

# A functional data analysis approach for the monitoring of ship CO<sub>2</sub> emissions

## *Abordagem de análise de dados funcionais para o monitoramento de emissões de CO<sub>2</sub> de navios*

Christian Capezza<sup>1</sup> , Fabio Centofanti<sup>1</sup> , Antonio Lepore<sup>1</sup> , Biagio Palumbo<sup>1</sup> 

<sup>1</sup>University of Naples Federico II, Department of Industrial Engineering, Naples, Italy.

E-mail: biagio.palumbo@unina.it

**How to cite:** Capezza, C., Centofanti, F., Lepore, A., & Palumbo, B. (2021). A functional data analysis approach for the monitoring of ship CO<sub>2</sub> emissions. *Gestão & Produção*, 28(3), e152. <https://doi.org/10.1590/1806-9649-2021v28e152>

**Abstract:** Sensing networks provide nowadays massive amounts of data that in many applications provide information about curves, surfaces and vary over a continuum, usually time, and thus, can be suitably modelled as functional data. Their proper modelling by means of functional data analysis approaches naturally addresses new challenges also arising in the statistical process monitoring (SPM). Motivated by an industrial application, the objective of the present paper is to provide the reader with a very transparent set of steps for the SPM of functional data in real-world case studies: i) identifying a finite dimensional model for the functional data, based on functional principal component analysis; ii) estimating the unknown parameters; iii) designing control charts on the estimated parameters, in a nonparametric framework. The proposed SPM procedure is applied to a real-case study from the maritime field in monitoring CO<sub>2</sub> emissions from real navigation data of a roll-on/roll-off passenger cruise ship, i.e., a ship designed to carry both passengers and wheeled vehicles that are driven on and off the ship on their own wheels. We show different scenarios highlighting clear and interpretable indications that can be extracted from the data set and support the detection of anomalous voyages.

**Keywords:** Profile monitoring; Functional principal component analysis; CO<sub>2</sub> emissions; Control charts; Statistical process monitoring.

**Resumo:** As redes de detecção fornecem hoje em dia grandes quantidades de dados que em muitas aplicações fornecem informações sobre curvas, superfícies e variam em um contínuo, geralmente o tempo, e, portanto, podem ser modelados adequadamente como dados funcionais. Sua modelagem adequada por meio de abordagens de análise de dados funcionais naturalmente aborda novos desafios que surgem também no monitoramento estatístico de processos (SPM). Motivado por uma aplicação industrial, o objetivo do presente artigo é fornecer ao leitor um conjunto muito transparente de etapas para o SPM de dados funcionais em estudos de caso do mundo real: i) identificar um modelo dimensional finito para os dados funcionais, com base na análise de componentes principais funcionais; ii) estimar os parâmetros desconhecidos; iii) desenhar cartas de controle sobre os parâmetros estimados, em uma estrutura não paramétrica. O procedimento SPM proposto é aplicado a um estudo de caso real do campo marítimo no

Received Sept. 28, 2020 - Accepted May 13, 2021

Financial support: None.



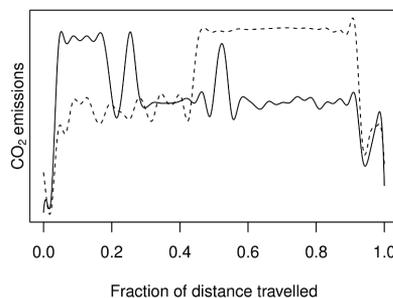
This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

monitoramento das emissões de CO<sub>2</sub> de dados de navegação reais de um navio de cruzeiro roll-on / roll-off de passageiros, ou seja, um navio projetado para transportar passageiros e veículos com rodas que são levados para dentro e para fora do navio em suas próprias rodas. Mostramos diferentes cenários destacando indicações claras e interpretáveis que podem ser extraídas do conjunto de dados e apoiar a detecção de viagens anômalas.

**Palavras-chave:** Monitoramento de perfis; Análise de componentes principais funcionais; CO<sub>2</sub> emissões; Gráficos de controle; Monitoramento estatístico do processo.

## 1 Introduction

In many applications, the development of data-acquisition systems allows the gathering of massive amount of data that can be suitably modelled as functional data, that is as functions varying over a continuum. Functional data analysis (FDA) refers to the set of statistical methods where the observation units are functional data. Thorough overviews of FDA techniques are provided by Ramsay & Silverman (2005); Horváth & Kokoszka (2012); Kokoszka & Reimherr (2017). More specific theoretical insight can be found in Hsing & Eubank (2015) and Bosq (2012). Each functional data observation is usually obtained from discrete measurements over the continuous domain. Thus, standard multivariate methods could be in principle applied, even though they fail when the number of observations is much less than the number of discrete measurements. This typical high dimensionality issue cannot be overlooked by collapsing or averaging measurements when the aim of the analysis is to monitor or control the stability over time of quality characteristics apt to be modelled as a functional data. This approach has been used extensively in the literature (Bocchetti et al., 2015; Erto et al., 2015; Lepore et al., 2019; Capezza et al., 2019), however, there is a serious risk of discarding valuable information. For instance, Figure 1 shows two CO<sub>2</sub> emission functions from the real-case study in the maritime field described in Section 3.



