Working paper

2023-04

Statistics and Econometrics
ISSN 2387-0303

Modelling physical activity profiles in COPD patients: a new approach to variable-domain functional regression models

Pavel Hernández-Amaro, María Durbán, M. Carmen Aguilera-Morillo,
Cristobal Esteban Gonzalez, Inma Arostegui

uc3m | Universidad Carlos III de Madrid

Modelling physical activity profiles in COPD patients: a new approach to variable-domain functional regression models

Pavel Hernández-Amaro[1*], María Durbán [1], M. Carmen Aguilera-Morillo[2],

Cristobal Esteban Gonzalez[3], Inma Arostegui[4]

[1] *Universidad Carlos III de Madrid,* [2] *Universitat Politècnica de València,* [3] *Osakidetza Basque Health Service,* [4] *University of the Basque Country UPV/EHU*

E-mail address for correspondence: pahernan@est-econ.uc3m.es

Summary

Motivated by the increasingly common technology for collecting data, like cellphones, smartwatches, etc, functional data analysis has been intensively studied in recent decades, and along with it, functional regression models. However, the majority of functional data methods in general and functional regression models, in particular, are based on the fact that the observed data present the same domain. When the data have variable domain it needs to be aligned or registered in order to be fitted with the usual modeling techniques adding computational burden. To avoid this, a model that contemplates the variable domain features of the data is needed, but this type of models are scarce and its estimation method presents some limitations. In this article, we propose a new scalar-on-function regression model for variable domain functional data that eludes the need for alignment and a new estimation methodology that we extend to other variable domain regression models.

*To whom correspondence should be addressed.

The efficiency of our proposal is demonstrated in a simulation study where we compare the obtained results with other existing methodologies. We illustrate our method with the analysis of data from the telEPOC study (Esteban *and others*, 2016).

Key words: Variable domain functional data; B-splines; Mixed models; COPD.

1. Introduction

Functional data analysis is a very active area of research and one of the fastest growing fields of statistical analysis. Prove of this are the great number of books and papers published in the past two decades, see for example Ramsay and Silverman (2005); Horváth and Kokoszka (2012), as well as the references therein. The interest in this area has been fueled by the technological advances that provide increasingly complex and high-dimensional data with functional nature.

In practice, functional data are usually found as discrete and often noisy observations of the true underlying function, measured at different locations in time, space, or other continuums. The domain where the data is observed is usually assumed to be the same across observations. Functional data where the domain is not the same for all the observations is named variable-domain functional data. This type of data can be found in many data sets and a variety of research fields like biology (Kulbaba *and others*, 2017), agriculture (Panayi *and others*, 2017), medicine (Gaynanova *and others*, 2022), among others.

Our particular motivation is the telEPOC study (Esteban *and others*, 2016). In this study a wide range of data from 119 patients suffering from Chronic Obstructive Pulmonary Disease (COPD) is collected, being the most important one the physical activity performed by each patient. The physical activity is measured as daily steps, with the particularity the number of days where steps are collected is different from patient to patient, varying from 64 days up to 1287 days, as shown in Figure 1. There are two major reasons for this: (i) the exact day of sign-up in the study is different from patient to patient (the inclusion dates go from 31-05-2010 to 07-12-2013) and (ii) the time each patient spends in the study is different, being one of the causes

that 21 patients died during the study period. Therefore, we are in presence of a variable-domain functional dataset.

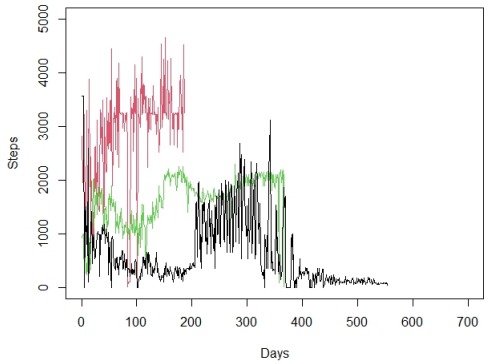

Fig. 1: Daily Steps of 3 different patients of the telEPOC study.

One of the goals of the telEPOC study is to determine how physical activity affects health in COPD patients. More precisely, the aim is to estimate the relationship between the number of hospitalizations due to COPD (scalar discrete r.v.) and physical activity (functional covariate). To accomplish this a scalar-on-function regression model should be required.

The scalar-on-function regression model was one of the first regression models extended to the case of functional data. The main theoretical aspects related to this model were studied in Cardot *and others* (1999) and in James (2002), in the more general framework of generalized linear models. This model has been widely used in the literature, leading to numerous applications and new methodological developments. Penalized versions of the functional generalized linear models can be seen in works, such as Cardot and Sarda (2005), Goldsmith *and others* (2011), Aguilera-Morillo *and others* (2013), among others.

This methodology is very useful for modeling functional data when the domain is constant across observations, but to deal with variable-domain functional data a previous transformation of the sample curves is needed. One of the most common choice is based on the registration of the sample curves to a common domain. This additional step presents some drawbacks in the case of

variable-domain functional data: it can add errors to the sample curves or lose some information given by the specific shape of the curves; the resulting estimation of the functional parameter will be more difficult or impossible to interpret since it will be a single curve estimated from the forced common domain; this registration procedure could be computational expensive in some cases. For some insights in registration of curves see Ramsay *and others* (2009).

Moreover, in a variable-domain dataset, the length of the sample paths is informative itself. Therefore, incorporating this information in the formulation of the functional regression model is essential for dealing with this type of data.

Is not until Gellar *and others* (2014) that variable domain functional data was successfully modeled while maintaining the variable domain features of the data. The variable domain functional regression model (VDFR) proposed in Gellar *and others* (2014) introduces the variable domain information directly into the design of the model by considering specific domains in the integration limits and a two-dimensional functional coefficient, meaning that the functional data will have different influence in the response variable according to its specific domain. The estimation method proposed by the authors is based on the basis representation of the functional parameter and the use of the well known relation of functional models with mixed model representation. However this approach presents some limitations. On the one hand, the basis representation of the functional covariate is not considered, leaving out the possibility of recover the true functional form of the data and not filtering the possible noise commonly present in the discrete observation of the sample curves. Additionally, this approach limits the use of thin-plate spline functions in the basis representation of the functional coefficient. As a consequence, only isotropic penalization can be used, which forces the use of the same degree of smoothness in both dimensions of the functional coefficient.

The goal of this work is to extend the VDFR model to what the authors call "fully functional variable-domain functional regression" (FF-VDFR). Both approaches (VDFR and FF-VDFR) assume that the predictor is a functional variable but they are conceptually different and hence

present some differences in the estimation procedure. The fully functional approach assumes the sample paths (raw data) belong to a finite-dimensional space spanned by a basis of functions in order to correctly filter the inherent noise on these. Moreover, a novel and flexible way to estimate the basis coefficients of the functional parameter is proposed, permitting the use of any kind of basis and considering an anisotropic penalty. Finally, the FF-VDFR model uses the connection to mixed model framework to gain computational efficiency by introducing the separation of overlapping penalties (SOP) algorithm, developed in Rodríguez-Álvarez *and others* (2019).

The rest of the document is organized as follows. In section 2 we present the FF-VDFR model with the corresponding estimation procedure. In section 3 we present a simulation study to evaluate the performance of the proposed method in compared to the existing methods. In section 4 we show the results of applying the proposed methodology to the telEPOC study. Finally we conclude with a discussion in Section 5.

2. Variable-domain functional regression

The main goal of telEPOC study is to explore the performance of the physical activity, measured as number of daily steps (X), among patients and to study its relation with the number of hospitalizations (Y) due to COPD. Then, we are focused on a regression problem where the response variable Y is a scalar and the predictor X is a function whose values varying over a continuous domain of varying length among subjects, i.e., $\{X_i(t) : t \in [d_i, T_i], \quad i = 1, \ldots, N.\}$. Essentially, this means that every curve can have observations points that fall in different domains. In our case study, the data can be left-aligned and ordered without affecting the information of the results, having for all the data the same initial point $t \in [0, T_i]$ and $T_i \leqslant T_{i+1}\, \forall i$.

