MOX-Report No. 12/2019

Control charts for monitoring ship operating conditions and CO2 emissions based on scalar-on-function regression

Capezza, C.; Lepore, A.; Menafoglio, A.; Palumbo, B.; Vantini, S.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

Control charts for monitoring ship operating conditions and CO_2 emissions based on scalar-on-function regression

Christian Capezza[1], Antonio Lepore[1], Alessandra Menafoglio[2], Biagio Palumbo[1], and Simone Vantini[2]

[1] *Department of Industrial Engineering, University of Naples Federico II, Piazzale Tecchio 80, 80125, Naples, Italy*

[2] *MOX - Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milan, Italy*

Abstract

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 procedure is effectively applied to a real-case study on monitoring of operating conditions (i.e., the multivariate functional data) and total CO_2 emissions (i.e., the scalar quality characteristic) at each voyage of a cruise ship.

Key Words: Functional Data Analysis; Multivariate Functional Principal Component Analysis; Profile Monitoring; Statistical Process Monitoring.

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 problem classically addressed by Mandel (1969), where a scalar quality characteristic is affected by a single scalar variable, is nowadays exacerbated by the capacity of storing massive amounts of data from multiple sources. This increases the complexity and the dimension of the information and naturally calls for an extension of the classical statistical methods toward new mathematical settings. In this perspective, in SPC of batch processes, Nomikos and MacGregor (1995a,b) and Kourti and MacGregor (1996) have introduced methods that address the problem of dimensionality reduction in order to monitor one or more quality characteristics on the basis of several covariates observed over a discrete time domain. In those works, the dimensionality reduction is mainly achieved by projection methods of a multivariate domain, such as principal component analysis (PCA), and indeed allows to jointly monitor also the covariates themselves. These multivariate methods have the potential to cope also with applications where the quality characteristic is described by a function (usually referred to as profile (Woodall et al., 2004)) and gave raise, in more recent years, to the new field in SPC known as *profile monitoring* (Noorossana et al., 2012; Colosimo and Pacella, 2007, 2010). Stimulated by many technological contexts with the increasing need of handling data that can be considered as varying over a continuous domain (Happ and Greven, 2016; Chen and Jiang, 2017), this new standpoint naturally unleashes cross-fertilization of SPC with functional data analysis (FDA) (Ramsay and 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 and Greven, 2016) and quality characteristics related to them. In this work, the quality characteristic is supposed to be a scalar and the 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 and 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. The mathematical and technological interpretation of these control charts can indeed benefit from the optimal choice of the functional principal components. As in the multivariate case, contribution plots shall be defined accordingly to help diagnosing which variables, among the functional covariates, are responsible for an out of control 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 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 signal an *out-of-control* (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, a discussion is provided on the natural extension of the proposed method to real-time monitoring in those cases where observations of the functional covariates are registered and made available also up to intermediate time domain points.

The proposed monitoring strategy is motivated and illustrated by means of a real-case study from the maritime field in monitoring CO_2 emissions from a roll-on/roll-off passenger (Ro-Pax) cruise ship navigation data, courtesy of the owner Grimaldi Group.

The paper is organized as follows: Section 2 describes the motivating example related to the problem of CO_2 emission in the maritime field; Section 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 4 presents the real-case study; Section 5 provides a discussion for a possible use of the proposed strategy in real-time monitoring; Section 6 draws conclusions.

2 A Motivating Example

In the last years, the problem of monitoring CO_2 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, 2012a,b,c, 2014; Smith et al., 2015) that require monitoring and verification of CO_2 emissions.

To respond to this compelling worldwide 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 (see e.g., Bialystocki and Konovessis (2016)), which are, in general, also function of time.

The maritime field constitutes a new challenging area for FDA and related methods. The problems addressed are, on one side, to build models that allow prediction of ship CO_2 emissions based on observational data that describe the ship operating conditions, and, on the other, to monitor operating conditions for detecting anomalies and diagnosing faults.

Naval 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 and 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 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 naval context only in the very last years (see e.g., Lepore et al. (2017) for a thorough comparison). However, most of these approaches involve only summary statistics of the complete sensor signal over each voyage and do not exploit the potential to monitor the entire voyage profile. Sensor signal data acquired on board of modern ships have in fact the potential to be used to build functional variables on a given route over different voyages and to timely support managerial decision-making.

3 Methodology

The scalar-on-function regression model is illustrated in Section 3.1, while the Phase I model estimation is described in Section 3.2. The Phase II monitoring procedure and fault diagnosis are introduced in Section 3.3 and 3.4, respectively.

3.1 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 $f = (f_1, \ldots, f_P)$ and $g = (g_1, \ldots, g_P)$, 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 $\|f\|_{\mathbb{H}} = \langle f, f \rangle_{\mathbb{H}}$.

Let us denote with $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \ldots, \tilde{X}_P)$ a random element that takes values

in $\mathbb{H}$, i.e., $\tilde{X}_1, \ldots, \tilde{X}_P$ are random elements that take values in $L^2(\mathcal{T})$, which are hereinafter referred to as *functional covariates*. Moreover, let $\tilde{\boldsymbol{X}}$ have mean function $\boldsymbol{\mu}^X = (\mu_1^X, \ldots, \mu_P^X)$, with $\mu_p^X(t) = \mathrm{E}(\tilde{X}_p(t))$ for every $t \in \mathcal{T}$, variance function $\boldsymbol{v}^X = \left(v_1^X, \ldots, v_P^X\right)$, where $v_p^X(t) = \mathrm{Var}(\tilde{X}_p(t))$, and correlation function $\boldsymbol{C} = \{C_{p_1p_2}\}_{p_1,p_2=1,\ldots,P}$, with $C_{p_1p_2}(t_1,t_2) = \mathrm{Corr}\left(\tilde{X}_{p_1}(t_1), \tilde{X}_{p_2}(t_2)\right) = \mathrm{Cov}(\tilde{X}_{p_1}(t_1), \tilde{X}_{p_2}(t_2)) v_{p_1}(t_1)^{-1/2} v_{p_2}(t_2)^{-1/2}$. To deal with infinite dimensionality of the data, $\tilde{\boldsymbol{X}}$ is decomposed through multivariate functional principal component analysis (MFPCA). However, as is known, this method is not scale-invariant, thus $\tilde{\boldsymbol{X}}$ are suitably scaled through the normalization approach proposed by Chiou et al. (2014a). The normalized functional covariates are denoted by $\boldsymbol{X} = (X_1, \ldots, X_P)$, and $X_p(t)$, $p = 1, \ldots, P$, are obtained as $v_p(t)^{-1/2}(\tilde{X}_p(t) - \mu_p^X(t))$. Trivially note that $\boldsymbol{X}$ has zero mean and covariance function that coincides with $\boldsymbol{C}$.

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

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

where $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P) \in \mathbb{H}$ are the coefficient to be estimated, and $\varepsilon_1, \ldots, \varepsilon_n$ 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. $\mathrm{E}(\varepsilon_i X_p(t)) = 0$ for each $i = 1 \ldots, n$, $p = 1, \ldots, P$, and $t \in \mathcal{T}$.

3.2 Phase I Model Estimation

Instead of using a random sample $\{(\boldsymbol{X}_i, y_i)\}_{i=1,\ldots,n}$, 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 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)$ in Eq.(1) can be estimated by solving the following least-squares problem

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

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 $\boldsymbol{X}$

$$\boldsymbol{X}(t) = \sum_{m=1}^{\infty} \xi_m \boldsymbol{\psi}_m(t), \tag{3}$$

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

$$\langle \boldsymbol{\psi}_{m_1}, \boldsymbol{\psi}_{m_2} \rangle_{\mathbb{H}} = \sum_{p=1}^{P} \langle \psi_{m_1 p}, \psi_{m_2 p} \rangle = \begin{cases} 1 & \text{if } m_1 = m_2 \\ 0 & \text{if } m_1 \neq m_2 \end{cases}. \tag{4}$$

The latter represent the eigenfunctions of the integral operator Γ with the correlation function $\boldsymbol{C}$ as kernel (i.e., the solutions of $\Gamma\boldsymbol{\psi}_m = \lambda_m \boldsymbol{\psi}_m$), with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. In Eq.(3), the multivariate functional principal component scores, or scores, $\xi_m = \langle \boldsymbol{X}, \boldsymbol{\psi}_m \rangle_{\mathbb{H}} = \sum_{p=1}^{P} \langle X_p, \psi_{mp} \rangle$ are random coefficients with $\mathrm{E}(\xi_m) = 0$, $\mathrm{E}(\xi_m^2) = \lambda_m$ and $\mathrm{E}(\xi_{m_1}\xi_{m_2}) = 0$ when $m_1 \neq m_2$.