**Figure 1.** Two CO<sub>2</sub> emission functions with similar means throughout the domain but different shapes.

If an approach where each function is replaced by its mean all over the domain is used, then these functions are summarized by very similar numbers (not reported for confidentiality reasons). However, it is clear from Figure 1 that these two functions show very different shapes and, thus, they should be treated accordingly. In a statistical process monitoring application, this could result in very serious lack of ability to detect anomalous observations.

Statistical process monitoring of functional data is known also as *profile monitoring*, where functional data are referred to as *profiles*. As in the classical univariate and multivariate setting, where data are represented by scalars or vectors, profile

monitoring has the task of continuously monitoring the quality characteristic and of triggering a signal when assignable sources of variations (i.e., special causes) act on it. When this happens, the process is said to be out of control (OC). Otherwise, when only normal sources of variation (i.e., common causes) apply, the process is said to be in-control (IC). As discussed by Woodall et al. (2004), all the approaches for profile monitoring share the following structure: i) identifying a finite dimensional model for the functional data; ii) estimating the unknown parameters; iii) designing control charts on the estimated parameters. In particular, the book of Noorossana et al. (2012) represents a comprehensive overview of profile monitoring methods. Pini et al. (2018) proposed a two-step profile monitoring approach where, firstly, the informative parts of the functional data to be monitored are selected by means of the inferential interval-wise testing procedure (Pini & Vantini, 2017) and then the monitoring procedure is performed on the basis of the information that the functions contain in the selected domains. Menafoglio et al. (2018) introduced a new approach for monitoring probability density functions based on simplicial functional principal component analysis. Grasso et al. (2016) presented a novel approach for profile monitoring that combines the functional principal component analysis and the use of parametric warping functions. More recently, Capezza et al. (2020) extended classical multivariate techniques to the monitoring of multivariate functional data and a scalar quality characteristic related to them. Centofanti et al. (2020) expand the Mandel's regression control chart (Mandel, 1969) to the functional setting, that is a control chart elaborated on the functional residuals obtained from a function-on-function regression of the quality characteristic profile on concurrent functional covariates. Other relevant contributions in this field include the work of Jin & Shi (1999), Colosimo & Pacella (2007), Colosimo & Pacella (2010), Grasso et al. (2017), and Bersimis et al. (2018).

Motivated by an industrial application, the objective of the present paper is to provide the reader with a very transparent set of steps for monitoring profiles in real-world case studies. In particular, the proposed method can be divided into three main steps. Firstly, the functional data are obtained from the raw data through a smoothing technique based on spline functions. Then, a functional principal component analysis (FPCA), that is the functional extension of the classical (non-functional) principal component analysis (PCA) (Jolliffe, 2011), is performed in order to extract the relevant principal component scores. Lastly, the retained principal component scores are used in a monitoring procedure that is based on the simultaneous application of the *Hotelling's*  $T^2$  and the *squared prediction error* ( $SPE$ ) control charts in a nonparametric framework.

A complete overview of smoothing techniques for functional data is provided by Ramsay & Silverman (2005), where methods based on least squares and roughness penalties are presented under a practical point of view. More generally, references on smoothing spline estimators for nonparametric regression are Wahba (1990); Green & Silverman (1993); Eubank (1999), and Gu (2013). A survey of FPCA, and its use in explanatory analysis, modeling and forecasting, and classification of functional data is provided by Shang (2014). The  $T^2$  and  $SPE$  control charts are widely used for multivariate statistical process monitoring (Montgomery, 2007). See Lowry & Montgomery (1995) for a review on multivariate control charts.

Finally, the proposed monitoring procedure is applied to a real-case study from the maritime field in monitoring CO<sub>2</sub> emissions during the navigation phase of a roll-on/roll-off passenger (Ro-Pax) cruise ship, i.e., a ship designed to carry both passengers and wheeled vehicles that are driven on and off the ship on their own wheels, whose data are courtesy of the owner Grimaldi Group.