In most of functional data problems, the functional predictor is assumed to have a common domain for all sample units. In this context, the sample information is usually given by $\{Y_i, X_i(t), C_i\}$, $i = 1, \ldots, N$, where C_i is a vector of non-functional covariates, Y_i is a scalar outcome following an exponential family distribution with mean μ_i and $\{X_i(t) : t \in T\}$ is the

functional predictor. From this information, the functional generalized linear model is given by

$$\eta_i = g(\mu_i) = \alpha + C_i\gamma + \int_0^T X_i(t)\beta(t)\,\mathrm{d}t, \tag{2.1}$$

with $g(\cdot)$ being the corresponding link function.

This model has been widely studied by authors such as Cardot *and others* (1999); James (2002); Cardot and Sarda (2005); Ramsay and Silverman (2005); Goldsmith *and others* (2011); Aguilera-Morillo *and others* (2013), between others, arising from the classical formulation to penalized versions that provide a smooth estimation of the functional parameter $\beta(t)$. The functional parameter represents the optimal way of weighting each sample curve across the full domain. Because of this, when the sample curves have different domains, the estimation of model 2.1 results in poor estimates of the functional parameter, which makes difficult to interpret the relationship between predictors and response variable.

As solution, Gellar *and others* (2014) proposed a formulation of the classical model 2.1 to deal with variable-domain functional data:

$$\eta_i = g(\mu_i) = \alpha + C_i\gamma + \frac{1}{T_i}\int_0^{T_i} X_i(t)\beta(t,T_i)\,\mathrm{d}t, \quad t \in [0,T_i], \tag{2.2}$$

where the univariate coefficient function $\beta(t)$ is now replaced by the bivariate coefficient function $\beta(t,T)$, that now depends on the time instant t and the data domain T. This functional parameter is now a surface and the curves obtained by fixing the variable $T = T_i$ represents the optimal function for $X_i(t)$ to express its contribution over $g(\mu_i)$. Moreover, the integration limits, previously fixed to be from 0 to T, are now subject-specific, avoiding then the necessity of a previous curve registration.

In practice, an additional problem of the functional regression is that we only have discrete observations x_{ik} of each sample curve $x_i(t)$ at a finite set of points $\{t_{ik} : k = 0,\ldots,m_i\}$. The authors deal with this problem by assuming the basis representation of the functional parameter in therms of thin-plate regression spline basis. Moreover, in order to get smooth estimations in both t and T_i directions, the basis coefficients of $\beta(t,T_i)$ are penalized with a second-order derivative

penalty. The estimation methodology proposed by Gellar *and others* (2014) falls within what we call "partially functional approach", since it does not take into consideration the functional form of the variable $X(t)$, working directly on the row data matrix (sample curves at the observation points), leaving out the possibility of recovering the true functional form of the data and not filtering the noise commonly present in its discrete observations.

Additionally, this approach limits the options of basis selection when making the basis representation of the functional coefficient, more particularly, leaves out the choice of B-spline basis as an optimal selection and then only isotropic penalization can be used. This might result in biased estimates of the functional coefficient, as it will be shown in the simulation section.

3. Fully functional Variable-domain functional regression

Partially functional approaches work on the discrete observations x_{ij} of each sample curve $X_i(t)$ at a set of points $\{t_{ik}, \quad k = 0, \dots T_i\}$. However, in practice it is very common to find functional datasets observed with error or noise. In that sense, a fully functional approach will perform a pre-smoothing of the sample curves, recovering the smooth functional form of the data by means of a basis representation of the sample curves. A review on the different ways to estimate the basis coefficients as well as the different penalties used and their performance is shown in Aguilera and Aguilera-Morillo (2013).

The main advantages of the proposed fully functional variable-domain functional regression model with respect to the approach described in Section 2 are the following: it allows to filter the inherent noise in the discrete observations of the sample curves, offering a better performance in the context of sparse data and partially observed data, because it makes an approximation of the missing data; it considers a more flexible representation of the functional parameter permitting that any basis can be chosen and then, anisotropic penalties can be used. Notice that when the discrete observations can be assumed free of error and the number of observations is sufficiently large, these two approaches perform very similarly.

3.1 Model formulation in terms of basis functions

Let Y be the scalar response variable and $X(t)$ the functional predictor. Let us consider $X(t)$ is a second order continuous-time stochastic process, with sample functions $\{X_i(t) : \quad t \in [d_i, T_i], \quad i = 1, \ldots, N\}$ in the Hilbert space $\mathscr{H}_1 = L^2(T)$ of integrable square functions, with the usual inner product.

Let us assume the basis representation of the sample curves and the functional coefficient as follows:

$$X_i(t) = \sum_{j=1}^{p_i} a_{ij}\phi_j(t) = \boldsymbol{\phi}_i'(t)\boldsymbol{a}_i,$$
$$\beta(t,T) = \sum_{l=1}^{q}\sum_{k=1}^{r} b_{lk}\varphi_l(t)\psi_k(T) = \boldsymbol{M}(t,T)\boldsymbol{b},$$

where $\boldsymbol{\phi}_i(t) = (\phi_1(t), \phi_2(t), \ldots, \phi_{p_i}(t))'$ and $\boldsymbol{M}(t,T)$ are the basis used in the representation of the functional data and the functional coefficient and $\boldsymbol{a}_i$ and $\boldsymbol{b}$ their respective basis coefficients. Notice that $\boldsymbol{M}(t,T)$ is a bivariate basis function resulting from the tensor product of $\boldsymbol{\varphi}(t)$ and $\boldsymbol{\psi}(T)$, with $\boldsymbol{\psi}(t) = (\psi_1(t), \psi_2(t), \ldots, \psi_r(t))'$ and $\boldsymbol{\varphi}(t) = (\varphi_1(t), \varphi_2(t), \ldots, \varphi_q(t))'$.

Here p_i, q and r are the respective number of basis of $\boldsymbol{\phi}_i(t), \boldsymbol{\varphi}(t)$ and $\boldsymbol{\psi}(t)$. For simplicity, hereinafter the same number of basis (p) is considered for the basis representations of all sample curves, i.e., $p_i = p$ and $\boldsymbol{\phi}_i(t) = \boldsymbol{\phi}(t)\ \forall i = 1, \ldots, N$, but this can be easily relaxed.

The choice of the basis is important. This decision is often data driven: if data have periodic trends a Fourier basis can be used; if data present a strong locally behavior and its derivatives are not of interest, wavelets basis are the common choice. In this paper B-splines basis (De Boor, 2001) have been considered, which is the common choice when the underlying signal is assumed to be smooth and their derivatives up to a certain order are needed.

By assuming the basis representation of both, sample curves and functional coefficient, the model in 2.2 turns into the following multivariate regression model:

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \boldsymbol{C\gamma} + \frac{1}{\boldsymbol{T}}\int_0^{\boldsymbol{T}} X(t)\beta(t,T)\,\mathrm{d}t = \boldsymbol{\alpha} + \boldsymbol{C\gamma} + \boldsymbol{A\Psi b} = \boldsymbol{B\theta}, \tag{3.3}$$

where $\boldsymbol{T}$ represents the vector considering the length of all curves, so for each sample curve the integration limits are different.