The coefficient $\boldsymbol{\beta}$ in the model in Eq.(1) can be expressed by using the same eigenbasis of $\mathbb{H}$

$$\boldsymbol{\beta}(t) = \sum_{m=1}^{\infty} b_m \boldsymbol{\psi}_m(t). \tag{5}$$

In this way, by substituting Eq.(3) and (5) into Eq.(1) we get

$$y_i = \beta_0 + \sum_{m=1}^{\infty} \langle \xi_{im}\boldsymbol{\psi}_m, b_m\boldsymbol{\psi}_m \rangle_{\mathbb{H}} + \varepsilon_i = \beta_0 + \sum_{m=1}^{\infty} \xi_{im} b_m + \varepsilon_i, \quad i = 1, \ldots, n, \tag{6}$$

where $\xi_{im} = \langle \boldsymbol{X}_i, \boldsymbol{\psi}_m \rangle_{\mathbb{H}}$ are the scores of the i-th observation $\boldsymbol{X}_i$. 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 Eq.(2) because of the finite number of available data. Therefore, we consider an M-dimensional approximation of $\boldsymbol{X}(t)$ in Eq.(3)

$$\boldsymbol{X}_M(t) = \sum_{m\in\mathcal{M}} \xi_m \boldsymbol{\psi}_m(t), \tag{7}$$

where $\mathcal{M} = \{m_1, \ldots, m_M\} \subset \mathbb{N}$ 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., $\mathcal{M} = \{1, \ldots, M\}$ (Chiou et al., 2014b). 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

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{[i]})^2, \tag{8}$$

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 4. By considering the approximation in Eq.(7) and taking into account Eq.(6), we can write the model in Eq.(1) as

$$y_i = \beta_0 + \sum_{m \in \mathcal{M}} \xi_{im} b_m + \varepsilon_i^M, \quad i = 1, \ldots, n, \tag{9}$$

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 $\boldsymbol{\beta}$ in Eq.(5) based on a set of n observations $(\tilde{\boldsymbol{X}}_i, y_i)$, we first estimate the mean function as $\hat{\boldsymbol{\mu}}^X(t) = \sum_{i=1}^n \tilde{\boldsymbol{X}}_i(t)/n$ and the variance function as $\hat{\boldsymbol{v}}^X(t) = \sum_{i=1}^n (\tilde{\boldsymbol{X}}_i(t) - \hat{\boldsymbol{\mu}}^X(t))^2/(n-1)$, then standardize $\tilde{\boldsymbol{X}}_i$ and obtain $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iP})$, 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 and Greven, 2016), 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 $\hat{\xi}_{im} = \langle \boldsymbol{X}_i, \hat{\boldsymbol{\psi}}_m \rangle_{\mathbb{H}}$. 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

$$\hat{b}_m = \frac{\sum_{i=1}^n y_i \hat{\xi}_{im}}{\sum_{i=1}^n \hat{\xi}_{im}^2}. \tag{10}$$

Accordigly, the estimate of $\boldsymbol{\beta}$ can be obtained as

$$\hat{\boldsymbol{\beta}}(t) = \sum_{m \in \mathcal{M}} \hat{b}_m \hat{\boldsymbol{\psi}}_m(t), \tag{11}$$

and the prediction of y_i results to be

$$\hat{y}_i = \hat{\beta}_0 + \sum_{m \in \mathcal{M}} \hat{\xi}_{im} \hat{b}_m, \tag{12}$$

Finally, we can also get an estimate of the variance σ^2 of the error in the model in Eq.(1)

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n - M - 1}, \tag{13}$$

where

$$e_i = y_i - \hat{y}_i \tag{14}$$

are the residuals.

3.3 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 $\{(\boldsymbol{X}_i, y_i)\}_{i=1,\dots,n}$, the scalar-on-function regression model has been estimated together with all the parameters described in Section 3.2. Moreover, suppose that a new observation $(\boldsymbol{X}^{new}, y^{new})$ is available from $(\boldsymbol{X}, y)$, where y^{new} is the new observation of the response variable and $\boldsymbol{X}^{new} = (X_1^{new}, \dots, X_P^{new})$ is the corresponding new standardized observation of $\tilde{\boldsymbol{X}}$. The new scores $\{\hat{\xi}_m^{new}\}_{m\in\mathcal{M}}$ can be calculated as

$$\hat{\xi}_m^{new} = \langle \boldsymbol{X}^{new}, \hat{\boldsymbol{\psi}}_m \rangle_{\mathbb{H}}, \quad m \in \mathcal{M}. \tag{15}$$

Note that the inner product in Eq. (15) assumes $\boldsymbol{X}^{new}$ to be completely observed over the domain $\mathcal{T}$ in order to calculate the statistics used for monitoring the operating conditions defined as in Eq. (16), (18), and (19). 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

$$T^2 = \sum_{m\in\mathcal{M}} \frac{(\hat{\xi}_m^{new})^2}{\hat{\lambda}_m}. \tag{16}$$

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 $\boldsymbol{X}^{new}$ with

$\boldsymbol{X}_M^{new} = (X_{M1}^{new}, \ldots, X_{MP}^{new})$

$$\boldsymbol{X}_M^{new}(t) = \sum_{m \in \mathcal{M}} \hat{\xi}_m^{new} \hat{\boldsymbol{\psi}}_m(t). \tag{17}$$

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 $\boldsymbol{X}^{new}(t) - \boldsymbol{X}_M^{new}(t) = \sum_{m \in \{1,\ldots,n-1\} \setminus \mathcal{M}} \hat{\xi}_m^{new} \hat{\boldsymbol{\psi}}_m(t)$, we can write

$$\begin{aligned} SPE = \|\boldsymbol{X}^{new} - \boldsymbol{X}_M^{new}\|_{\mathbb{H}}^2 = \\ = \left\langle \sum_{m_1 \in \{1,\ldots,n-1\} \setminus \mathcal{M}} \hat{\xi}_{m_1}^{new} \hat{\boldsymbol{\psi}}_{m_1}, \sum_{m_2 \in \{1,\ldots,n-1\} \setminus \mathcal{M}} \hat{\xi}_{m_2}^{new} \hat{\boldsymbol{\psi}}_{m_2} \right\rangle_{\mathbb{H}} = \\ = \sum_{m_1 \in \{1,\ldots,n-1\} \setminus \mathcal{M}} \sum_{m_2 \in \{1,\ldots,n-1\} \setminus \mathcal{M}} \hat{\xi}_{m_1}^{new} \hat{\xi}_{m_2}^{new} \langle \hat{\boldsymbol{\psi}}_{m_1}, \hat{\boldsymbol{\psi}}_{m_2} \rangle_{\mathbb{H}} = \\ = \sum_{m \in \{1,\ldots,n-1\} \setminus \mathcal{M}} (\hat{\xi}_m^{new})^2. \end{aligned} \tag{18}$$

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.

Note that if the principal components retained in the model are chosen (as suggested in Section 3.2) as those having the most influence on the response, the T^2 control chart can be technologically interpreted as controlling the variability of the covariates that impacts on the response. The SPE control chart monitors the remaining (outside the model) variability of the functional covariates that does not affect the scalar response. This perspective shows a more straightforward interpretation of the OCs issued by the two control charts.

