis: A normalization approach', *Statistica Sinica* **24**, 1571–1596.
- Chiou, J.-M., Yang, Y.-F. & Chen, Y.-T. (2016), 'Multivariate functional linear regression and prediction', *Journal of Multivariate Analysis* **146**, 301–312.
- Chiou, J.-M., Zhang, Y.-C., Chen, W.-H. & Chang, C.-W. (2014), 'A functional data approach to missing value imputation and outlier detection for traffic flow data', *Transportmetrica B: Transport Dynamics* **2**(2), 106–129.
- Clarke, K. A. (2005), 'The phantom menace: Omitted variable bias in econometric research', *Conflict management and peace science* **22**(4), 341–352.
- Colosimo, B. M. & Pacella, M. (2007), 'On the use of principal component analysis to identify systematic patterns in roundness profiles', *Quality and reliability engineering international* **23**(6), 707–725.
- Colosimo, B. M. & Pacella, M. (2010), 'A comparison study of control charts for statistical monitoring of functional data', *International Journal of Production Research* **48**(6), 1575–1601.
- Corbett, J. J. & Koehler, H. W. (2003), 'Updated emissions from ocean shipping', *Journal of Geophysical Research* **108**(D20).
- De Jong, S. (1993), 'SIMPLS: an alternative approach to partial least squares regression', *Chemometrics and intelligent laboratory systems* **18**(3), 251–263.
- Dietterich, T. G. (2000), Ensemble methods in machine learning, in 'International workshop on multiple classifier systems', Springer, pp. 1–15.
- Draper, N. R. & Smith, H. (2014), *Applied regression analysis*, John Wiley & Sons.
- Duchesne, C. & MacGregor, J. F. (2004), 'Establishing multivariate specification regions for incoming materials', *Journal of quality technology* **36**(1), 78–94.
- Efron, B. & Gong, G. (1983), 'A leisurely look at the bootstrap, the jackknife, and cross-validation', *The American Statistician* **37**(1), 36–48.

- Elith, J., Leathwick, J. R. & Hastie, T. (2008), ‘A working guide to boosted regression trees’, *Journal of Animal Ecology* **77**(4), 802–813.
- Erto, P., Lepore, A., Palumbo, B. & Vitiello, L. (2015), ‘A procedure for predicting and controlling the ship fuel consumption: Its implementation and test’, *Quality and Reliability Engineering International* **31**(7), 1177–1184.
- European Commission (2012), ‘Commission Regulation (EU) No 601/2012 of the European Parliament and of the Council of 21 June 2012 on the monitoring and reporting of greenhouse gas emissions pursuant to Directive 2003/87/EC of the European Parliament and of the Council’, *Official Journal of the European Union* **L181**, 30–104.
- European Commission (2013), ‘Proposal for a regulation of the monitoring, reporting and verification of carbon dioxide emissions from maritime transport and amending Regulation (EU) No 525/2013’.
- European Commission (2015), ‘Regulation (EU) 2015/757 of the European Parliament and of the Council of 29 April 2015 on the monitoring, reporting and verification of carbon dioxide emissions from maritime transport, and amending directive 2009/16/EC’, *Official Journal of the European Union* **L123**, 55–76.
- Fagerholt, K., Laporte, G. & Norstad, I. (2010), ‘Reducing fuel emissions by optimizing speed on shipping routes’, *Journal of the Operational Research Society* **61**(3), 523–529.
- Filzmoser, P., Liebmann, B. & Varmuza, K. (2009), ‘Repeated double cross validation’, *Journal of Chemometrics: A Journal of the Chemometrics Society* **23**(4), 160–171.
- Freund, Y., Schapire, R. & Abe, N. (1999), ‘A short introduction to boosting’, *Journal-Japanese Society For Artificial Intelligence* **14**(771-780), 1612.
- Freund, Y., Schapire, R. E. et al. (1996), Experiments with a new boosting algorithm, in ‘ICML’, Vol. 96, pp. 148–156.
- Gabrielli, G. & von Karman, T. (1950), ‘What price speed?—specific power required for propulsion of vehicles’, *Mechanical Engineering* **72**(10).