The paper is structured as follows. Section 2 introduces the proposed procedure. In particular, in Section 2.1 we discuss how to obtain the functional data from the raw observations through data smoothing techniques; Section 2.2 describes the FPCA, and Section 2.3 introduces the monitoring procedure based on the  $T^2$  and  $SPE$  control charts. The real-case study in the shipping industry is presented in Section 3. Section 4 concludes the paper. All computations and plots have been obtained by using the software environment R (R Core Team, 2020), where the proposed procedure is implemented through the package `funcharts` (Capezza et al., 2021)

## 2 Methodology

As stated before, the proposed methodology for profile monitoring is composed of three main steps:

1. data smoothing: the raw observations are converted to functional data;
2. FPCA: the infinite dimensional problem is translated into a finite dimensional one by means of an optimal functional data approximation;
3. monitoring procedure: the principal component scores are used as input to build the  $T^2$  and  $SPE$  control charts.

In the following Sections 2.1, 2.2 and 2.3, these steps are illustrated.

### 2.1 Data smoothing

Data are collected by devices in a discrete fashion, that is as  $n$  discrete observed curves  $\{Y_i(t_j), j=1, \dots, p\}_{i=1, \dots, n}$ , where  $\{t_j\}_{j=1, \dots, p}$  are the observation points in a given closed interval  $\mathcal{T} \subseteq \mathbb{R}$ . Hence, appropriate methods are required to convert discrete raw data  $\{Y_i(t_j)\}$  into functional data  $\{X_i(t)\}$  computable for any  $t \in \mathcal{T}$ , which are random realizations of a functional quality characteristic. If the discrete data are assumed without any measurement error, functional data can be theoretically drawn up by merely connecting the whole set of points  $\{Y_i(t_j), j=1, \dots, p\}_{i=1, \dots, n}$ . However, this does not represent the ordinary situation. When measurement error is present, each discrete observation is expressed as

$$Y_i(t_j) = X_i(t_j) + \varepsilon_{ij} \tag{1}$$

where  $\varepsilon_{ij}$  are zero mean random errors with equal variances. Trivially, note that Equation 1 degenerates in the previous case when the variance  $\varepsilon_{ij}$  tends to zero. From Equation 1, data smoothing techniques aim to recover the functional data by discarding exogenous perturbation due to error terms  $\varepsilon_{ij}$ . Functional data are intrinsically infinite dimensional; that is, infinite values are needed to completely specify them, precisely, the values at each possible argument  $t \in \mathcal{T}$ . To this end, a common approach consists of representing each functional datum  $\{X_i(t)\}$  by introducing a basis function system, i.e., a set of  $K$  known, linearly independent functions  $\Phi = (\phi_1, \dots, \phi_K)^T$  that have the property that we can approximate arbitrarily well any function by taking a weighted sum

or linear combination of a sufficiently large number of these functions (Ramsay & Silverman, 2005). Then, we have

$$X_i(t) = \sum_{l=1}^K c_{il} \phi_l(t) = c_i^T \Phi(t), \quad t \in \mathcal{T} \tag{2}$$

where  $c_i = (c_{i1}, \dots, c_{iK})^T$  is the coefficient vector for each curve. Then, the problem of recovering the functional data  $\{X_i(t)\}$  in Equation 2 reduces to the estimation of the unknown coefficient vectors  $c_i$  for every  $i = 1, \dots, n$ . In particular, the coefficient vector  $c_i$  is estimated as  $\hat{c}_i$  by minimizing the following penalized sum of squares error

$$\hat{c}_i = \underset{c \in \mathbb{R}^K}{\operatorname{argmin}} \left\{ \sum_{j=1}^p \left( Y_i(t_j) - c^T \Phi(t_j) \right)^2 + \lambda c^T R c \right\}, \tag{3}$$

where  $\lambda > 0$  is a smoothing parameter and  $R$  is a matrix whose  $(i, j)$ -th entry is  $\int_{\mathcal{T}} \phi_i^{(m)}(t) \phi_j^{(m)}(t) dt$ , with  $\phi^{(m)}$  the  $m$ -th derivative of  $\phi$ . Finally, the functional data we are interested in are as follows

$$\hat{X}_i(t) = \hat{c}_i^T \Phi(t), \quad t \in \mathcal{T}. \tag{4}$$

Note that, to obtain the functional data as in Equation 4, some choices should be made; these are discussed in the following. As basis functions  $\Phi$ , the B-spline basis system is the most common choice in case of non-periodic functional data because it has good computational properties and great flexibility (Ramsay & Silverman, 2005). This implicitly assumes that the curves considered are well approximated by a spline function. Splines are optimal in the sense of being the smoothest possible functions interpolating the data (Green and Silverman, 1993). Spline functions divide the functional domain into subintervals, by means of break points. Over any subinterval, the spline is a polynomial of specific order  $q$ , with  $q-1$  non-zero derivatives and matching proper derivative constraints between adjacent polynomials (De Boor et al., 1978). The smoothing parameter  $\lambda$  in Equation 3 is chosen as that corresponding to the minimum value assumed by the generalized cross-validation (GCV) criterion, which is a well-known method to tradeoff between variance and bias. This criterion considers the degrees of freedom of the estimated curve that vary according to  $\lambda$ . See Ramsay & Silverman (2005) for further details. The penalty on the right-hand side of Equation 3 is computed by setting  $m=2$  in the elements of the matrix  $R$ , i.e., by penalizing the function roughness, which is defined as the integrated squared second derivative of  $X_i(t)$ , calculated as  $\int_{\mathcal{T}} [X^{(2)}(t)]^2 dt = c^T R c$ . The number  $K$  of basis functions is not crucial (Cardot et al., 2003), unless it is sufficiently large to capture the local behavior of the functional data.