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

$$y^{new} - \hat{y}^{new} = y^{new} - \hat{\beta}_0 - \sum_{m \in \mathcal{M}} \hat{\xi}_m^{new} \hat{b}_m. \tag{19}$$

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

$$L_{\alpha_y} = t_{n-M-1,1-\alpha_y/2}\left[\hat{\sigma}^2\left(1+\frac{1}{n-1}\sum_{m\in\mathcal{M}}\frac{(\hat{\xi}_m^{new})^2}{\hat{\lambda}_m}\right)\right]^{1/2} = \\ = t_{n-M-1,1-\alpha_y/2}\left[\hat{\sigma}^2\left(1+\frac{T^2}{n-1}\right)\right]^{1/2} \quad (20)$$

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 nonparametrically 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 Eq.(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 case 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

$$(1-\alpha_{T^2})(1-\alpha_{SPE})(1-\alpha_y) = 1-\alpha. \quad (21)$$

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

$$\alpha_{T^2} + \alpha_{SPE} + \alpha_y = \alpha. \quad (22)$$

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

$$\alpha_{T^2} = \alpha_{SPE} = \alpha_y = 1/3. \tag{23}$$

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

$$\alpha_{T^2} = \alpha_{SPE} = \alpha/4, \quad \alpha_y = \alpha/2. \tag{24}$$

3.4 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 and MacGregor (1996), the overall contribution of each functional variable to the Hotelling statistic T^2 can be defined by observing that

$$T^2 = \sum_{m\in\mathcal{M}} \frac{\hat{\xi}_m^{new}}{\hat{\lambda}_m}\hat{\xi}_m^{new} = \sum_{m\in\mathcal{M}} \frac{\hat{\xi}_m^{new}}{\hat{\lambda}_m}\langle \boldsymbol{X}^{new}, \hat{\boldsymbol{\psi}}_m\rangle_{\mathbb{H}} = \sum_{p=1}^{P}\sum_{m\in\mathcal{M}} \frac{\hat{\xi}_m^{new}}{\hat{\lambda}_m}\langle X_p^{new}, \hat{\psi}_{mp}\rangle. \tag{25}$$

Then, we can write

$$CONT_p^{T^2} = \sum_{m\in\mathcal{M}} \frac{\hat{\xi}_m^{new}}{\hat{\lambda}_m}\langle X_p^{new}, \hat{\psi}_{mp}\rangle, \quad p = 1, \ldots, P. \tag{26}$$

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

$$SPE = \sum_{p=1}^{P} \|X_p^{new} - X_{Mp}^{new}\|^2, \tag{27}$$

can be analogously defined as

$$CONT_p^{SPE} = \|X_p^{new} - X_{Mp}^{new}\|^2, \quad p = 1, \ldots, P. \tag{28}$$

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 the $(1 - \alpha_{SPE}/P)$ 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.

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) to be monitored and for predicting and monitoring CO_2 emissions, (i.e., the response variable). In view of the recent regulations discussed in Section 2, monitoring of CO_2 emissions is of great interest in this sector, in order to timely plan energy efficiency improvement operations and react to anomalies.

Section 4.1 describes the variables chosen as functional covariates and scalar response. Section 4.2 illustrates the preprocessing step required to obtain functional covariates from the data acquired from the multi-sensor system. Section 4.3 shows the implementation details for estimation of the scalar-on-function regression model (Phase I) and illustrates a scenario in which the control charts are applied to Phase II monitoring.

Table 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

4.1 Variable Description

Each observation refers to a specific voyage at given route and direction. The name of the ship, route, and voyage dates are omitted for confidentiality reasons. Table 1 shows the $P = 9$ variables used as functional covariates in the scalar-on-function regression model in Eq.(1). All the variables are measured during the navigation phase, which 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 response variable of the scalar-on-function regression model (Eq.(1)) is the total CO_2 emissions in navigation phase per each voyage, measured in tons. The 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 *cumulative sailing time* variable, measured in hours (h) is the cumulative voyage navigation time. The *speed over ground* (SOG) variable, measured in knots (kn), is the ratio between the sailed distance over ground, i.e., the distance travelled by the vessel during navigation phase, and the cumulative sailing time. The former is measured in nautic miles (NM) and calculated from latitude and longitude data acquired by the by the GPS sensor using 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* variable is useful for discovering anomalies or malfunctioning in the main engines, when one of the engines is

turned off. The *Distance from the nominal route* variable, measured in Nautic Miles (NM), is calculated as the distance, at each domain 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 the wind direction relative to the ship Ψ, measured in radiants, 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).

4.2 Preprocessing and Registration

In the proposed case study, functional data have been 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 $\tilde{\boldsymbol{X}}_i = (\tilde{X}_{i1}, \ldots, \tilde{X}_{iP})$ of the functional covariates $\tilde{\boldsymbol{X}}$ at each voyage $i = 1, \ldots, n$. For each $i = 1, \ldots, n$ and $p = 1, \ldots, P$, $\tilde{X}_{ip}$ can be obtained from the discrete data x_{ipn}, $n = 1, \ldots, N_i$, using a cubic B-spline basis with equally spaced knots

$$\tilde{X}_{ip}(t) = \sum_{q=1}^{Q} c_{iqp} \phi_q(t), \quad i = 1, \ldots, n, \quad p = 1, \ldots, P, \quad t \in \mathcal{T}, \tag{29}$$

where $\phi_1, \ldots, \phi_Q$ are the B-spline basis functions and c_{iqp} are the basis coefficients. Functional data have been obtained by smoothing data with a roughness penalty on the integrated squared second derivative. The R package `fda` (Ramsay et al., 2015) has been used for the purpose. In particular, after analysing the generalized cross-validation values and considering the number of observations along the domain, 100 bases with equally spaced knots and a roughness penalty on the integrated squared second derivative equal to 10^{-10} have been chosen.

Even if time is naturally prone to be chosen as functional domain, total traveling time could vary from voyage to voyage. Thus, a more sensitive choice is to use the fraction of distance travelled over the voyage as the

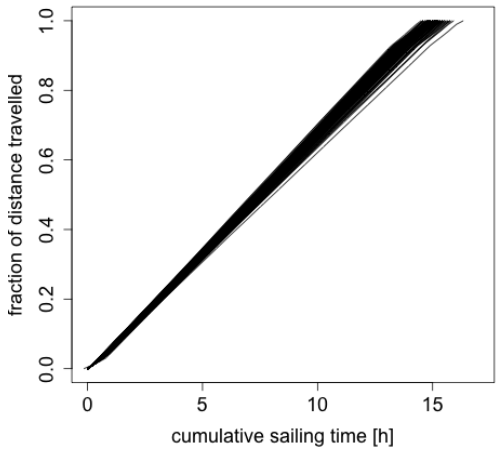


Figure 1. Warping functions that map at each voyage the cumulative sailing time to the common domain $\mathcal{T} = [0, 1]$, which is the fraction of distance travelled over the voyage.

common domain $\mathcal{T} = (0, 1)$ of the functional data (e.g., Abramowicz et al. (2018)). Switching from time to the fraction of distance travelled over the voyage can be seen as a landmark registration (Ramsay et al., 2009) of the functional data set from the function specific temporal domain (a_i, b_i) to the common domain $(0, 1)$ with the group of affine transformations with positive slope as the group of the warping functions and the voyage starting and ending points as landmarks.