- Geladi, P. & Kowalski, B. R. (1986), ‘Partial least-squares regression: a tutorial’, *Analytica chimica acta* **185**, 1–17.
- Goldberg, Y., Ritov, Y. & Mandelbaum, A. (2014), ‘Predicting the continuation of a function with applications to call center data’, *Journal of Statistical Planning and Inference* **147**, 53–65.
- Gunst, R. F. & Mason, R. L. (1977), ‘Biased estimation in regression: an evaluation using mean squared error’, *Journal of the American Statistical Association* **72**(359), 616–628.
- Happ, C. & Greven, S. (2018), ‘Multivariate functional principal component analysis for data observed on different (dimensional) domains’, *Journal of the American Statistical Association* **113**(522), 649–659.
- Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The elements of statistical learning*, Springer.

- He, Q. P. & Wang, J. (2011), ‘Statistics pattern analysis: A new process monitoring framework and its application to semiconductor batch processes’, *AIChE journal* **57**(1), 107–121.
- Helland, I. S. (1988), ‘On the structure of partial least squares regression’, *Communications in statistics-Simulation and Computation* **17**(2), 581–607.
- Hoerl, A. E. & Kennard, R. W. (1970), ‘Ridge regression: Biased estimation for nonorthogonal problems’, *Technometrics* **12**(1), 55–67.
- IMO (2000), ‘IMO Standard Marine Communication Phrases (SMCP)’.
- IMO (2009), ‘Guidelines for voluntary use of the ship Energy Efficiency Operational Indicator (EEOI), MEPC.1/Circ.684’.
- IMO (2012a), ‘Air Pollution and Greenhouse Gas (GHG) Emissions from International Shipping, MARPOL Annex 6. London, U.K.’.
- IMO (2012b), ‘Guidelines for the development of a Ship Energy Efficiency Management Plan (SEEMP), MEPC.213(63) Annex 9. London, U.K.’.
- IMO (2012c), ‘Guidelines on survey and certification of the Energy Efficiency Design Index (EEDI), MEPC.214(63) Annex 10. London, U.K.’.
- IMO (2012d), ‘Guidelines on the method of calculation of the attained Energy Efficiency Design Index (EEDI) for new ships, MEPC.212 Annex 8. London, U.K.’.
- IMO (2014), ‘2014 Guidelines on survey and certification of the Energy Efficiency Design Index (EEDI), London, U.K.’.
- ISO (2015), ‘ISO 15016:2015 Ships and marine technology — Guidelines for the assessment of speed and power performance by analysis of speed trial data’.
- ITTC (2008), *Dictionary of Ship Hydrodynamics*.
- ITTC (2017), ‘Recommended procedures and guidelines: Preparation, conduct and analysis of speed/power trials.’.
- Jackson, J. E. (2005), *A user’s guide to principal components*, John Wiley & Sons.
- Jackson, J. E. & Mudholkar, G. S. (1979), ‘Control procedures for residuals associated with principal component analysis’, *Technometrics* **21**(3), 341–349.
- Jolliffe, I. T. (2002), *Principal component analysis*, Springer.
- Jørgensen, K., Mevik, B.-H. & Næs, T. (2007), ‘Combining designed experiments with several blocks of spectroscopic data’, *Chemometrics and intelligent laboratory systems* **88**(2), 154–166.
- Jørgensen, K., Segtnan, V., Thyholt, K. & Næs, T. (2004), ‘A comparison of methods for analysing regression models with both spectral and designed variables’, *Journal of Chemometrics: A Journal of the Chemometrics Society* **18**(10), 451–464.
- Ki-Moon, B. (2008), Kyoto protocol reference manual, in ‘United Nations Framework Convention on Climate Change’.

- Kim, J.-G., Kim, H.-J., Jun, H. B. & Kim, C.-M. (2016), ‘Optimizing ship speed to minimize total fuel consumption with multiple time windows’, *Mathematical Problems in Engineering* **2016**.
- Kim, J.-G., Kim, H.-J. & Lee, P. T.-W. (2014), ‘Optimizing ship speed to minimize fuel consumption’, *Transportation Letters* **6**(3), 109–117.