Notice that the matrix of coefficients $\boldsymbol{A}$ is a block diagonal matrix, where the i-th block of the diagonal is the estimated vector of basis coefficients $\boldsymbol{a}_i'$. For the estimation of these basis coefficients, and following the results in Aguilera and Aguilera-Morillo (2013), we use penalized least squares with a discrete penalty based on the second order differences between adjacent coefficients of the B-splines (Eilers and Marx, 1996). Finally the matrix of inner products $\boldsymbol{\Psi}$ is a block column matrix of weighted inner products with the i-th block being a weighted inner product between the basis $\boldsymbol{\phi}(t)$ and $\boldsymbol{M}(t,T_i)$: $\boldsymbol{\Psi}_{Np\times qr}=(\boldsymbol{\Psi}_1,\ldots,\boldsymbol{\Psi}_N)'$ where $\boldsymbol{\Psi}_i=\frac{1}{T_i}\langle\boldsymbol{\phi}(t),\boldsymbol{M}(t,T_i)\rangle$,

$$\boldsymbol{A}=\begin{pmatrix}\boldsymbol{a}_1' & 0 & \ldots & 0\\ 0 & \boldsymbol{a}_2' & 0 & \ldots\\ \ldots & \ldots & \ldots & \ldots\\ 0 & 0 & \ldots & \boldsymbol{a}_N'\end{pmatrix}_{N\times Np}\qquad \boldsymbol{\Psi}=\begin{pmatrix}\frac{1}{\boldsymbol{T}_1}\int_1^{T_1}\boldsymbol{\phi}(t)\boldsymbol{M}(t,T_1)\,\mathrm{d}t\\ \frac{1}{\boldsymbol{T}_2}\int_1^{T_2}\boldsymbol{\phi}(t)\boldsymbol{M}(t,T_2)\,\mathrm{d}t\\ \vdots\\ \frac{1}{\boldsymbol{T}_N}\int_1^{T_N}\boldsymbol{\phi}(t)\boldsymbol{M}(t,T_N)\,\mathrm{d}t\end{pmatrix}_{Np\times qr}.$$

The elements of the inner products matrix are given by a new operation named partial inner product defined very recently in Masak *and others* (2022) and which is detailed below.

Proposition 1: Let $\mathscr{H}_1=L^2(T)$ and $\mathscr{H}_2=L^2(F)$ be two separable Hilbert spaces as in Section 1 with $F=\{T:T_{min}\leqslant T\leqslant T_{max}\}$ being the space corresponding to all the different values of the data domains.

Let $\mathscr{H}=\mathscr{H}_1\otimes\mathscr{H}_2$ where $\otimes$ represents the tensor product and let $f(t)$, $u(T)$ and $h(t,T)$ be functions in $\mathscr{H}_1$, $\mathscr{H}_2$ and $\mathscr{H}$, respectively.

Then the partial inner products are two unique bi-linear operators $K_1:\mathscr{H}\times\mathscr{H}_1\to\mathscr{H}_2$ and $K_2:\mathscr{H}\times\mathscr{H}_2\to\mathscr{H}_1$ defined by:

$$K_1(T)_{h,f}=\int_T f(t)h(t,T)\,\mathrm{d}t$$
$$K_2(t)_{h,u}=\int_F u(T)h(t,T)\,\mathrm{d}T.\blacksquare$$

Finally, the elements of the new matrix of inner products are given by $\boldsymbol{\Psi}_{Np\times qr}=(\boldsymbol{\Psi}_1,\ldots,\boldsymbol{\Psi}_N)'$ are $\boldsymbol{\Psi}_i=\dfrac{1}{T_i}K_1(T_i)_{M,\phi}=\dfrac{1}{T_i}\displaystyle\int_{T_i}\boldsymbol{\phi}(t)\boldsymbol{M}(t,T_i)\,\mathrm{d}t$.

Numerically we approximate these integrals by the Composite Simpson method. The key difficulty is to perform the integration only in the t dimension while maintaining the proper two dimensional structure of the basis $M(t, T_i)$. In order to overcome this problem, for each iteration of the integration, a matrix $\boldsymbol{M}_i$ is obtained by performing the Kronecker product of two matrices $\boldsymbol{\varphi}_i$ and $\boldsymbol{\psi}_i$. The matrix $\boldsymbol{\psi}_i$ is the result of evaluating the basis $\boldsymbol{\psi}(T)$ in the corresponding domain T_i, i.e., $\boldsymbol{\psi}_i = \boldsymbol{\psi}(T_i)$. The matrix $\boldsymbol{\varphi}_i$ is the result of evaluate the basis $\boldsymbol{\varphi}(t)$ in a set of points determined by the integration method, this set of points change according with the domain of every curve. The basis $\boldsymbol{\phi}(t)$ is evaluated in the same set of points as the basis $\boldsymbol{\varphi}(t)$ in every iteration resulting in a matrix $\boldsymbol{\phi}_i$. Finally when the matrix $\boldsymbol{M}_i$ is recalculated, the product between $\boldsymbol{M}_i$ and $\boldsymbol{\phi}_i$ is performed.

Notice that the matrix $\boldsymbol{\psi}_i$ is the i-th row of a more general matrix $(\boldsymbol{\psi})_{N\times r}$, associated to all the different domains present in the data: $\boldsymbol{T} = [T_1, \ldots, T_N]$ and then, for subject i, we select the corresponding row. Two or more different sample curves can have the same domain; in this case the corresponding row of the matrix $\boldsymbol{\psi}$ will be the same for all of them.

We use B-splines for all our basis representations because of their desirable properties, but is not a restriction.

3.2 Model estimation through a mixed model representation

The multivariate regression model (3.3) falls into the category of generalized linear models and therefore the maximum likelihood method is used in order to estimate the model parameters. In our motivational example the response variable follows a Poisson distribution with likelihood:

$$L(\boldsymbol{\theta}, \boldsymbol{y}) = \sum_{i=1}^{N} y_i \eta_i - exp\left\{\sum_{i=1}^{N} \eta_i\right\}.$$

Since the functional coefficient has been represented using a B-spline basis, the smoothness of the resulting estimated coefficient is determined by the basis dimension. To avoid the problem

of choosing the optimal number of basis functions we follow the penalized likelihood approach by Eilers and Marx (1996) with the final penalized likelihood equation:

$$L_p(\boldsymbol{\theta}, \boldsymbol{y}) = L(\boldsymbol{\theta}, \boldsymbol{y}) - \frac{1}{2}\boldsymbol{\theta}'\boldsymbol{P}\boldsymbol{\theta},$$

where $L(\boldsymbol{\theta}, \boldsymbol{y})$ is the likelihood of $\boldsymbol{Y}$ and $\boldsymbol{P}$ is the penalty term. Penalties are, in general, based on derivatives of curves (Wood, 2017) or differences between adjacent B-splines coefficients (Eilers and Marx, 1996). We take here this second approach.

Considering that the functional parameter is two dimensional an anisotropic two dimensional penalization is used, allowing to control the smoothness of the functional coefficient independently for each dimension. The penalization added is:

$$\boldsymbol{P} = \lambda_t(\boldsymbol{I}_r \times \boldsymbol{D}_t'\boldsymbol{D}_t') + \lambda_T(\boldsymbol{I}_q \times \boldsymbol{D}_T'\boldsymbol{D}_T), \tag{3.4}$$

where the matrices $\boldsymbol{D}_t$ and $\boldsymbol{D}_T$ are second order differences matrices, where $\times$ represents the Kronecker product.

This penalized approach make the choice of the number of basis not relevant (provided that the size of the basis is large enough), controlling the smoothness through the smoothing parameters λ_t and λ_T.