4.3 Phase I Model Estimation and Phase II Monitoring

The reference data set has $n = 140$ observations and is used to estimate the control limits for the Phase II monitoring of 30 consecutive voyages. The estimation of multivariate functional principal components and corresponding scores have been obtained through the R package `MFPCA` (Happ, 2018). As explained in Section 3.2, the choice of $\mathcal{M}$, i.e., the set of components to retain in the model, has been carried out by considering both the variability of the covariates explained by the principal components, reported in Figure 2, and the PRESS statistic calculated by Eq.(8), reported in Figure 3, as a function of the first m retained principal components in Eq.(9). As previously stated in Section 3.2, we get $\mathcal{M} = \{1, 2, 5, 6\}$ as the set of principal components

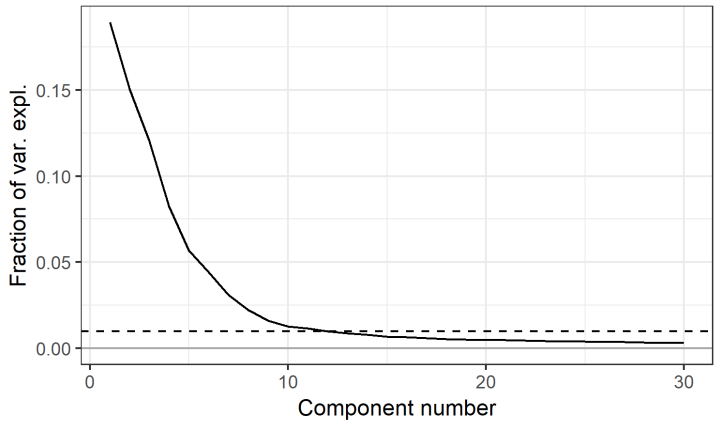


Figure 2. Fraction of the variance of the functional covariates explained by the multivariate functional principal components (solid line) and threshold (dashed line) equal to 0.01.

that achieve the higher reductions of the PRESS statistic and percentages of variance explained larger than the threshold value (0.01) fixed in Figure 2. The normality assumption for the errors is supported by the Shapiro-Wilk test on the errors (p-value $= 0.11$).

To give the same importance to the functional covariates and the scalar response, the functional control charts have been built by choosing the Bonferroni correction as $\alpha_{T^2} = \alpha_{SPE} = \alpha/4$ and $\alpha_y = \alpha/2$ as proposed in Eq.(24). The upper control limits of the Hotelling T^2 and the SPE control charts have been calculated as the Phase I empirical $(1-\alpha/2)$ quantiles of the corresponding statistic. The limits for the response prediction error control chart have been calculated using Eq.(20). Figures 4 to 6 show the three control charts described in Section 3.4 used for Phase II monitoring of the upcoming voyages.

Figures 7a and 7b report the boxplots of the functional covariates contribution to the Hotelling T^2 and SPE statistics, respectively. From those figures it is clear that the contributions are not identically distributed and then different contribution limits for each variable must be set for both the Hotelling T^2 and the SPE statistics. Contribution limits have been obtained by using the Bonferroni-like correction discussed in Section 3.4 as the Phase I empirical $(1 - \alpha/(4P))$ quantiles of the corresponding contribution.

OC points signaled by T^2 or SPE control charts are investigated by means of the corresponding contribution plot and the most paradigmatic cases are illustrated and discussed in what follows.

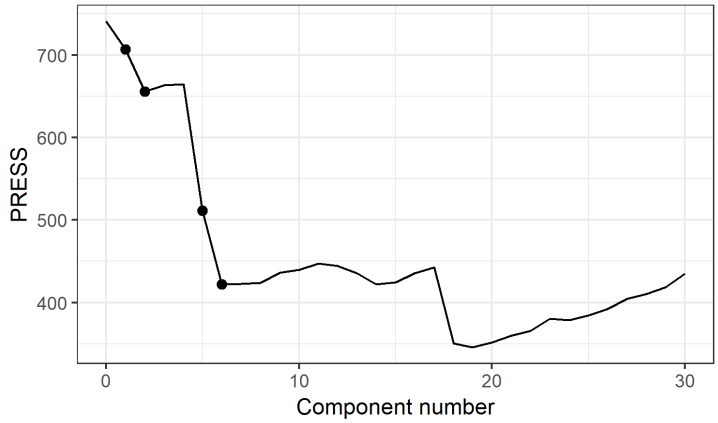


Figure 3. PRESS statistic calculated by Eq.(8) as a function of the the first m retained principal components. Trivially, when $m = 0$, the PRESS is obtained on the basis of the predictions calculated as the sample mean of the response variable. The points correspond to the multivariate functional principal components retained in the model.

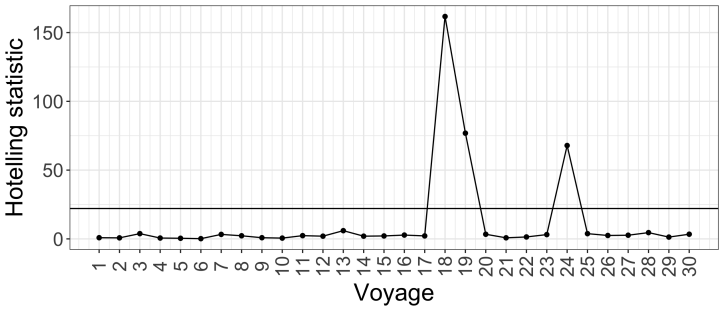


Figure 4. Hotelling T^2 control chart used for monitoring the functional covariates of the Phase II voyages.

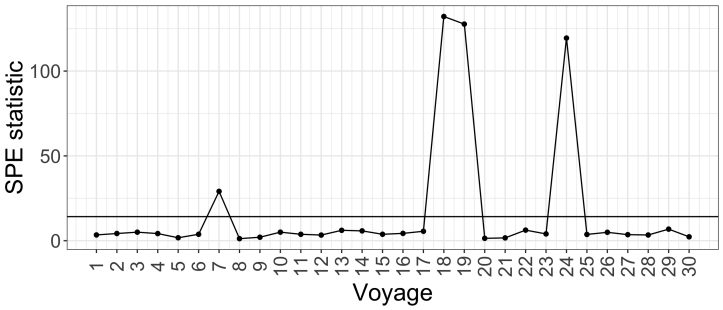


Figure 5. SPE control chart used for monitoring the functional covariates of the Phase II voyages.

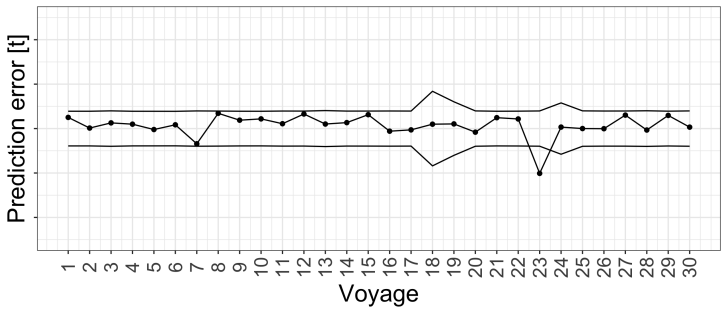
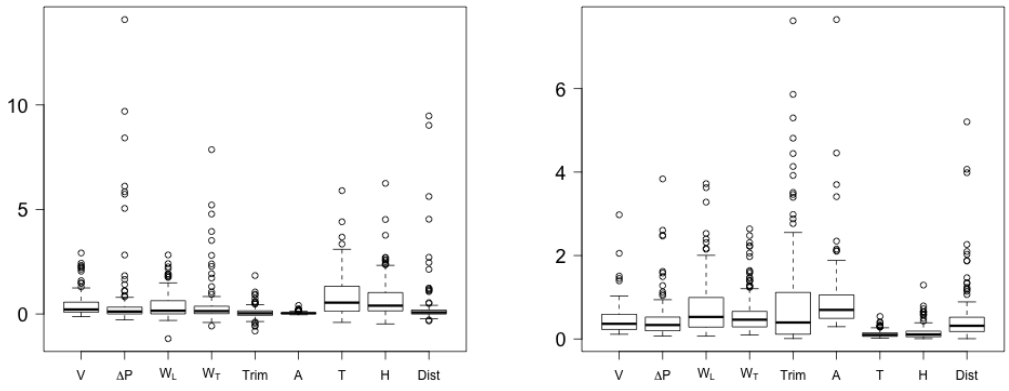


Figure 6. Response prediction error control chart used for monitoring the response variable of the Phase II voyages.



(a) (a) (b) (b)

Figure 7. 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.