## 2.2 Functional principal component analysis

FPCA is a key method aimed at reducing the infinite dimensionality of the functional data, by retaining a finite number  $L$  of *principal component scores* or simply *scores*  $\{\xi_{il}\}_{l=1,\dots,L}$ , which explain the largest part of the sample variability, for each functional observation  $\{X_i(t)\}$ , obtained as described in Section 2.1 defined for  $t \in \mathcal{T}$ . Without loss of generality, in what follows, let us assume that  $\{X_i(t)\}$  have zero mean, or that they are centered by subtracting the functional sample mean. Then, scores are defined as

$$\xi_{il} = \int_{\mathcal{T}} \psi_l(t) X_i(t) dt \tag{5}$$

where  $\{\psi_l\}_{l=1,\dots,L}$  are weight functions referred to as *functional principal components* (FPCs) or simply *principal components*. The FPCs are subject to size restrictions of normalization and orthogonality, i.e.,  $\int_{\mathcal{T}} \psi_l(t)^2 dt = 1$  and  $\int_{\mathcal{T}} \psi_i(t) \psi_j(t) dt = 0$ , for  $i \neq j$ . In this way, each weight function provides new information with respect to those brought by previous FPCs. FPCs are calculated by an iterative algorithm which at each step finds the weight function that maximizes the following mean square of the scores, i.e., their sample variance,

$$\psi_l = \operatorname{argmax}_{\psi} \sum_{i=1}^n \xi_{il}^2 = \operatorname{argmax}_{\psi} \sum_{i=1}^n \left( \int_{\mathcal{T}} \psi(t) X_i(t) dt \right)^2, \quad l=1,\dots,L, \tag{6}$$

under the normalization and orthogonality constraints. Moreover, the FPCs in Equation 6 correspond to eigenfunctions of the covariance function of the process  $X$  (Ramsay & Silverman, 2005). Let us consider the function

$$\hat{X}_i^{PC}(t) = \sum_{l=1}^L \xi_{il} \psi_l(t), \quad t \in \mathcal{T} \tag{7}$$

that is the linear combination of the FPCs and the scores. It can be demonstrated that  $\hat{X}_i^{PC}(t)$  is the best  $L$ -dimensional approximation of  $X_i$  in terms of mean squared error, i.e., the quantity  $E \left[ \int_{\mathcal{T}} \left( \hat{X}_i^{PC}(t) - X_i(t) \right)^2 dt \right]$  is minimum over all the other  $L$ -dimensional linear combinations.

The choice of the number  $L$  of retained components depends on different necessities. Generally, the FPCs retained are chosen such that they explain at least a given percentage of the total variability. However, more sophisticated methods could be used as well (Jolliffe, 2011).

## 2.3 Monitoring procedure

In this step, the information provided by FPCA is used to continuously monitor the functional quality characteristic  $X$  over time. To this aim, two functional control charts

are introduced based on the following  $T^2$  and the  $SPE$  statistics. The  $T^2$  statistic is as follows

$$T_i^2 = \sum_{l=1}^L \frac{\xi_{il}^2}{\lambda_l} \tag{8}$$

where  $\lambda_1, \dots, \lambda_L$  are the variances of the scores  $\xi_{i1}, \dots, \xi_{iL}$  introduced in Equation 5 and correspond to the eigenvalues of the covariance function of  $X$ . The statistic  $T^2$  is the square distance of the projection of  $X$  from the origin of the space spanned by the FPCs  $\{\psi_l\}$ . Changes along directions orthogonal to such space are monitored by means of the  $SPE_i$  statistic, defined for each  $i$  as

$$SPE_i = \int_{\mathcal{T}} \left( X_i(t) - \hat{X}_i^{PC}(t) \right)^2 dt \tag{9}$$

where  $\hat{X}_i^{PC}$  are defined in Equation 7.