Finally, we use the mixed model reparametrization of a penalized spline to estimate the parameters of the FF-VDFR model. This transformation allows the estimation of all parameters in the model, including the smoothing parameters, simultaneously. A brief description of this reparametrization is done next to help the reader understand the used methodology. For a more detailed insight into the mixed model reparametrization of a penalized spline when functional data do not present variable domain see Lee (2010). Our aim is to transform

$$\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\theta} \Rightarrow \boldsymbol{X}\boldsymbol{\nu} + \boldsymbol{Z}\boldsymbol{\delta}, \quad \boldsymbol{\delta} \sim N(0, \boldsymbol{G}), \tag{3.5}$$

with $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the model matrices, $\boldsymbol{\nu}$ and $\boldsymbol{\delta}$ are the fixed and random effects respectively, and $\boldsymbol{G}$ is the variance-covariance matrix of the random effects which depend on two variance components τ_t^2 and τ_T^2. This reparametrization is done through a transformation matrix $\boldsymbol{T}$ based on the SVD factorization of the product of the differences matrices $\boldsymbol{D}_i'\boldsymbol{D}_i$. Let

$$\boldsymbol{D}_i'\boldsymbol{D}_i = [\boldsymbol{U}_{in}|\boldsymbol{U}_{is}] \begin{bmatrix} \boldsymbol{0}_2 & \\ & \tilde{\boldsymbol{\Sigma}}_i \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_{in}' \\ \boldsymbol{U}_{is}' \end{bmatrix}$$

be the SVD factorization of the matrix $\boldsymbol{D}_i'\boldsymbol{D}_i$, for $i = \{t, T\}$ where $\boldsymbol{U}_{in}$ and $\boldsymbol{U}_{is}$ are the eigenvectors associated with the zero and non-zero eigenvalues respectively. Then the transformation matrix $\boldsymbol{T}$ is define as:

$$\boldsymbol{T} = [\boldsymbol{T}_n|\boldsymbol{T}_s] = [\boldsymbol{U}_{Tn} \times \boldsymbol{U}_{tn}|\boldsymbol{U}_{Ts} \times \boldsymbol{U}_{tn} : \boldsymbol{U}_{Tn} \times \boldsymbol{U}_{ts} : \boldsymbol{U}_{Ts} \times \boldsymbol{U}_{ts}]\,.$$

Other options for the transformation matrix $\boldsymbol{T}$ are possible but the one proposed in this paper allows to recover the estimated original functional parameter $\widehat{\boldsymbol{\theta}}$ from the estimated mixed model coefficients thanks to the orthogonality property of the matrix $\boldsymbol{T}$ and, hence, recover the estimated functional coefficient $\widehat{\beta}(t, T)$. Using this transformation matrix the model is reparametrized as follows:

$$\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\theta} = \boldsymbol{B}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\nu} + \boldsymbol{Z}\boldsymbol{\delta},$$

where $\boldsymbol{BT} = [\boldsymbol{BT_n}|\boldsymbol{BT_s}] = [\boldsymbol{X}|\boldsymbol{Z}]$, $\boldsymbol{T}'\boldsymbol{\theta} = \boldsymbol{\omega}$ with $\boldsymbol{\omega}' = (\boldsymbol{\nu}', \boldsymbol{\delta}')$ and the variance-covariance matrix $\boldsymbol{G}$ is obtained from applying this transformation to the penalization used before, $\boldsymbol{G}^{-1} = \boldsymbol{T}'\boldsymbol{P}\boldsymbol{T}$ with

$$\boldsymbol{G}^{-1} = \begin{pmatrix} \frac{1}{\tau_T^2}\tilde{\boldsymbol{\Sigma}}_T \times \boldsymbol{I}_2 & & \\ & \frac{1}{\tau_t^2}\boldsymbol{I}_2 \times \tilde{\boldsymbol{\Sigma}}_t & \\ & & \frac{1}{\tau_T^2}\tilde{\boldsymbol{\Sigma}}_T \times \boldsymbol{I}_{q-2} + \frac{1}{\tau_t^2}\boldsymbol{I}_{r-2} \times \tilde{\boldsymbol{\Sigma}}_t \end{pmatrix}, \tag{3.6}$$

where for the variance components we have the relations $\tau_t^2 = \dfrac{1}{\lambda_t}$ and $\tau_T^2 = \dfrac{1}{\lambda_T}$ (Brumback *and others*, 1999)

Finally, penalized quasi-likelihood (Breslow and Clayton, 1993) is used to estimate the mixed

model coefficient. In order to speed up computations, the SOP algorithm has been applied (Rodríguez-Álvarez *and others*, 2019).

4. Simulation study

In order to evaluate the performance of the FF-VDFR model, a simulation study has been carried out in this section and the obtained results have been compared with the one offered by the VDFR and the usual scalar-on-function (SOF) regression models. For the SOF model, a previous registration of the curves was performed. The simulation scheme is inspired by the one performed in Gellar *and others* (2014).

4.1 Simulation scenarios

For simplicity, only models with one functional covariate and no non-functional covariates have been considered. In this study 100 data sets have been simulated for each combination of the following parameters in a total of $3 \times 2 \times 2 \times 2 \times 4 = 96$ different scenarios:

- Three sample sizes: $N = \{100, 200, 500\}$.
- Two different types of outcomes: continuous data and count data. In both cases the following linear predictor is used:

$$\eta_i = \frac{1}{T_i} \sum_{t=1}^{T_i} X_i(t)\beta(t, T_i), \quad t = 1, \ldots, T_i \leqslant 100,$$

  and $Y_i = \eta_i + \epsilon_i$, $\epsilon_i \sim N(0, 1)$ for the continuous outcome and $Y_i \sim \text{Poiss}(\mu_i)$ with $\mu_i = \exp(\eta_i)$ for the count data.
- Two different distribution for the data domain T_i: Uniform($T_i \sim U(10, 100)$) and Negative Binomial ($T_i \sim NegBin(1, p = 0, 04)$).

  The domain of the sample curves is set to ensure that every curve have a minimum of 10 and a maximum of 100 observations in both distribution settings (we considered 10 to be an

acceptable minimum to consider the observed points as observation of the true underlying functional data $X_i(t)$). When T_i is simulated using a negative binomial distribution we truncate the generated values to belong in the interval $(10, 100)$, when a generated value is lower than 10 (higher than 100) this is set by default as 10 (100).

- Two levels of noise for the true functional covariate.

  The true functional covariate $X_i(t)$ is simulated according to the following:

$$X_i(t) = u_i + \sum_{k=1}^{10} \left\{ v_{ik1} \cdot sen\left(\frac{2\pi k}{100} t\right) + v_{ik2} \cdot \cos\left(\frac{2\pi k}{100} t\right) \right\} + \delta_i(t),$$

  with $u_i \sim N(0,1)$, $v_{ik1}, v_{ik2} \sim N(0, \frac{4}{k^2})$, $\delta_i(t) \sim N(0, \sigma_x)$, $t = 1, \ldots, T_i \leqslant 100$ and $\sigma_x = \{0, 1\}$. Here $\sigma_x = 0$ indicates that the true functional data has been considered as smooth curves and $\sigma_x = 1$ indicates that the true functional data has been considered as noisy curves.

- Four different possibilities for the functional coefficient $\beta(t, T)$ defined by

$$\beta_1(t, T_i) = 10\frac{t}{T_i} - 5 \qquad \beta_2(t, T_i) = \left(1 - \frac{2T_i}{T}\right) \times \left(5 - 40\left(\frac{t}{T_i} - 0.5\right)^2\right)$$
$$\beta_3(t, T_i) = 5 - 10\left(\frac{T_i - t}{T}\right) \quad \beta_4(t, T_i) = sen\left(\frac{2\pi T_i}{T}\right) \times \left(5 - 10\left(\frac{T_i - t}{T}\right)\right),$$

  where $T = \max\{T_1, \ldots, T_N\} = T_N$.

For the basis representation of the functional data in the FF-VDFR and SOF models 25 cubic B-splines have been used ($p = 25$). For the functional coefficient and in the case of the SOF model, 25 cubic B-splines were used while for the FF-VDFR model, 25 basis functions were considered for both marginal basis ($q = r = 25$), resulting in a bi-dimensional basis of size 625. The penalties used for the estimation of FF-VDFR and SOF models are based on a matrix of difference of order 2. The VDFR model considers a thin plate basis of size 89 for the functional coefficient, which is the maximum size allowed by the software, and an isotropic penalty based

on second-order derivatives.