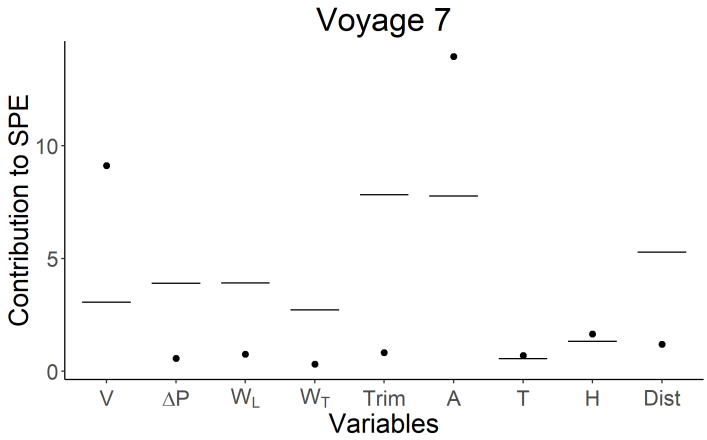


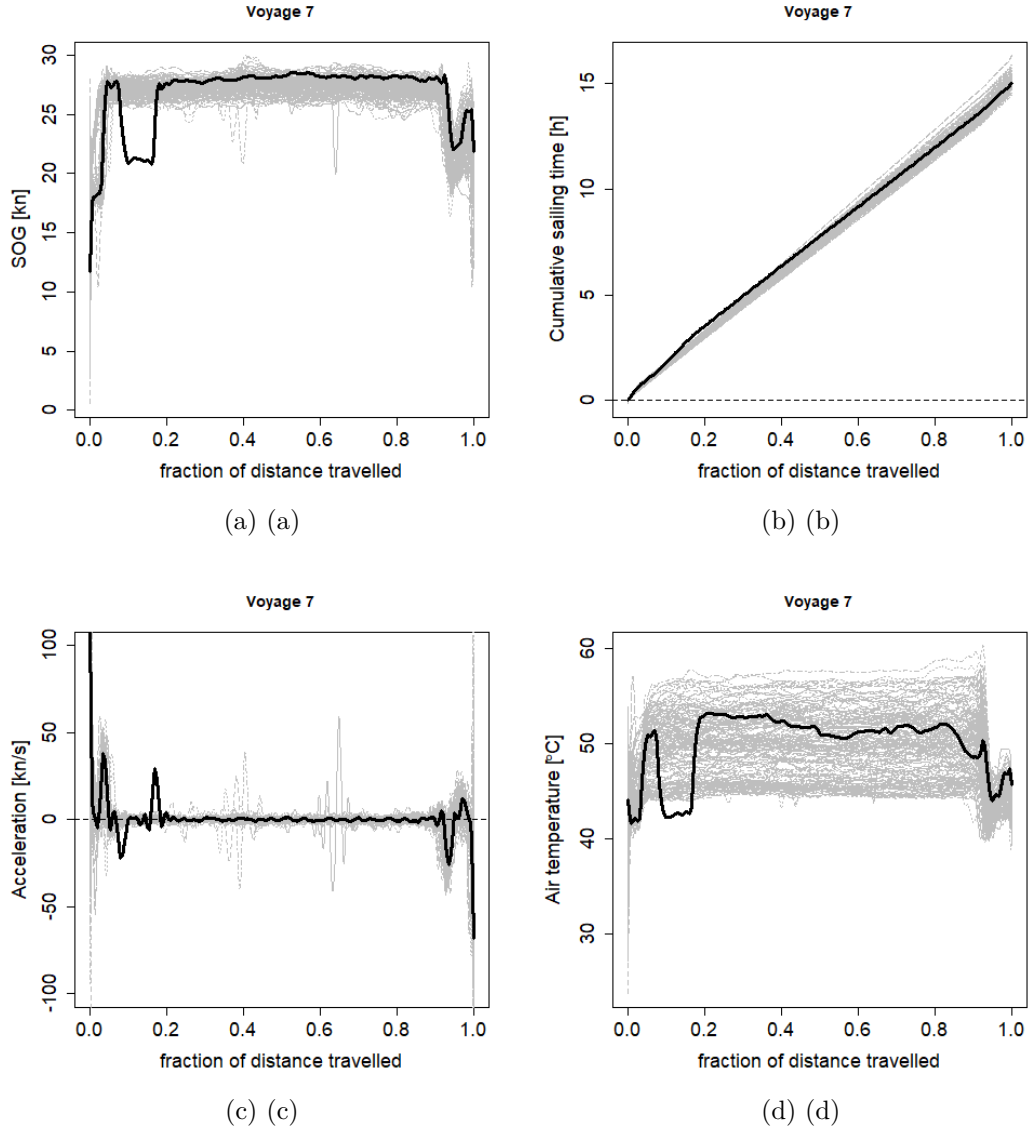
Figure 8. Contribution of the functional covariates to the SPE statistic for voyage 7. The points are the contributions of the variables, while the black dashes are the limits calculated on the basis of the reference voyages.

4.3.1 Voyage 7

The first OC voyage is the voyage 7 signaled by the SPE statistic. The contribution to the SPE statistic of the functional covariates is shown in Figure 8. The main variables responsible for the OC condition are the acceleration (A) and SOG (V), while also for the cumulative sailing time (H) and mean air temperature of the engines (T), the contributions are larger than their limit. The plots of the functional covariates can be then explored, as shown in Figure 9. From the SOG profile in Figure 9a, it is clear that the ship was sailing at a speed lower than usual for a short initial fraction of the voyage. This assuredly affected the sailing time (Figure 9b). In fact, by sailing at a SOG higher than average after the slowdown, the ship completed the voyage without delay. Accordingly, the acceleration variable (Figure 9c) shows the variations in SOG, and the mean air temperature of the engines (Figure 9d) reflects the same behaviour of the SOG.

4.3.2 Voyage 18

Voyage 18 is highlighted as OC in both T^2 and SPE functional control charts. Note that a high value of the Hotelling T^2 statistic results in larger intervals for the response prediction error control chart. In Figure 10, the contribution to the Hotelling T^2 statistic signaled the SOG (V), the power difference between port and starboard propeller shafts (ΔP), and the cumu-



(a) (a) (b) (b)

(c) (c) (d) (d)

Figure 9. Observations of the functional covariates signaled as anomalous by the functional control charts for monitoring voyage 7, i.e. (a) SOG (V), (b) cumulative sailing time (H), (c) acceleration (A), and (d) mean air temperature of the engines (T). In each plot, the black line indicates the functional observation for the voyage 7, while in grey the reference functional observations are plotted.

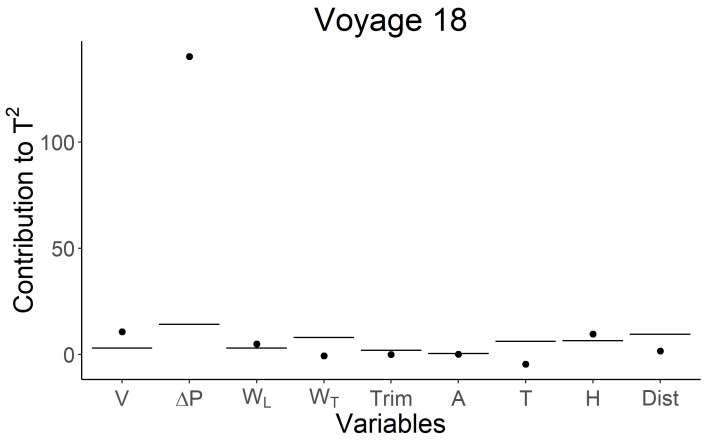


Figure 10. Contribution of the functional covariates to the Hotelling T^2 statistic for voyage 18. The points are the contributions of the variables, while the black dashes are the limits calculated on the basis of the reference voyages.