- Kneip, A. & Liebl, D. (2020), ‘On the optimal reconstruction of partially observed functional data’, *to appear in Annals of Statistics* .
- Kourti, T. & MacGregor, J. F. (1995), ‘Process analysis, monitoring and diagnosis, using multivariate projection methods’, *Chemometrics and intelligent laboratory systems* **28**(1), 3–21.
- Kourti, T. & MacGregor, J. F. (1996), ‘Multivariate spc methods for process and product monitoring’, *Journal of Quality Technology* **28**(4), 409–428.
- Krämer, N. (2007), ‘An overview on the shrinkage properties of partial least squares regression’, *Computational Statistics* **22**(2), 249–273.
- Kraus, D. (2015), ‘Components and completion of partially observed functional data’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**(4), 777–801.
- Lepore, A., Palumbo, B. & Capezza, C. (2018), ‘Analysis of profiles for monitoring of modern ship performance via partial least squares methods’, *Quality and Reliability Engineering International* **34**(7), 1424–1436.
- Lepore, A., Reis, M. S., Palumbo, B., Rendall, R. & Capezza, C. (2017), ‘A comparison of advanced regression techniques for predicting ship CO₂ emissions’, *Quality and Reliability Engineering International* **33**(6), 1281–1292.
- Lewis, E. V. (1988), *Principles of Naval Architecture Second Revision – Volume II: Resistance, Propulsion and Vibration*, The Society of Naval Architects and Marine Engineers.
- Løvøll, G. & Kadal, J. C. (2014), ‘Big data-the new data reality and industry impact’, *DNV GL* .
- Lu, R., Turan, O. & Boulougouris, E. (2013), Voyage optimization, prediction of ship specific fuel consumption for energy efficient shipping, *in* ‘Low Carbon Shipping Conference, London’, pp. 1–11.
- Lu, R., Turan, O., Boulougouris, E., Banks, C. & Incecik, A. (2015), ‘A semi-empirical ship operational performance prediction model for voyage optimization towards energy efficient shipping’, *Ocean Engineering* **110**, 18–28.
- MacGregor, J. F. & Kourti, T. (1995), ‘Statistical process control of multivariate processes’, *Control Engineering Practice* **3**(3), 403–414.
- MAN Diesel & Turbo (2011), ‘Basic principles of ship propulsion’.
- Mandel, B. (1969), ‘The regression control chart’, *Journal of Quality Technology* **1**(1), 1–9.
- Meng, Q., Du, Y. & Wang, Y. (2016), ‘Shipping log data based container ship fuel efficiency modeling’, *Transportation Research Part B: Methodological* **83**, 207–229.

- Milligan, P., Martinsky, F., Good, K. & Nelson, B. (2006), 'EPA Evaluation report: EPA Can Improve Emission Factors Development and Management'.
- Molinero, C. M. & Mitsis, S. N. (1984), 'Budgeting fuel consumption in a cruise liner', *European journal of operational research* **18**(2), 172–183.
- Montgomery, D. C. (2007), *Introduction to statistical quality control*, John Wiley & Sons.
- Montgomery, D. C., Peck, E. A. & Vining, G. G. (2012), *Introduction to linear regression analysis*, John Wiley & Sons.
- Montgomery, D. C. & Runger, G. C. (2010), *Applied statistics and probability for engineers*, John Wiley & Sons.
- Murphy, A., Weston, S. & Young, R. (2012), 'Reducing fuel usage and CO₂ emissions from tug boat fleets: Sea trials and theoretical modelling', *International Journal of Maritime Engineering* **154**, A31–A34.
- Nomikos, P. & MacGregor, J. F. (1995a), 'Multi-way partial least squares in monitoring batch processes', *Chemometrics and intelligent laboratory systems* **30**(1), 97–108.
- Nomikos, P. & MacGregor, J. F. (1995b), 'Multivariate spc charts for monitoring batch processes', *Technometrics* **37**(1), 41–59.