All simulations were implemented in R Core Team (2013). The package **SOP** (Rodriguez-Alvarez and Oviedo de la Fuente, 2021) have been used for the mixed model reparametrization of the multivariate regression model and its estimation. The estimation of the SOF and VDFR models have been performed using the **refund** package (Goldsmith *and others*, 2021).

4.2 Performance criteria

We evaluate the performance of the above mentioned models with respect to two important aspects. The first one is the prediction ability. To this end, a cross-validation 10-fold approach has been carried out. The measure that we use for this prediction errors is the mean of the root mean square error (RMSE) calculated for every fold:

$$RMSE_j = \sqrt{\frac{\sum_{i=1}^{N_j}(Y_{ij} - \hat{Y}_{ij})^2}{N_j}}, \quad j = 1, \ldots, 10,$$

where Y_{ij} is the i-th response variable in the j-th fold, $\hat{Y}_{ij}$ its corresponding estimation and N_j is the number of responses in the j-th fold.

Another important aspect is the ability of correctly estimate the true functional coefficient $\beta(t, T)$. To this end the average mean square error (AMSE) is considered as follows:

$$AMSE^r = \frac{1}{T(T+1)} \sum_{k=10}^{T} \sum_{t=1}^{k} \left\{ \beta(t,k) - \hat{\beta}(t,k) \right\}^2,$$

where $T = \max\{T_1, \ldots, T_N\} = T_N$ and $\hat{\beta}(t, k)$ is the estimated functional coefficient.

Notice that with the SOF model it is not possible to calculate the AMSE since this models only considers one fixed domain for the functional coefficient.

4.3 Results

In this section we comment the results obtained in the simulation study for all the scenarios, but due to lack of space only tables and figures for the scenarios where T_i is generated from the negative binomial distribution and the response variable was simulated from a Poisson distribution are shown. The rest of the tables and figures can be found in the supplementary material.

Table 1 and Table 2 show the mean and standard deviation (in parenthesis) of the RMSE and AMSE, respectively, for all the possible true coefficient functions and when the true functional data are smooth or noisy. The lowest values are highlighted.

	N=100							
	Smooth data				Noisy data			
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
VDFR	1.343 (0.256)	1.145 (0.141)	3.143 (1.331)	1.924 (0.532)	1.364 (0.279)	1.155 (0.147)	2.793 (1.047)	1.991 (0.617)
SOF	5.35 (3.762)	1.813 (0.851)	3.196 (0.867)	2.798 (1.147)	5.128 (3.039)	1.803 (0.979)	5.199 (2.471)	2.543 (0.881)
FF-VDFR	**1.135 (0.134)**	**1.118 (0.12)**	**2.423 (0.898)**	**1.862 (0.568)**	**1.137 (0.136)**	**1.148 (0.142)**	**2.121 (0.667)**	**1.824 (0.553)**
	N=200							
	Smooth data				Noisy data			
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
VDFR	1.193 (0.104)	1.1 (0.076)	2.693 (0.59)	1.827 (0.341)	1.206 (0.117)	1.102 (0.077)	2.633 (0.768)	1.831 (0.38)
SOF	5.269 (2.865)	2.316 (1.146)	4.153 (0.88)	2.923 (0.828)	5.78 (3.422)	2.279 (1.123)	5.641 (2.315)	2.953 (0.974)
FF-VDFR	**1.12 (0.101)**	**1.092 (0.077)**	**2.311 (0.607)**	**1.69 (0.343)**	**1.111 (0.1)**	**1.096 (0.084)**	**1.858 (0.33)**	**1.699 (0.39)**
	N=500							
	Smooth data				Noisy data			
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
VDFR	1.129 (0.053)	1.087 (0.053)	2.721 (0.66)	1.675 (0.216)	1.133 (0.053)	**1.09 (0.053)**	2.33 (0.452)	**1.692 (0.243)**
SOF	3.569 (0.85)	2.616 (1.089)	3.848 (0.705)	3.035 (0.848)	3.523 (0.831)	2.487 (1.067)	9.386 (1.221)	3.143 (0.893)
FF-VDFR	**1.101 (0.056)**	**1.084 (0.056)**	**2.374 (0.508)**	**1.65 (0.288)**	**1.096 (0.05)**	1.091 (0.061)	**1.964 (0.255)**	1.695 (0.323)

Table 1: Mean (standard deviation) of 100 measures of RMSE for all the scenarios where the domain follows a negative binomial distribution and the response follows a Poisson distribution.

Regarding the RMSE we can see that the FF-VDFR model outperforms all others in all the scenarios shown in the table except in two cases, when the true functional coefficient is $\beta_2(t,T)$ and $\beta_4(t,T)$, the true functional data is noisy and the sample size is 500.

Notice that even in these scenarios the performance is very similar even when worse performance of the FF-VDFR model is expected. This is because the observed data correspond to the noisy curves regardless of whether the true functional data was smooth or noisy. For this reason, it is expected that filtering the noise by making a basis representation of the sampled curves will

	N=100							
	Smooth data				Noisy data			
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
VDFR	0,0187 (0,0086)	**0,0155 (0,0052)**	0,0118 (0,0068)	0,0196 (0,013)	0,0197 (0,009)	**0,0163 (0,0054)**	0,0118 (0,0053)	**0,0215 (0,0126)**
FF-VDFR	**0,0079 (0,0075)**	0,0176 (0,0074)	**0,008 (0,0126)**	**0,018 (0,0199)**	**0,008 (0,0073)**	0,0177 (0,0067)	**0,0083 (0,0103)**	0,0229 (0,0231)
	N=200							
	Smooth data				Noisy data			
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
VDFR	0,0246 (0,0039)	0,0217 (0,0057)	0,0142 (0,0051)	0,0254 (0,0101)	0,0243 (0,0044)	0,022 (0,0063)	0,013 (0,0043)	0,0264 (0,0107)
FF-VDFR	**0,0083 (0,0094)**	**0,0193 (0,0098)**	**0,0085 (0,0058)**	**0,021 (0,0084)**	**0,0096 (0,0088)**	**0,019 (0,0093)**	**0,0055 (0,0053)**	**0,0208 (0,0076)**
	N=500							
	Smooth data				Noisy data			
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
VDFR	0,0291 (0,0079)	0,0455 (0,0138)	0,0115 (0,0199)	0,0229 (0,0165)	0,0281 (0,0094)	0,0451 (0,0142)	0,0106 (0,01)	0,0208 (0,0189)
FF-VDFR	**0,0093 (0,0033)**	**0,0269 (0,0145)**	**0,01 (0,0046)**	**0,0195 (0,0101)**	**0,0093 (0,0039)**	**0,0228 (0,0141)**	**0,008 (0,0047)**	**0,017 (0,0077)**

Table 2: Mean (standard deviation) of 100 measures of AMSE when the domain follows a negative binomial distribution and the response follows a Poisson distribution.

improve the performance of the estimation when the true functional data is smooth but will be counterproductive when the true functional data is noisy.

Regarding the AMSE, the FF-VDFR model outperforms the VDFR model in all the scenarios but three, all corresponding to the smallest sample size $N = 100$, where both methods perform similarly. Notice that the differences in cases where the FF-VDFR model outperforms the VDFR model can be significant, for example in the scenarios where the true functional coefficient is $\beta_1(t, T)$.

These results by itself could be misleading, because they do not take into consideration the distribution of all the error measures. Figures 2 and 3 show the violin box-plots for the RMSE and AMSE measures, respectively, for the scenarios when the true functional coefficient is $\beta_3(t, T)$. In these figures, all values that fall outside the interval $(q_1 - 1,5 \cdot s \ ; \ q_3 + 1,5 \cdot s)$ have been excluded, with q_1 and q_3 being the first and third quartile, respectively, and s the standard deviation of the corresponding scenario.

The results shown in the figures reveal that the FF-VDFR model outperforms all other models for the RMSE and AMSE measurements, in terms of lower error and variability.