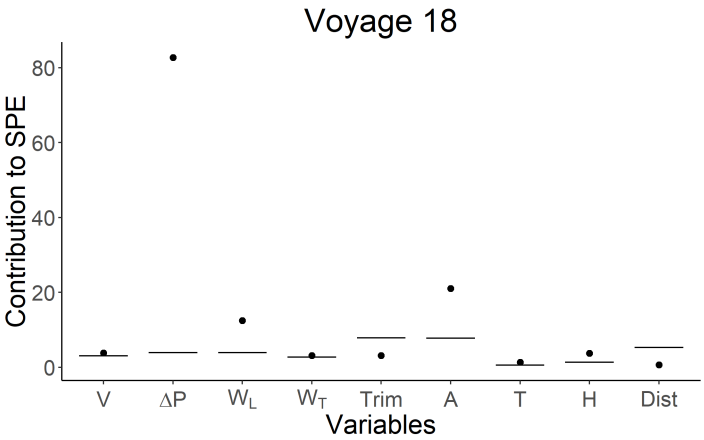


Figure 11. Contribution of the functional covariates to the SPE statistic for voyage 18. The points are the contributions of the variables, while the black dashes are the limits calculated on the basis of the reference voyages.

lative sailing time (H) as anomalous variables. In Figure 11, the contribution to the SPE statistic also indicates the longitudinal wind component (W_L) and the acceleration (A) variables. The plots of the functional covariates can be then exploited, as shown in Figure 12. As the voyage 7, this voyage is characterized by an atypical SOG profile (Figure 12a), with a lower average value throughout the entire voyage and alternation of accelerations and decelerations (Figure 12c). This affected the sailing time (Figure 12b), which shows a strong delay of the ship. By looking at the profile for the power difference between port and starboard propeller shafts (ΔP) in Figure 12d, 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 (Figure 12e).

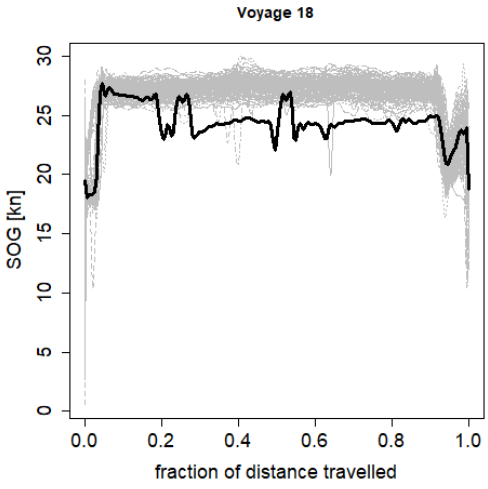
4.3.3 Voyage 23

For voyage 23, the functional control charts on the covariates do not signal any anomaly, while the response prediction error control chart indicates that the total CO_2 emissions were lower than the predicted value. In this case, possible causes are to be investigated outside the set of variables that have been chosen as covariates.

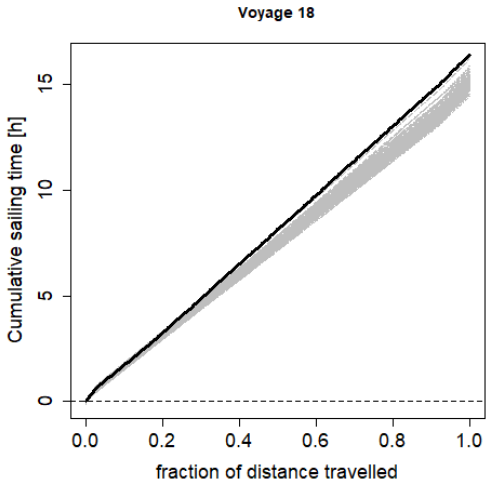
5 Discussion: real-time monitoring

Up to this point, the proposed procedure has been presented by assuming that the functional covariates have been fully observed in $\mathcal{T} = (0, 1)$. In what follows we discuss the capability of the proposed procedure for real-time monitoring of functional covariates and scalar response, i.e., before the end of a voyage. To do this, let us denote with t^* the current instant at which we want to perform the real-time monitoring and with k^* the fraction of travelled distance at t^*. The response variable to be monitored, denoted by y^*, is the total CO_2 emissions up to t^*. Accordingly, the functional covariates need to be warped into the domain $(0, k^*)$. However, the total distance travelled at the end of the voyage is not known yet at t^*, the following steps are therefore required to calculate k^*:

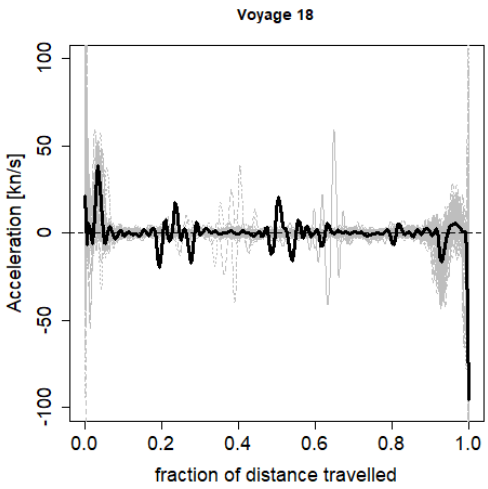
1. consider the current GPS position of the ship p^*, given by its longitude and latitude;


Voyage 18
SOG [kn]
0
5
10
15
20
25
30
0.0
0.2
0.4
0.6
0.8
1.0
fraction of distance travelled

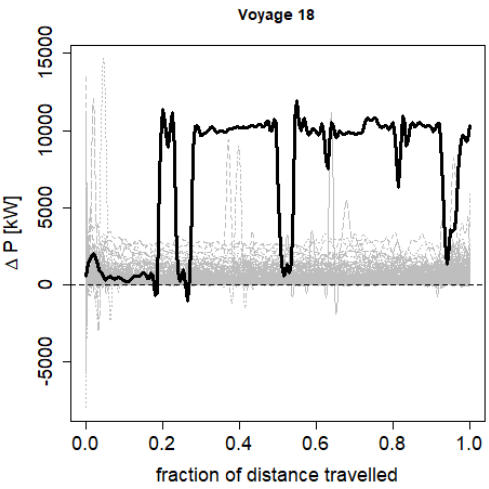
(a) (a)


Voyage 18
Cumulative sailing time [h]
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
fraction of distance travelled

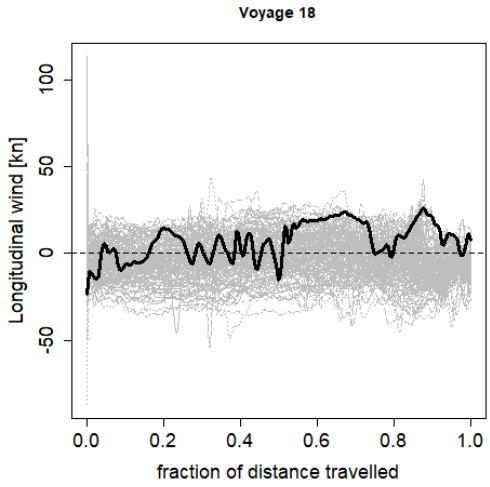
(b) (b)


Voyage 18
Acceleration [kn/s]
-100
-50
0
50
100
0.0
0.2
0.4
0.6
0.8
1.0
fraction of distance travelled

(c) (c)


Voyage 18
Δ P [kW]
-5000
0
5000
10000
15000
0.0
0.2
0.4
0.6
0.8
1.0
fraction of distance travelled

() (d)



(e) (e)

Figure 12. Observations of the functional covariates signaled as anomalous by the functional control charts for monitoring voyage 18, i.e. (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 functional observation for the voyage 18, while in grey the reference functional observations are plotted.

2. identify the point $\overline{p}^*$ on the nominal route as that with minimal distance from the current position of the ship at the considered instant t^* (Figure 13);

3. calculate the fraction of travelled distance k^* as the ratio between the length d^* of the nominal route from the departure port to $\overline{p}^*$ and the length of the whole nominal route d, i.e. $k^* = d^*/d \in (0, 1)$.

Thus, at given instant t^*, functional covariates for the new partially-observed voyage can be warped into the domain $(0, k^*)$ by considering the map $f : (0, t^*) \to (0, k^*)$, which associates to each $t \in (0, t^*)$ the fraction of travelled distance as $k(t) = d(t)/d$, where $d(t)$ denotes the distance travelled until $t \leqslant t^*$.