- Noorossana, R., Saghaei, A. & Amiri, A. (2012), *Statistical analysis of profile monitoring*, John Wiley & Sons.
- Pal, S. K. & Mitra, P. (2004), *Pattern recognition algorithms for data mining*, Chapman and Hall/CRC.
- Perera, L. P. & Mo, B. (2016), 'Emission control based energy efficiency measures in ship operations', *Applied Ocean Research* **60**, 29–46.
- Petersen, J. P., Jacobsen, D. J. & Winther, O. (2012), 'Statistical modelling for ship propulsion efficiency', *Journal of marine science and technology* **17**(1), 30–39.
- Phatak, A. & De Jong, S. (1997), 'The geometry of partial least squares', *Journal of Chemometrics* **11**(4), 311–338.
- Qi, X. & Luo, R. (2019), 'Nonlinear function on function additive model with multiple predictor curves', *Statistica Sinica* **29**, 719–739.
- Ramsay, J. O., Hooker, G. & Graves, S. (2009), *Functional data analysis with R and MATLAB*, Springer Science & Business Media.
- Ramsay, J. O. & Silverman, B. W. (2002), *Applied functional data analysis: methods and case studies*, Vol. 77, Springer New York.
- Ramsay, J. O. & Silverman, B. W. (2005), *Functional data analysis*, Wiley Online Library.
- Ramsay, J., Wickham, H., Graves, S. & Hooker, G. (2018), *fda: Functional Data Analysis*. R package version 2.4.8.
- Rasmussen, M. A. & Bro, R. (2012), 'A tutorial on the lasso approach to sparse modeling', *Chemometrics and Intelligent Laboratory Systems* **119**, 21–31.

- Reis, M. S., Rendall, R., Palumbo, B., Lepore, A. & Capezza, C. (2020), ‘Predicting ships’ CO₂ emissions using feature-oriented methods’, *Applied Stochastic Models in Business and Industry* **36**(1), 110–123.
- Reiss, P. T., Goldsmith, J., Shang, H. L. & Ogden, R. T. (2017), ‘Methods for scalar-on-function regression’, *International Statistical Review* **85**(2), 228–249.
- Rendall, R., Lu, B., Castillo, I., Chin, S.-T., Chiang, L. H. & Reis, M. S. (2017), ‘A unifying and integrated framework for feature oriented analysis of batch processes’, *Industrial & Engineering Chemistry Research* **56**(30), 8590–8605.
- Rendall, R. & Reis, M. S. (2018), ‘Which regression method to use? Making informed decisions in “data-rich/knowledge poor” scenarios—The Predictive Analytics Comparison framework (PAC)’, *Chemometrics and Intelligent Laboratory Systems* **181**, 52–63.
- Ronen, D. (1982), ‘The effect of oil price on the optimal speed of ships’, *Journal of the Operational Research Society* **33**(11), 1035–1040.
- Scheipl, F., Gertheiss, J., Greven, S. et al. (2016), ‘Generalized functional additive mixed models’, *Electronic Journal of Statistics* **10**(1), 1455–1492.
- Scheipl, F., Staicu, A.-M. & Greven, S. (2015), ‘Functional additive mixed models’, *Journal of Computational and Graphical Statistics* **24**(2), 477–501.
- Schrady, D. A., Smyth, G. K. & Vassian, R. B. (1996), Predicting Ship Fuel Consumption: Update, Technical report, Naval Postgraduate School, Monterey, California, Department of Operations Research.
- Searle, S. R. & Khuri, A. I. (2017), *Matrix algebra useful for statistics*, John Wiley & Sons.
- Seasholtz, M. B. & Kowalski, B. (1993), ‘The parsimony principle applied to multivariate calibration’, *Analytica Chimica Acta* **277**(2), 165–177.
- Šidák, Z. (1967), ‘Rectangular confidence regions for the means of multivariate normal distributions’, *Journal of the American Statistical Association* **62**(318), 626–633.
- Smilde, A. K. (1997), ‘Comments on multilinear pls’, *Journal of Chemometrics* **11**(5), 367–377.