From the total of the 96 simulated scenarios, the FF-VDFR model outperformed all the others in 72 scenarios (75%) in terms of the RMSE. From the scenarios when the proposed methodology was not the best one, 55% corresponds with noisy true functional data from which

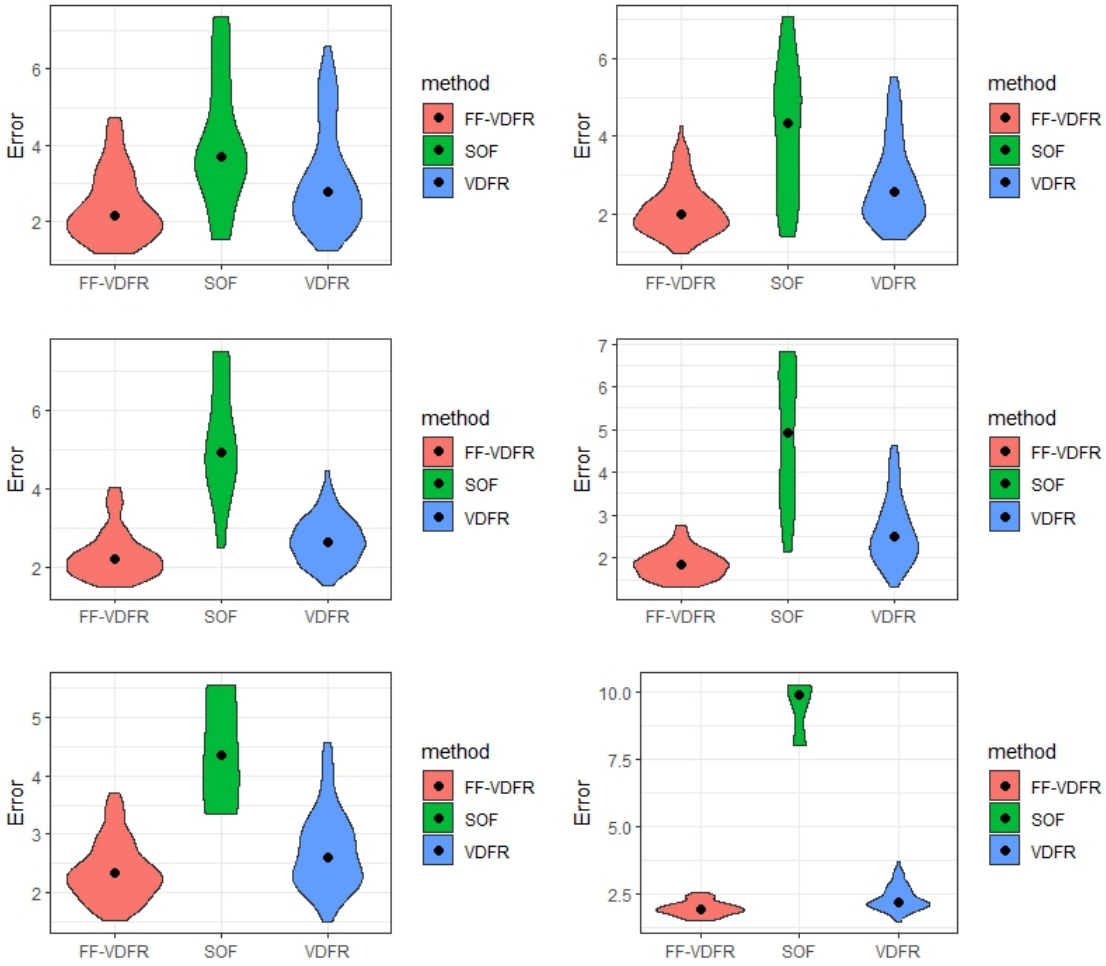

Fig. 2: Violin box-plots of the RMSE when the domain follows a negative binomial distribution and the response follows a Poisson distribution and the true functional coefficient is $\beta_3(t,T)$. Left column corresponds with the true functional data being smooth while the right column corresponds with its noisy counterpart. The up, middle, and bottom rows represent sample sizes of $N = 100, 200, 500$, respectively. The dot in the middle of the boxes represents the median value.

a worse performance was expected.

Regarding the AMSE, of the total 96 scenarios the FF-VDFR outperformed the VDFR model in 82 scenarios (85,5%). And from the scenarios where this model did not offer the best performance 50% of the cases correspond with noisy true functional data.

In summary, the FF-VDFR model outperforms all the others in the majority of the scenarios

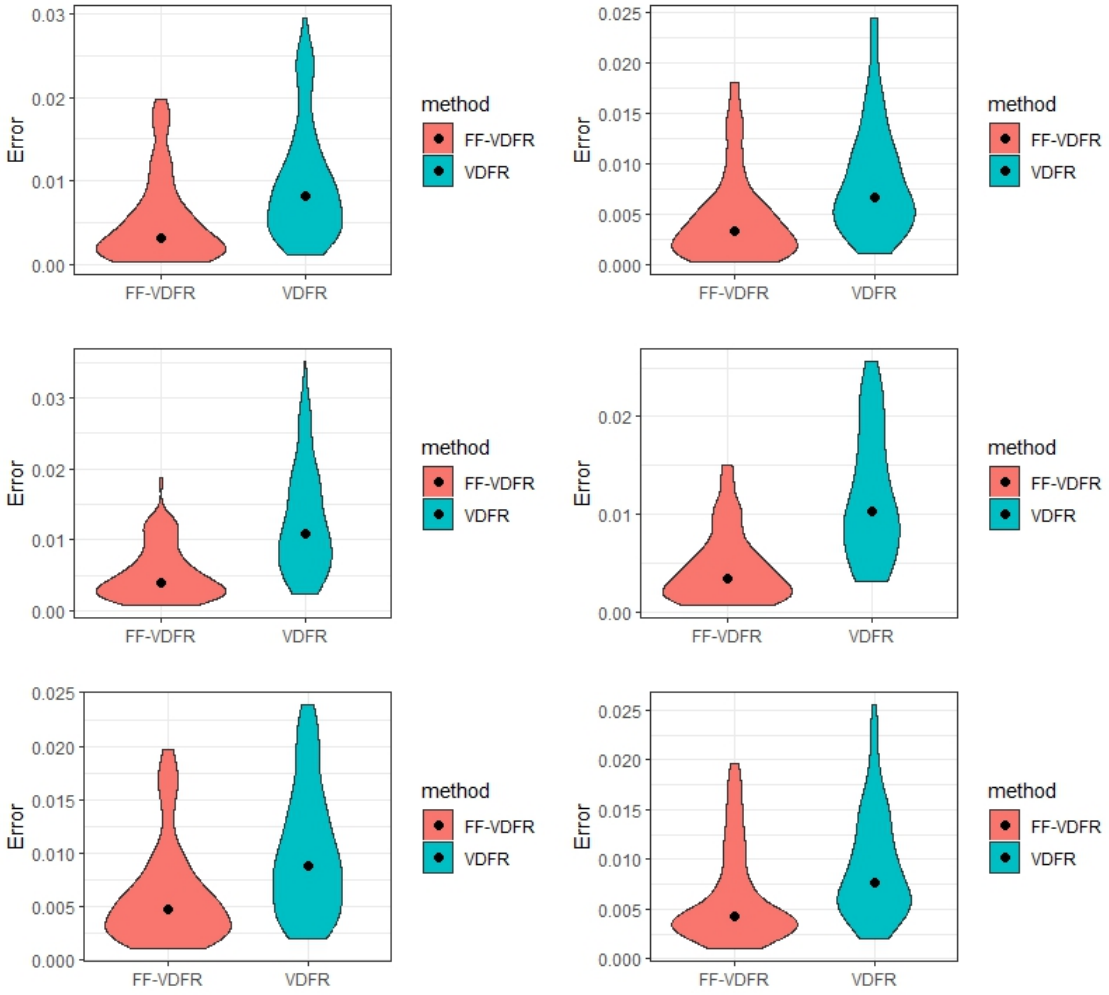

Fig. 3: Violin box-plots of the AMSE when the domain follows a negative binomial distribution and the response follows a Poisson distribution and the true functional coefficient is $\beta_3(t,T)$. Left column corresponds with the true functional data being smooth while the right column corresponds with its noisy counterpart. The up, middle, and bottom rows represent sample sizes of $N = 100, 200, 500$, respectively. The dot in the middle of the boxes represents the median value.

respecting both evaluation criteria used. Furthermore, most of the scenarios when the proposed model was not the best in performance correspond with the true functional data being noisy from which a worse performance of the FF-VDFR model is expected. But even in the 48 scenarios of noisy true functional data the FF-VDFR model is competitive, outperforming both the VDFR and the SOF models in terms of RMSE in 73% of the scenarios and outperforming the VDFR

models in terms of the AMSE in 85,4% of the scenarios.