In this paper, we focus on prospective (Phase II) monitoring. Thus, a set of IC data must be preliminarily obtained in the design phase of the control charts (Phase I). Let us assume that the functional observations  $\{X_i\}$  are acquired under IC conditions; principal components  $\{\psi_l\}$  and eigenvalues  $\lambda_1, \dots, \lambda_L$  shall be estimated from the sample covariance function. Let us denote the corresponding estimates by  $\{\hat{\psi}_l\}$  and  $\hat{\lambda}_1, \dots, \hat{\lambda}_L$ . The control limits for both the  $T^2$  (Equation 8) and the  $SPE$  (Equation 9) control charts are obtained as the  $(1-\alpha)$ -quantiles of the empirical distribution of the two statistics, based on the estimated  $\{T_i^2\}$  and  $\{SPE_i\}$ . Other methods could be also used, either as in Centofanti et al. (2020), where the distribution of the two statistics is estimated through a kernel density estimation approach, or as in Nomikos & MacGregor (1995), where a parametric approach is considered. However, all these methods are expected to provide the same results for large sample size.

The parameter  $\alpha$  is chosen by using the Bonferroni correction  $\alpha = \alpha^* / L$ , where  $\alpha^*$  is the overall type I error probability, in order to control the *family-wise error rate* (FWER). Other corrections are also possible, such as the Sidák correction (Lehmann & Romano, 2006)  $\alpha = 1 - (1 - \alpha^*)^{0.5}$ . In Phase II, let  $X^*(t)$ ,  $t \in \mathcal{T}$ , denote a new observation of the functional quality characteristic. Then, the new estimated scores are calculated as  $\hat{\xi}_l^* = \int_{\mathcal{T}} \hat{\psi}_l(t) X^*(t) dt$ , for  $l = 1, \dots, L$ , where  $\{\hat{\psi}_l\}$  are the estimated FPCs. The

new realization of the  $T^2$  and  $SPE$  statistics are calculated as

$$T^{2*} = \sum_{l=1}^L \frac{\hat{\xi}_l^{*2}}{\hat{\lambda}_l} \tag{10}$$

and

$$SPE^* = \int_{\mathcal{T}} \left( X^*(t) - \hat{X}^{PC*}(t) \right)^2 dt, \tag{11}$$

where  $\hat{X}^{PC*}(t) = \sum_{l=1}^L \hat{\xi}_l^* \hat{\psi}_l(t)$ , with  $t \in \mathcal{T}$ . An OC signal is issued if at least one of  $T^{2*}$

(Equation 10) and  $SPE^*$  (Equation 11) violates the control limits. The larger the portion of variability of the functional data explained by the first  $L$  FPCs retained into the FPCA model, the more coherent is the following interpretation of  $T^{2*}$  and  $SPE^*$  statistics. In fact, as the former is based on the first scores, we expect that the larger the value of  $T^{2*}$ , the larger the deviation in *magnitude* from the reference mean of the new functional observation. Accordingly, as the latter is based on the last FPCs (see (9)), we expect that a new profile with large  $SPE^*$ , which roughly monitors the *shape* of the current functional observation, exhibits non-negligible deviation in the covariance structure from that estimated on the reference data set.

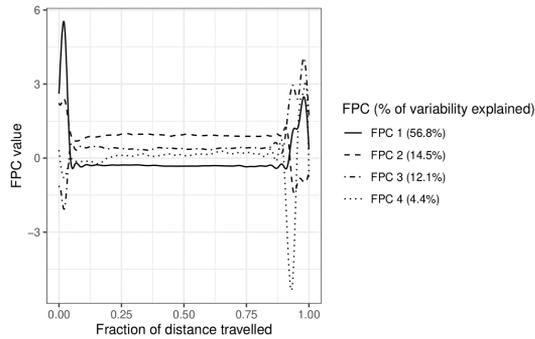
### 3 A real-case study in the shipping industry

We illustrate the proposed monitoring procedure by means of a real-case study from the maritime field in monitoring CO<sub>2</sub> emissions during the navigation phase of a roll-on/roll-off passenger (Ro-Pax) cruise ship. The data analyzed in this paper are a courtesy of the owner Grimaldi Group. Information about ports, name of the ship and CO<sub>2</sub> emissions are omitted for confidentiality reasons. Two years of data are available with five-minute frequency. In the proposed application, we focus only on one route sailed by the ship to link two ports. The available data set contains the discrete values for 194 voyages of CO<sub>2</sub> emissions due to propulsion, which is the functional quality characteristic to be monitored at the end of each voyage. The functional domain for each voyage is the fraction of total distance travelled from the beginning of the voyage, which is a dimensionless quantity between zero and one.

The first 146 voyages are used as training data set to perform FPCA and estimate control chart limits. Then, the following 48 voyages are sequentially numbered using a voyage number (VN) and monitored as described in Section 2. Note that, since we focus on Phase II monitoring only, we do not report details about Phase I, which was devoted to filter out from the training data set data that do not reflect standard navigation conditions and thus may introduce bias in the estimation of model parameters and control chart limits.