The reference data set at t^* can be then obtained by truncating the reference observations of covariates at k^*, so that the new functional domain is $(0, k^*)$. The data set so obtained can be used to repeat the Phase I model estimation, to set new limits for the monitoring statistics for every k^* as described in Section 3.2. It is worth noting that by repeating the model estimation, the set $\mathcal{M}$ of retained principal components may vary.

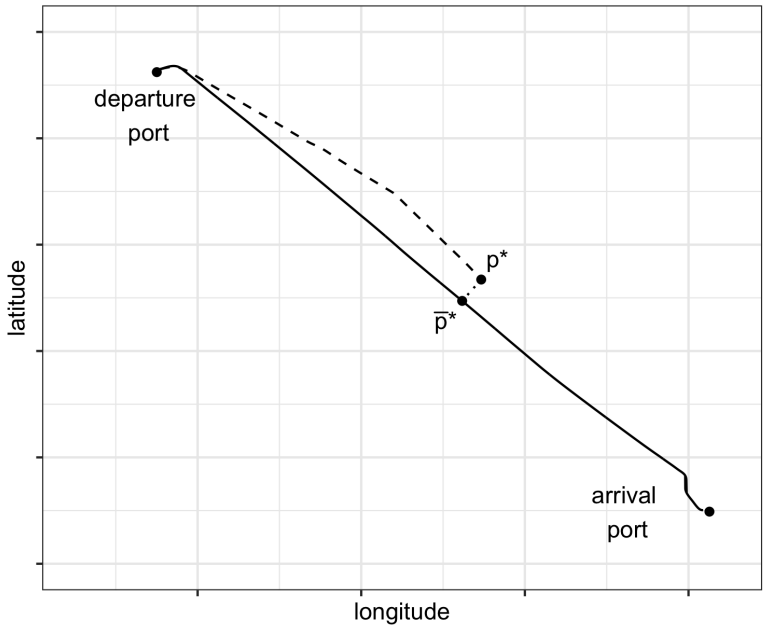


Figure 13. Graphical example showing how $k^* = d^*/d$ is determined for a new voyage. The dashed curve represents the route travelled by the ship up to the current GPS position, which is the point labeled as p^*. The solid curve represents the nominal route and its point nearest to p^* is labeled as $\overline{p}^*$. d^* is the length of the portion of the solid curve from the departure port to $\overline{p}^*$. d is the total length of the solid curve.

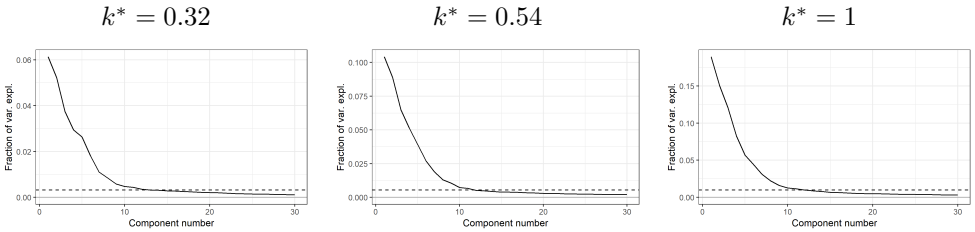


Figure 14. Fraction of the variance of the functional covariates explained by the multivariate functional principal components (solid line) and threshold (dashed line) calculated on the basis of the Phase I dataset with functional data observed up to different fractions of travelled distance ($k^* = 0.32, 0.54, 1$).

Ideally, the Phase II monitoring procedure discussed in Section (ii) can be then performed at every k^* i.e., at every t^*. However, in this work, as example, the real-time monitoring is performed only at two randomly chosen values of k^*, say 0.32, and 0.54. Figures 14 shows the fraction of variance explained by the multivariate functional principal components in the models estimated at $k^* = 0.32$, $k^* = 0.54$, and $k^* = 1$, respectively. Note that the plot on the right is Figure 2, reported again for comparison purpose. Analogously, Figure 15 shows the PRESS statistic calculated as a function of the number of multivariate functional principal components in the models estimated at $k^* = 0.32$, $k^* = 0.54$, and $k^* = 1$, respectively. The plot on the right is Figure 3. Moreover, Figure 16 reports the three control charts for all the Phase II voyages considered in Section 4 corresponding to the three k^* values. For comparison purpose, the last column reports again the proposed control charts for the fully-observed ($k^* = 1$) voyage (i.e., those already shown in Figures 4 to 6). Consistently with the lower predictive power due to the partial observation of the data, the following four dynamic scenarios can be encountered: OC states that are promptly signaled before the end of a voyage (e.g., SPE statistic for voyage 7); OC states that are not signaled before the end of a voyage (e.g., T^2 statistic for voyage 24); temporary OC that come back IC at the end of a voyage (e.g., prediction error in voyage 24); IC states that persist during the entire voyage. From the last row of Figure 16, we can observe that control limits of the response prediction error control chart become wider as k^* increases due to the fact that the response variable (total CO_2 emissions) is cumulative.

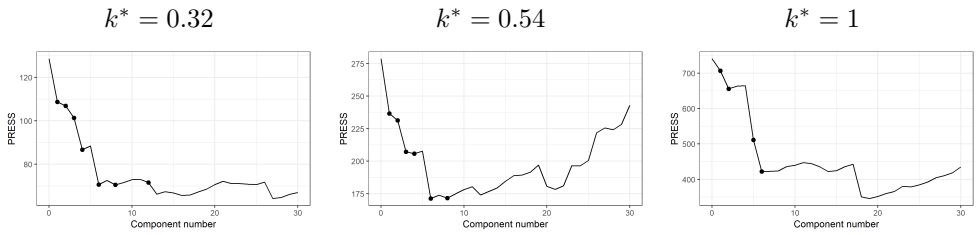


Figure 15. PRESS statistic calculated as a function of the first m retained principal components on the basis of the Phase I dataset with functional data observed up to different fractions of travelled distance ($k^* = 0.32, 0.54, 1$).

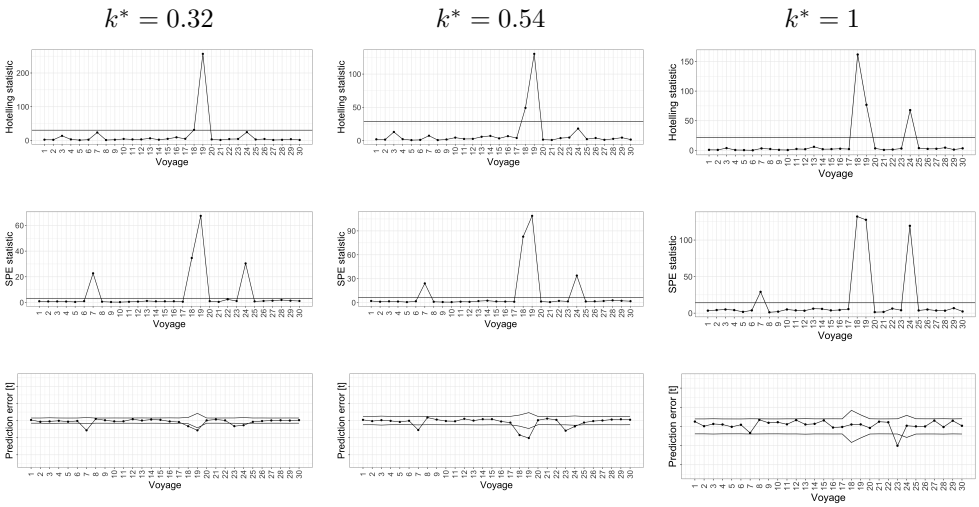


Figure 16. Hotelling T^2, SPE and prediction error control charts for Phase II voyages at different fractions of travelled distance ($k^* = 0.32, 0.54, 1$).

6 Conclusions

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 monitoring procedure for functional covariates and a scalar response related to them is proposed. To the best of the authors' knowledge, this is absolutely new in both statistical and maritime field. A suitable warping function is proposed to register 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 through a normalization approach into a multivariate functional linear 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. The discussion of the optimal choice of the set of functional principal components to keep, with the aim of considering also the variability in the covariates that is useful for the prediction of the response, is indeed beneficial for the mathematical and technological interpretation of out-of-control alarms. In case of an out-of-control signal, the corresponding contribution plots are demonstrated to be powerful tools for supporting diagnosis of faults.