In the next section, we show the results of applying our methodology to our motivational case study the telEPOC study.

5. Case study: The telEPOC dataset

In this section, we apply the proposed methodology to the telEPOC Study set to determine the possible relationship between physical activity and the number of hospitalizations due to COPD in patients.

5.1 telEPOC Study

The telEPOC Study (Esteban *and others*, 2016) was carried out at the Galdakao-Usansolo University Hospital (Biscay, Spain). Patient collection was done between the years 2010 and 2013, and the study includes five years of follow-up. The main goal of the study was to evaluate the efficacy of a telemonitoring-based program (telEPOC) in COPD patients with frequent hospitalizations. A total of 119 patients defined as those with frequent hospitalizations previous to inclusion were selected for telemonitoring at home.

Moreover, one of the goals of the study was to analyze the effect of performing physical activity on the health of the patients, in particular on the rate of hospitalizations due to COPD. The performance of daily physical activity was measured as the number of daily steps taken by each patient during their time in the study, which was included in the telemonitoring process. This has been the motivation of the work we present in this article.

Patients were included in the study at different time points. However, we are not interested in the effect, if any, that the different dates of admission on patients may have. For that reason, and a clearer analysis, we have aligned to the left all the collected variables making all patients begin the study at “day 1”.

Daily physical activity was analyzed for the 119 patients included initially in the study. The daily steps were recorded only for 112 out of the 119 patients. Besides, some of the measurements of the patients were out of the acceptable range of steps that a person can reach during the day, which were considered as missing data. Two more patients showed too many irregularities in the measurements, therefore they were eliminated. All the missing values for the daily steps were replaced by the mean of the previous and next days. Finally, we worked with a sample of 110 patients with a complete follow-up of daily physical activity. This situation reinforces the idea that observed data present errors and a previous smoothing will provide better results.

The study also collected clinical variables at baseline as possible covariates of interest: smoking habits, age, gender, previous hospitalizations due to COPD, anxiety, and depression symptomatology, among others. The number of hospitalizations due to COPD during the time in the study, the mortality, and the time spent in the study were recorded as potential outcomes. For a detailed explanation of the data collected in the study as well as the enrollment procedures and criteria of acceptance, we refer the readers to Esteban *and others* (2016).

5.2 Methodology

The response variable is the number of hospitalizations suffered by each patient and the functional covariate is the daily physical activity for each patient, measured as the daily steps they performed. Then, a fully functional variable domain functional Poisson regression model has been considered. However, the length of the follow-up depends on the patient, and so, the annual rate of hospitalizations was selected as response variable, instead of the number of hospitalizations, in order to avoid the cumulative effect of time in the study. All basis used for this data set are B-splines basis, the number of basis used for the functional covariate was 25 and the number of basis used for the bidimensional coefficient was 625 (25 for each marginal basis).

AIC criteria has been used to select other covariates in the model with adjusting purposes. The final model presents one functional covariate and four baseline non-functional covariates

namely: gender, previous hospitalizations, anxious symptomatology and depressive symptomatology. Nevertheless, for the interpretation of the results, we will focus on the effect of the functional variable on the annual rate of hospitalizations, adjusting by the rest of covariates in the model.

A negative value of the functional coefficient $\widehat{\beta}(t,T)$ will imply a positive influence on the patient's health, meaning that physical activity is helping to reduce the annual rate of hospitalizations. On the other hand, a positive value should not be interpreted as physical activity worsening the patient's health; that is, physical activity is not yet helping to reduce the rate of hospitalizations. However, interpretation should be cautious, without evidence of significant effect in any direction.

5.3 Results

In this section we will focus on the results obtained for the estimated functional coefficient $\widehat{\beta}(t,T)$, which reflects directly the relationship between the daily number of steps and the annual rate of hospitalizations due to COPD, adjusted by the baseline covariates in the model.

The estimated functional coefficient is a surface, where the curve resulting of fixing a value of T = Ti represents the influence of physical activity on the annual rate of hospitalizations for patients who have performed this physical activity during T_i days.

Figure 4 shows the heat map of the surface $\widehat{\beta}(t,T)$ for all the periods where patients carried out physical activity. The heat map presents a common feature: doing physical activity regularly during more than 6 months help to reduce the mean annual rate of hospitalizations due to COPD. This conclusion is based on the fact that all the curves longer than 180 days (the vertical dotted line) end up being negative represented with a cold color (green or blue), meaning a favorable influence of physical activity in the reduction of the annual rate of hospitalizations due to COPD.

For patients whose corresponding curve is shorter than 6 months, represented at the bottom of the heat map and marked with the red letter B, we can see how having performed physical

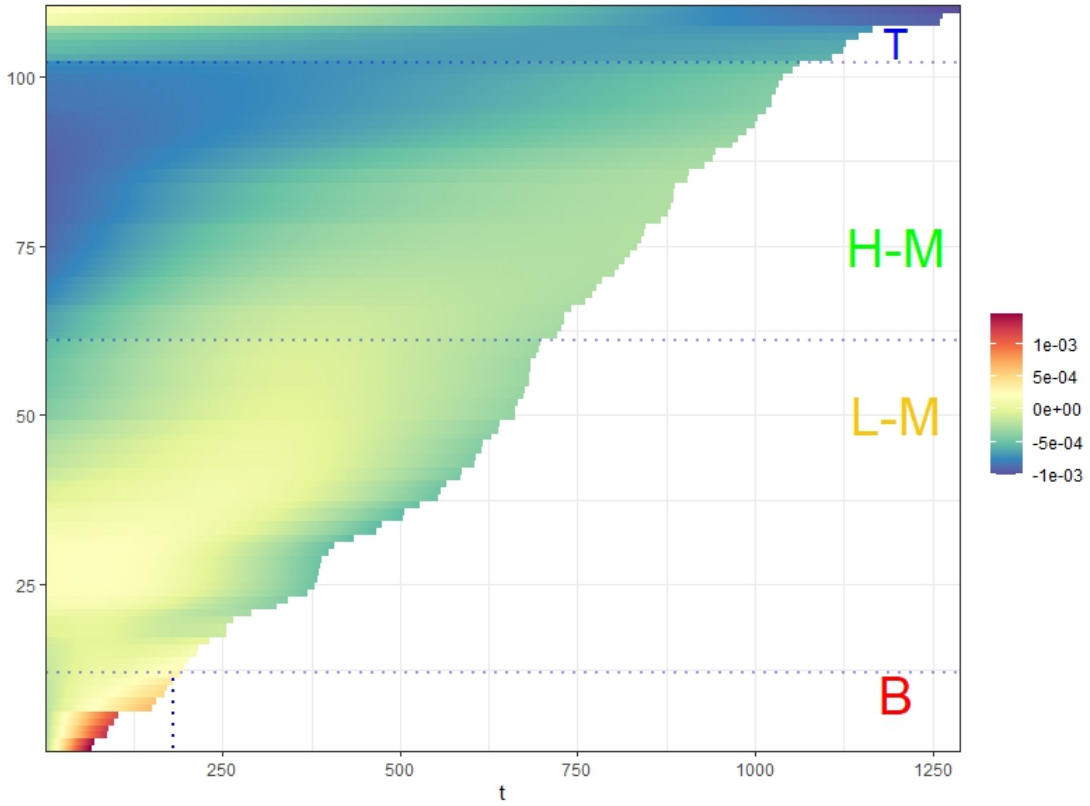

Fig. 4: Functional coefficient $\beta(t, T_i)$ for patients with T_i days in the study.

activity this amount of time is not sufficient to reduce the annual rate of hospitalizations since the end of the corresponding curves in the heat map is a warm color (red or orange) meaning a positive value of the curve.