For each voyage, functional data are drawn from discrete observations by means of 50 B-spline basis functions and equally spaced knots. Functional data are smoothed by penalizing the integrated squared second derivative and by choosing the smoothing parameter through GCV criterion, as discussed in Section 2.1.

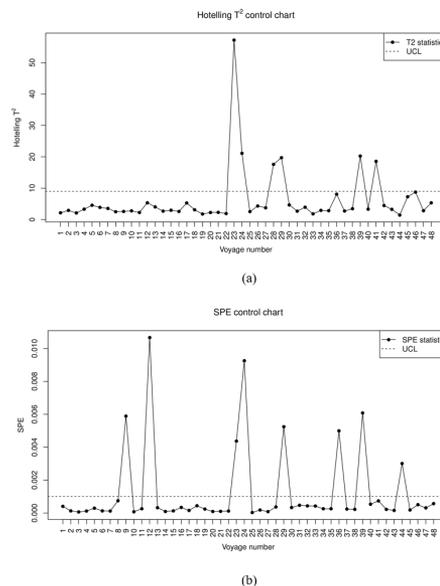
Then, FPCA is applied on the training data set. Figure 2 reports the first four FPCs and the percentage of variability explained by each FPC.



**Figure 2.** First four functional principal components (FPCs). For each FPC, the percentage of variability explained is reported in the legend.

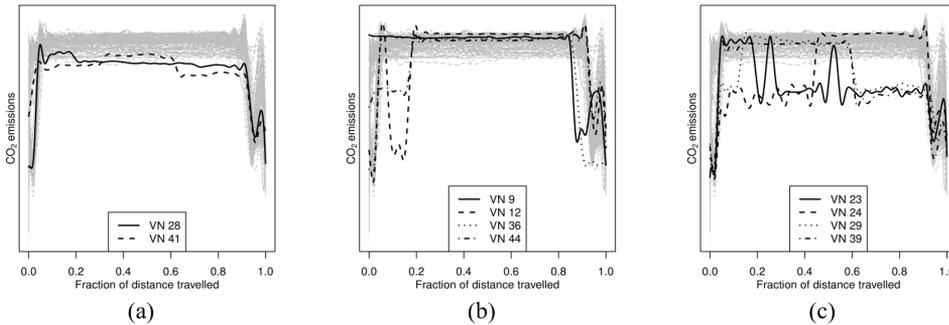
As an example, note that the first component explains 56.8% of the total variability in the data, which is mainly attributed to the beginning and the end parts of the voyage. The second component explains 14.5% of the variability, which is attributed to the average value of CO<sub>2</sub> emissions alongside the voyage, whereas the third component explains 12.1% of the variability and attributes the main weight to the end part of the voyage. Starting from the fourth component, the explained variability is less than 5% and interpretation becomes cumbersome.

For the reasons discussed above, in this application it is convenient to retain the first  $L=3$  FPCs, which explain together 83.4% of the total variability in the data, to approximate functional data and use the corresponding scores to calculate the statistic, while the residual functions  $X^*(t) - \hat{X}^{PC*}(t)$  can be used to calculate the *SPE* statistic as in Equation 11. Once control limits are estimated based on the  $T^2$  and *SPE* statistics calculated on the training data set, it is possible to use control charts to monitor new voyages. Figure 3 shows the  $T^2$  and *SPE* control charts used for Phase II monitoring.



**Figure 3.** (a)  $T^2$  and (b) *SPE* phase II control charts. In each control chart, points joint by solid line indicate monitoring statistic values at each voyage, while dashed lines indicate upper control limit (UCL), at  $\alpha = 0.05$ .

Several scenarios are possible, and it is interesting to notice that the use of both control charts supports the interpretation of the type of anomalies encountered. In Figure 4 we report OC profiles against those of the training data set, which are plotted for ease of comparison as grey lines.



**Figure 4.** OC CO<sub>2</sub> emission profiles (black lines) are superimposed on phase I reference ones (grey lines) and grouped by OC in (a) in  $T^2$  control chart only; (b)  $SPE$  control chart only; (c) both  $T^2$  and  $SPE$  control charts.

In particular, note that VN 28 and 41 are OC in the  $T^2$  control chart only (Figure 4a), VN 9, 12, 36, and 44 are OC in the  $SPE$  control chart only (Figure 4b), whereas VN 23, 24, 29, and 39 are OC in both control charts (Figure 4c).