- Smith, T., Jalkanen, J., Anderson, B., Corbett, J., Faber, J., Hanayama, S., O’Keeffe, E., Parker, S., Johansson, L., Aldous, L., Raucci, C., Traut, M., Ettinger, S., Nelissen, D., Lee, D., Ng, S., Agrawal, A., Winebrake, J., Hoen, M., Chesworth, S. & Pandey, A. (2015), ‘Third imo ghg study 2014’, *International Maritime Organization (IMO)*, London, UK .
- Smola, A. J. & Schölkopf, B. (2004), ‘A tutorial on support vector regression’, *Statistics and computing* **14**(3), 199–222.
- St Amand, D. G. (2012), ‘Optimal economic speed and the impact on marine ghg emissions-saving money and the world at the same time’, *SNAME Transactions* .
- Strobl, C., Malley, J. & Tutz, G. (2009), ‘An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests.’, *Psychological methods* **14**(4), 323.

- Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.
- Tracy, N. D., Young, J. C. & Mason, R. L. (1992), 'Multivariate control charts for individual observations', *Journal of quality technology* **24**(2), 88–95.
- Trodden, D., Murphy, A., Pazouki, K. & Sargeant, J. (2015), 'Fuel usage data analysis for efficient shipping operations', *Ocean Engineering* **110**, 75–84.
- Van Manen, J., Van Ossanen, P. & Lewis, E. (1988), 'Principles of naval architecture, second revision, volume II: resistance, propulsion, and vibration', *Society of Naval Architects and Marine Engineers* .
- Veness, C. (2007), 'Calculate distance, bearing and more between two latitude/longitude points'.
- Vining, G. (2009), 'Technical Advice: Phase I and phase II control charts', *Quality Engineering* **21**(4), 478–479.
- Wang, J. & He, Q. P. (2010), 'Multivariate statistical process monitoring based on statistics pattern analysis', *Industrial & Engineering Chemistry Research* **49**(17), 7858–7869.
- Wang, J.-L., Chiou, J.-M. & Mueller, H.-G. (2016), 'Functional data analysis', *Annual Review of Statistics and Its Application* **3**, 257–295.
- Wold, S., Esbensen, K. & Geladi, P. (1987), 'Principal component analysis', *Chemometrics and intelligent laboratory systems* **2**(1-3), 37–52.
- Wold, S., Ruhe, A., Wold, H. & Dunn, III, W. (1984), 'The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses', *SIAM Journal on Scientific and Statistical Computing* **5**(3), 735–743.
- Wold, S., Sjöström, M. & Eriksson, L. (2001), 'PLS-regression: a basic tool of chemometrics', *Chemometrics and intelligent laboratory systems* **58**(2), 109–130.
- Wood, S. N. (2017), *Generalized additive models: an introduction with R*, CRC press.
- Woodall, W. H. (2017), 'Bridging the gap between theory and practice in basic statistical process monitoring', *Quality Engineering* **29**(1), 2–15.
- Woodall, W. H., Spitzner, D. J., Montgomery, D. C. & Gupta, S. (2004), 'Using control charts to monitor process and product quality profiles', *Journal of Quality Technology* **36**(3), 309.
- Yao, F., Müller, H.-G. & Wang, J.-L. (2005), 'Functional data analysis for sparse longitudinal data', *Journal of the American Statistical Association* **100**(470), 577–590.
- Yuan, J., Ng, S. H. & Sou, W. S. (2016), 'Uncertainty quantification of CO₂ emission reduction for maritime shipping', *Energy Policy* **88**, 113–130.
- Zaman, I., Pazouki, K., Norman, R., S, Y. & Coleman, S. (2017), 'Development of automatic mode detection system by implementing the statistical analysis of ship data to monitor the performance', *International Journal of Maritime Engineering* **159**, A225–A235.

Zhang, Z. & Folmer, H. (1995), ‘The choice of policy instruments for the control of carbon dioxide emissions’, *Intereconomics* **30**(3), 133–142.

Zou, H. & Hastie, T. (2005), ‘Regularization and variable selection via the elastic net’, *Journal of the royal statistical society: series B (statistical methodology)* **67**(2), 301–320.