A more exhaustive examination of the heat map shows several areas where the behavior is similar between patients. The first one already mentioned corresponds with the bottom part of the map; above that can be seen the low-mid area where the patient's influence begins almost null but turns positive towards the end, reflected in the negative values of the curves, this area is marked with the yellow letters L-M. The area directly above is the high-mid region and corresponds to patients whose curves start out negative, reflecting a positive influence on their health, but then increasing their value without ever becoming positive, this area is marked with the green letters H-M. And finally, a small area at the top of the heat map reveals curves that start near zero,

i.e., small or null influence over the patient's health, but rapidly decrease meaning that these patients soon begin to notice a positive influence of the physical activity on their health, this region is marked with the blue letter T.

In summary, we may conclude that within this group of patients, performing physical activity helps to reduce the annual rate of hospitalizations due to COPD. More specifically, it was shown that patients that perform physical activity for at least 6 year will see a reduction in their annual rate of hospitalizations.

6. Discussion

In this paper a fully functional approach for the variable domain functional generalized lineal model is proposed. This approach is based on assuming the basis representation of both the functional data and the functional coefficient. As a consequence, the functional model turns into a multivariate model, which has been reparametrized to a mixed model to gain computational efficiency. We refer to this new approach as fully functional variable domain functional regression model (FF-VDFR). This methodology can be seen as the extension of the methodology proposed in Aguilera-Morillo *and others* (2013) to the variable domain context.

The performance of the FF-VDFR model was tested via a simulation study and compared with the usual scalar-on-function model and the VDFR model, showing that the FF-VDFR model outperforms all the others in the evaluation criteria used.

The methodology presented was developed in order to analyse the influence of physical activity on the annual rate of hospitalizations in COPD patients. The analysis showed that a steady performance of physical activity for at least 6 months helps in the reduction of the annual rate of hospitalizations due to COPD.

The proposed methodology solves some of the limitations existing in previous approaches such as the optimality of the anisotropic penalties or the free choice of basis.

A future extension of the FF-VDFR model is the case of the function-on-function regression

models for a situation in which the regressors and/or the response variable present variable domain.

7. Acknowledgments

This work is supported by the grant ID2019-104901RB-I00 from the Spanish Ministry of Science, Innovation and Universities MCIN/AEI/10.13039/501100011033. This support is gratefully acknowledged.

8. Software

Software in the form of R code and complete documentation is available on request from the corresponding author (pahernan@est-econ.uc3m.es) and can also be found in `https://github.com/Pavel-Hernadez-Amaro/V.D.F.R.M-new-estimation-approach.git`

References

Aguilera, A. M. and Aguilera-Morillo, M. C. (2013, 10). Comparative study of different B-spline approaches for functional data. *Mathematical and Computer Modelling* **58**(7-8), 1568–1579.

Aguilera-Morillo, M. Carmen, Aguilera, Ana M., Escabias, Manuel and Valderrama, Mariano J. (2013, 6). Penalized spline approaches for functional logit regression. *Test* **22**(2), 251–277.

Breslow, N E and Clayton, D G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Technical Report* 421.

Brumback, Babette A, Ruppert, David and Wand, M P. (1999). Variable Selection

and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior: Comment. *Journal of the American Statistical Association* **94**(447), 794–797.

Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *Statistics and Probability Letters* **45**, 11–22.

Cardot, Hervé and Sarda, Pacal. (2005, 1). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis* **92**(1), 24–41.

De Boor, C. (2001). A practical guide to splines. *Springer*.

Eilers, Paul H C and Marx, Brian D. (1996). Flexible Smoothing with B-splines and Penalties. *Technical Report* 2.

Esteban, Cristóbal, Moraza, Javier, Iriberri, Milagros, Aguirre, Urko, Goiria, Begoña, Quintana, José M., Aburto, Myriam and Capelastegui, Alberto. (2016, 11). Outcomes of a telemonitoring-based program (telEPOC) in frequently hospitalized COPD patients. *International Journal of COPD* **11**(1), 2919–2930.

Gaynanova, Irina, Punjabi, Naresh and Crainiceanu, Ciprian. (2022, 1). Modeling continuous glucose monitoring (CGM) data during sleep. *Biostatistics* **23**(1), 223–239.

Gellar, Jonathan E., Colantuoni, Elizabeth, Needham, Dale M. and Crainiceanu, Ciprian M. (2014, 10). Variable-Domain Functional Regression for Modeling ICU Data. *Journal of the American Statistical Association* **109**(508), 1425–1439.

Goldsmith, Jeff, Bobb, Jennifer, Crainiceanu, Ciprian M., Caffo, Brian and Reich, Daniel. (2011, 12). Penalized functional regression. *Journal of Computational and Graphical Statistics* **20**(4), 830–851.

Goldsmith, Jeff, Scheipl, Fabian, Huang, Lei, Wrobel, Julia, Di, Chongzhi, Gellar,

Jonathan, Harezlak, Jaroslaw, McLean, Mathew W, Swihart, Bruce, Xiao, Luo, Crainiceanu, Ciprian *and others*. (2021). refund: Regression with Functional Data.

Horváth, Lajos and Kokoszka, Piotr. (2012). *Inference for Functional Data with Applications*.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society. Series B* **64**(3), 411–432.

Kulbaba, Mason W., Clocher, Ilona C. and Harder, Lawrence D. (2017, 11). Inflorescence characteristics as function-valued traits: Analysis of heritability and selection on architectural effects. *Journal of Systematics and Evolution* **55**(6), 559–565.

Lee, Dae-Jin. (2010). Smoothing mixed models for spatial and spatio-temporal data [Ph.D. Thesis]. Universidad Carlos III de Madrid.

Masak, T, Sarkar, S and Panaretos, V M. (2022). Separable expansions for covariance estimation via the partial inner product. *Biometrika*, 1–23.

Panayi, Efstathios, Peters, Gareth W and Kyriakides, George. (2017). Statistical Modelling for Precision Agriculture: A case study in optimal environmental schedules for Agaricus Bisporus production via variable domain functional regression. *PLoS One* **12**(9).

R Core Team. (2013). R: A Language and Environment for Statistical Computing.

Ramsay, James, Hooker, Giles and Graves, Spencer. (2009). *Functional Data Analysis with R and MATLAB*. Springer New York.

Ramsay, James O and Silverman, Bernard W. (2005). Applied Functional Data Analysis: Methods and Case Studies. *Technical Report*.

Rodríguez-Álvarez, María Xosé, Durban, Maria, Lee, Dae Jin and Eilers, Paul H.C. (2019, 5). On the estimation of variance parameters in non-standard generalised

linear mixed models: application to penalised smoothing. *Statistics and Computing* **29**(3), 483–500.

Rodriguez-Alvarez, Maria Xose and Oviedo de la Fuente, Manuel. (2021). SOP: Generalised Additive P-Spline Regression Models Estimation.

Wood, Simon N. (2017, 7). P-splines with derivative based penalties and tensor product smoothing of unevenly distributed data. *Statistics and Computing* **27**(4), 985–989.