In Figure 4a, profiles of VN 28 and 41 show a clear deviation in *magnitude* only, that is the CO<sub>2</sub> emissions plot below the average. Strictly speaking, it is worth noting that lower CO<sub>2</sub> emissions, which are in fact desired, often are trivially associated to voyages sailed at lower-than-usual speed over ground that, in turn, imply other types of undesired costs for the shipping company due to arrival delay. Therefore, it is crucial that the proposed control charting procedure can signal these profiles. In Figure 4b, VN 9, 12, 36, and 44 show that, during most of the voyage, the CO<sub>2</sub> emissions were not particularly different from the reference profiles in the training data set. However, some non-negligible slowdowns are highlighted in brief parts of these voyages. The most important deviation from reference behavior occurs in fact at the beginning of VN 12, which shows the largest  $SPE$  in Figure 4b. The other voyages seem to postpone the acceleration phase at the beginning of the voyage, or to anticipate the slowdown at the end of the voyage, then they show lower amounts of CO<sub>2</sub> emissions. More generally, these voyages show a different *shape* from standard Phase I profiles. Finally, In Figure 4c, with respect to the other voyages we discussed above, the CO<sub>2</sub> emissions for VN 23, 24, 29, and 39 show much larger deviations from the reference profiles in terms of both *magnitude* and *shape*. All these voyages have a lower-than-usual amount of CO<sub>2</sub> emissions for most of the voyage (deviation in *magnitude*), but also show sudden accelerations/slowdowns (deviation in *shape*). This is plausibly due to bad weather conditions, which has forced the ship to sail at an unusual navigation speed profile.

Even though no action could be taken in most of the cases discussed above, we believe this real-case study motivates the use of a functional data approach that allows the definition of two non-trivial monitoring statistics. By plotting profiles related to out-of-control signals, one can identify the domain portions where anomalies have occurred. However, in this application, we are not directly interested in a real-time feedback control and immediate actions during a voyage. Instead, the proposed control charts focus on tracking automatically (and possibly the whole fleet) for the following out-of-control signals, or patterns and trends which may identify malfunctioning in the

engines, the need for hull cleaning, or for any other energy efficiency initiative. Explicitly note that, in this application, CO<sub>2</sub> emission profile at each voyage is the actual key performance indicator that the shipping company is interested in. This means that monitoring through the ship speed over ground alone, which may appear as an easier job than monitoring CO<sub>2</sub> emissions directly, might not be adequate. While, theoretically, it is trivially known that the larger the speed, the larger the CO<sub>2</sub> emissions, the true relationship between observational data is affected also by many covariates, such as weather conditions, displacement, trim. This should be properly accounted for in a complex model, which however is beyond the scope of this application. A different research question can be for example: “*given the value of the covariates, is the quality characteristic as expected?*”, as faced by Capezza et al. (2020) and Centofanti et al. (2020).

## 4 Conclusions

In this paper, we showed benefits and practical applicability of a functional data analysis approach in real-world case studies, with a very transparent set of steps. The two most important advantages over the classical multivariate approach are the possibility (i) of analyzing data theoretically defined over a continuum domain even when, over different observations, the discrete measurements are unequally spaced and may be more numerous than functional data observations; (ii) of assuming smoothness, i.e., that data points can borrow information from their neighbors. The smoothness assumption (ii) is very reasonable in most of the practical cases and supports model interpretability as it implies e.g., smooth eigenfunctions in the functional principal component analysis.

The proposed functional control charting scheme is shown to be able to monitor CO<sub>2</sub> emissions in practice from real navigation data and to support the detection and interpretation of anomalous voyage profiles. Different scenarios have validated the capability of distinguishing, with respect to the reference profiles characterizing the standard operating conditions, the type of deviation based on which control charts has issued the out of control signal.