The proposed procedure is also shown by means of response prediction error control chart for monitoring ship CO_2 emissions and plausibly indicating if an anomaly occurs in the scalar-on-function linear model, i.e., outside the ship operational conditions monitored on board. 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 in order 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.

Finally, a discussion has been provided to illustrate the potential of the proposed monitoring procedure in giving indications and making predictions even if observations are still not complete, which would greatly help shipping practitioners in managerial decision making.

References

Abramowicz, K.; Häger, C. K.; Pini, A.; Schelin, L.; Sjöstedt de Luna, S.; and 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*.

Bialystocki, N. and 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), pp. 157–166.

Bocchetti, D.; Lepore, A.; Palumbo, B.; and Vitiello, L. (2015). "A statistical approach to ship fuel consumption monitoring". *Journal of Ship Research*, 59(3), pp. 162–171.

Chen, L.-H. and Jiang, C.-R. (2017). "Multi-dimensional functional principal component analysis". *Statistics and Computing*, 27(5), pp. 1181–1192.

Chiou, J.-M.; Chen, Y.-T.; and Yang, Y.-F. (2014a). "Multivariate functional principal component analysis: A normalization approach". *Statistica Sinica*, pages 1571–1596.

Chiou, J.-M.; Yang, Y.-F.; and Chen, Y.-T. (2016). "Multivariate functional linear regression and prediction". *Journal of Multivariate Analysis*, 146, pp. 301–312.

Chiou, J.-M.; Zhang, Y.-C.; Chen, W.-H.; and Chang, C.-W. (2014b). "A functional data approach to missing value imputation and outlier detection for traffic flow data". *Transportmetrica B: Transport Dynamics*, 2(2), pp. 106–129.

Colosimo, B. M. and Pacella, M. (2007). "On the use of principal component analysis to identify systematic patterns in roundness profiles". *Quality and reliability engineering international*, 23(6), pp. 707–725.

Colosimo, B. M. and Pacella, M. (2010). "A comparison study of control charts for statistical monitoring of functional data". *International Journal of Production Research*, 48(6), pp. 1575–1601.

Erto, P.; Lepore, A.; Palumbo, B.; and Vitiello, L. (2015). "A Procedure for Predicting and Controlling the Ship Fuel Consumption: Its Implementation and Test". *Quality and Reliability Engineering International*, 31(7), pp. 1177–1184.

European Commission (2015). "Proposal for a regulation of the monitoring,reporting and verification of carbon dioxide emissions frommaritime transport and amending regulation (EU) no 525/2013.European Commission Transportation".

Happ, C. (2018). *MFPCA: Multivariate Functional Principal Component Analysis for Data Observed on Different Dimensional Domains*. R package version 1.3.

Happ, C. and Greven, S. (2016). "Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains". *Journal of the American Statistical Association*.

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 the method of calculation of the attained EnergyEfficiency 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.".

Jolliffe, I. (2002). *Principal Component Analysis*. Springer.

Kourti, T. and MacGregor, J. F. (1996). "Multivariate SPC methods for process and product monitoring". *Journal of quality technology*, 28(4), pp. 409–428.

Lepore, A.; Reis, M. S. d.; Palumbo, B.; Rendall, R.; and Capezza, C. (2017). "A comparison of advanced regression techniques for predicting ship CO_2 emissions". *Quality and Reliability Engineering International*.

Lu, R.; Turan, O.; Boulougouris, E.; Banks, C.; and Incecik, A. (2015). "A semi-empirical ship operational performance prediction model for voyage optimization towards energy efficient shipping". *Ocean Engineering*, 110, pp. 18–28.

Mandel, B. (1969). "The regression control chart". *Journal of Quality Technology*, 1(1), pp. 1–9.

Montgomery, D. C. (2007). *Introduction to statistical quality control*. John Wiley & Sons.

Montgomery, D. C.; Peck, E. A.; and Vining, G. G. (2012). *Introduction to linear regression analysis*, volume 821. John Wiley & Sons.

Nomikos, P. and MacGregor, J. F. (1995a). "Multi-way partial least squares in monitoring batch processes". *Chemometrics and intelligent laboratory systems*, 30(1), pp. 97–108.

Nomikos, P. and MacGregor, J. F. (1995b). "Multivariate SPC charts for monitoring batch processes". *Technometrics*, 37(1), pp. 41–59.

Noorossana, R.; Saghaei, A.; and Amiri, A. (2012). *Statistical analysis of profile monitoring*, volume 865. John Wiley & Sons.

Perera, L. P. and Mo, B. (2016). "Emission control based energy efficiency measures in ship operations". *Applied Ocean Research*, 60, pp. 29–46.

Petersen, J. P.; Jacobsen, D. J.; and Winther, O. (2012). "Statistical modelling for ship propulsion efficiency". *Journal of marine science and technology*, 17(1), pp. 30–39.

Ramsay, J.; Wickham, H.; Graves, S.; and Hooker, G. (2015). *fda: Functional Data Analysis*. R package version 2.4.4.

Ramsay, J. O.; Hooker, G.; and Graves, S. (2009). *Functional data analysis with R and MATLAB*. Springer Science & Business Media.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Wiley Online Library.

Reiss, P. T.; Goldsmith, J.; Shang, H. L.; and Ogden, R. T. (2017). "Methods for Scalar-on-Function Regression". *International Statistical Review*, 85(2), pp. 228–249.

Šidák, Z. (1967). "Rectangular confidence regions for the means of multivariate normal distributions". *Journal of the American Statistical Association*, 62(318), pp. 626–633.

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.; and Pandey, A. (2015). "Third IMO GHG study 2014". *International Maritime Organization (IMO), London, UK*.

Wang, J.-L.; Chiou, J.-M.; and Mueller, H.-G. (2016). "Functional Data Analysis". *Annual Review of Statistics and Its Application*, 3, pp. 257–295.

Woodall, W. H.; Spitzner, D. J.; Montgomery, D. C.; and Gupta, S. (2004). "Using control charts to monitor process and product quality profiles". *Journal of Quality Technology*, 36(3), pp. 309.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

09/2019 Antonietti, P.F.; Facciola', C.; Verani, M.
Unified analysis of Discontinuous Galerkin approximations of flows in fractured porous media on polygonal and polyhedral grids

10/2019 Abramowicz, K.; Pini, A.; Schelin, L.; Sjostedt de Luna, S.; Stamm, A.; Vantini, S.
Domain selection and family-wise error rate for functional data: a unified framework

11/2019 Benacchio, T.; Klein, R.
A semi-implicit compressible model for atmospheric flows with seamless access to soundproof and hydrostatic dynamics

08/2019 Prouse, G.; Stella, S.; Vergara, C.; Engelberger, S.; Trunfio, R.; Canevascini, R.; Quarteroni, A.;
Computational analysis of turbulent haemodynamics in radiocephalic arteriovenous fistulas with different anastomotic angles

05/2019 Gasperoni, F.; Ieva, F.; Paganoni, A.M.; Jackson, C.; Sharples, L.
Evaluating the effect of healthcare providers on the clinical path of Heart Failure patients through a novel semi-Markov multi-state model

07/2019 Dal Santo, N.; Manzoni, A.
Hyper-reduced order models for parametrized unsteady Navier-Stokes equations on domains with variable shape

06/2019 Pagani, S.; Manzoni, A.; Carlberg, K.
Statistical closure modeling for reduced-order models of stationary systems by the ROMES method

04/2019 Delpopolo Carciopolo, L.; Formaggia, L.; Scotti, A.; Hajibeygi, H.
Conservative multirate multiscale simulation of multiphase flow in heterogeneous porous media

03/2019 Ratti, L.; Verani, M.
A posteriori error estimates for the monodomain model in cardiac electrophysiology

02/2019 Micheletti, S.; Perotto, S.; Soli, L.
Topology optimization driven by anisotropic mesh adaptation: towards a free-form design