## References

- Bersimis, S., Sgora, A., & Psarakis, S. (2018). The application of multivariate statistical process monitoring in non-industrial processes. *Quality Technology & Quantitative Management*, 15(4), 526-549. <http://dx.doi.org/10.1080/16843703.2016.1226711>.
- 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. <http://dx.doi.org/10.5957/jsr.2015.59.3.162>.
- Bosq, D. (2012). *Linear processes in function spaces: theory and applications* (Lecture Notes in Statistics). New York: Springer.
- Capezza, C., Centofanti, F., Lepore, A., Menafoglio, A., Palumbo, B., & Vantini, S. (2021). *Funcharts: functional control charts. R package version 1.0.0*. Vienna: R Foundation for Statistical Computing. Retrieved in 2020, September 28, from <https://CRAN.R-project.org/package=funcharts>
- 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.  
<http://dx.doi.org/10.1016/j.trd.2018.11.009>.
- Capezza, C., Lepore, A., Menafoglio, A., Palumbo, B., & Vantini, S. (2020). Control charts for monitoring ship operating conditions and CO emissions based on scalar-on-function regression. *Applied Stochastic Models in Business and Industry*, 36(3), 477-500.  
<http://dx.doi.org/10.1002/asmb.2507>.
- Cardot, H., Ferraty, F., & Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13, 571-591.
- Centofanti, F., Lepore, A., Menafoglio, A., Palumbo, B., & Vantini, S. (2020). Functional regression control chart. *Technometrics*, 1-14.  
<http://dx.doi.org/10.1080/00401706.2020.1753581>.
- 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. <http://dx.doi.org/10.1002/qre.878>.
- 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. <http://dx.doi.org/10.1080/00207540802662888>.
- De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C., & De Boor, C. (1978). *A practical guide to splines* (Vol. 27). New York: Springer-Verlag. <http://dx.doi.org/10.1007/978-1-4612-6333-3>.
- 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. <http://dx.doi.org/10.1002/qre.1864>.
- Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. Boca Raton: CRC Press. <http://dx.doi.org/10.1201/9781482273144>.
- Grasso, M., Colosimo, B. M., & Tsung, F. (2017). A phase I multi-modelling approach for profile monitoring of signal data. *International Journal of Production Research*, 55(15), 4354-4377. <http://dx.doi.org/10.1080/00207543.2016.1251626>.
- Grasso, M., Menafoglio, A., Colosimo, B. M., & Secchi, P. (2016). Using curve-registration information for profile monitoring. *Journal of Quality Technology*, 48(2), 99-127. <http://dx.doi.org/10.1080/00224065.2016.11918154>.
- Green, P. J., & Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Boca Raton: Chapman & Hall/CRC. <http://dx.doi.org/10.1201/b15710>.
- Gu, C. (2013). *Smoothing spline ANOVA models* (Vol. 297). New York: Springer Science & Business Media. <http://dx.doi.org/10.1007/978-1-4614-5369-7>.
- Horváth, L., & Kokoszka, P. (2012). *Inference for functional data with applications*. New York: Springer Science & Business Media. <http://dx.doi.org/10.1007/978-1-4614-3655-3>.
- Hsing, T., & Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. West Sussex: John Wiley & Sons. <http://dx.doi.org/10.1002/9781118762547>.
- Jin, J., & Shi, J. (1999). Feature-preserving data compression of stamping tonnage information using wavelets. *Technometrics*, 41(4), 327-339. <http://dx.doi.org/10.1080/00401706.1999.10485932>.
- Jolliffe, I. (2011). *Principal component analysis*. Cham: Springer.

- Kokoszka, P., & Reimherr, M. (2017). *Introduction to functional data analysis*. Boca Raton: CRC Press. <http://dx.doi.org/10.1201/9781315117416>.
- Lehmann, E. L., & Romano, J. P. (2006). *Testing statistical hypotheses*. New York: Springer Science & Business Media.
- Lepore, A., Palumbo, B., & Capezza, C. (2019). Orthogonal LS-PLS approach to ship fuel-speed curves for supporting decisions based on operational data. *Quality Engineering*, 31(3), 386-400. <http://dx.doi.org/10.1080/08982112.2018.1537445>.
- Lowry, C. A., & Montgomery, D. C. (1995). A review of multivariate control charts. *IIE Transactions*, 27(6), 800-810. <http://dx.doi.org/10.1080/07408179508936797>.
- Mandel, B. (1969). The regression control chart. *Journal of Quality Technology*, 1(1), 1-9. <http://dx.doi.org/10.1080/00224065.1969.11980341>.
- Menafoglio, A., Grasso, M., Secchi, P., & Colosimo, B. M. (2018). Profile monitoring of probability density functions via simplicial functional PCA with application to image data. *Technometrics*, 60(4), 497-510. <http://dx.doi.org/10.1080/00401706.2018.1437473>.
- Montgomery, D. C. (2007). *Introduction to statistical quality control*. Hoboken: John Wiley & Sons.
- Nomikos, P., & MacGregor, J. F. (1995). Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37(1), 41-59. <http://dx.doi.org/10.1080/00401706.1995.10485888>.
- Noorossana, R., Saghaei, A., & Amiri, A. (2012). *Statistical analysis of profile monitoring*. Hoboken: John Wiley & Sons.
- Pini, A., & Vantini, S. (2017). Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, 29(2), 407-424. <http://dx.doi.org/10.1080/10485252.2017.1306627>.
- Pini, A., Vantini, S., Colosimo, B. M., & Grasso, M. (2018). Domain-selective functional analysis of variance for supervised statistical profile monitoring of signal data. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, 67(1), 55-81. <http://dx.doi.org/10.1111/rssc.12218>.
- R Core Team. (2020). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
- Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis*. New York: Wiley. <http://dx.doi.org/10.1007/b98888>.
- Shang, H. L. (2014). A survey of functional principal component analysis. *AStA. Advances in Statistical Analysis*, 98(2), 121-142. <http://dx.doi.org/10.1007/s10182-013-0213-1>.
- Wahba, G. (1990). *Spline models for observational data* (Vol. 59). Philadelphia: Society for Industrial and Applied Mathematics. <http://dx.doi.org/10.1137/1.9781611970128>.
- 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), 309-320. <http://dx.doi.org/10.1080/00224065.2004.11980276>.