MOX-Report No. 31/2022

Weighted functional data analysis for the calibration of ground motion models in Italy

Bortolotti, T; Peli, R.; Lanzano, G; Sgobba, S.; Menafoglio, A

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

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

Weighted functional data analysis for the calibration of ground motion models in Italy

Teresa Bortolotti[1*], Riccardo Peli[1], Giovanni Lanzano[2], Sara Sgobba[2], and Alessandra Menafoglio[1]

[1]MOX, Department of Mathematics, Politecnico di Milano, Milano, Italy
[2]Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Milano, Italy
[*]teresa.bortolotti@polimi.it

Abstract

Motivated by the crucial implications of Ground Motion Models (GMM) in terms of seismic hazard analysis and civil protection planning, this work extends a scalar ground motion model for Italy to the framework of Functional Data Analysis. The inherent characteristic of seismic data to be incomplete over the observation domain entails embedding the analysis in the context of partially observed functional data. This work proposes a novel methodology that combines pre-existing techniques of data reconstruction with the definition of observation-specific functional weights, which enter the estimation process to reduce the impact that the reconstructed parts of the curves have on the final estimates. The classical methods of smoothing and concurrent functional regression are extended to include weights. The advantages of the proposed methodology are assessed on synthetic data. Eventually, the weighted functional analysis performed on seismological data is shown to provide a natural smoothing and stabilization of the spectral estimates of the GMM.

Keywords: Functional Data Analysis, Weighted analysis, Partially observed functional data, Ground motion model

1 Introduction

In the field of seismic hazard assessment, Ground Motion Models (Douglas and Edwards, 2016) estimate the distribution of ground motion intensity measures, conditionally on parameters that are descriptive of a certain seismic scenario. Earthquake-induced ground motion is typically measured at the recording sites by the seismic response spectrum, which is the peak response of a set of damped harmonic oscillators to the seismic force, each characterised by its natural period of oscillation T (Newmark and Hall, 1982). Consequently, the intensity measure can either be seen as single ordinates defined with respect to T, or as profiles along the range of vibration periods. This gives rise to the threefold possibility of inserting the analysis of ground motion in a scalar (*e.g.* Bindi et al. (2011), Lanzano et al. (2019), Kotha

et al. (2016), Boore et al. (2014)), multivariate (*e.g.* Worden et al. (2018), Huang and Galasso (2019)), or functional context (Menafoglio et al., 2020). A scalar approach ignores the correlation between the ordinates of the spectrum. In considering this correlation, on the other hand, multivariate approaches inevitably suffer from the curse of dimensionality. Embedding ground motion models in the context of Functional Data Analysis (FDA, Ramsay and Silverman (2005), Horváth and Kokoszka (2012)) solves the shortcomings of both scalar and multivariate approaches, by moving the focus from the period-specific intensity measures to their continuous profile over the domain of vibration periods.

By exploiting the methodologies of FDA on the seismic recorded data, this work aims to provide a functional extension of ITA18, the ground motion model that provides scalar spectral ordinates for shallow crustal earthquakes in Italy, proposed in Lanzano et al. (2019) and calibrated on the same database.

The peculiarity of the data of ground motion analysed in this work is that their processing is manual. The non-automatic handling of the recordings results in high-pass corner frequencies that differ from datum to datum, generating the problem that a non-negligible number of curves are only observed on subsets of the whole domain. Since such data are effective in populating the dataset with information that produces robust regression results, and since there is seismological interest in doing inference over the entire period domain, we are reluctant in erasing data from the dataset, or in reducing the domain of analysis similarly to what Menafoglio et al. (2020) did in their work. Rather, we are motivated in embedding the analysis in the context of partially observed functional data.

Most classical methodologies of FDA do not generalize to the case of data that are not completely observed over the domain. Recently, *ad-hoc* techniques for partially observed functional data arose aiming to obtain estimates of the mean and of the covariance operator (Yao et al. (2005), Kraus (2015)), to perform functional principal component analysis (Stefanucci et al. (2018), Kraus and Stefanucci (2018), Yao et al. (2005)), and to impute missing trajectories to the unobserved parts of the domain (*e.g.* Kraus (2015), Kneip and Liebl (2020)). This work exploits these last techniques to reconstruct the missing observations of the acceleration spectra, in order to preserve the formulation of the functional ground motion model over the entire period domain.

The methodological novelty proposed in this work fits downstream of curves reconstruction. The idea is to build a workflow of analysis that keeps track of the fact that the degree of uncertainty associated to the discrete observations of the curves may be variable within the single functional datum. In particular, the analysis should associate less confidence to the parts of a curve that underwent reconstruction, with respect to those that are originally observed. This requires to modify the optimization criteria for smoothing and functional regression, in a way that greater weight is given to the estimation errors made on the observed values of a curve, and less weight to those made on the reconstructed values. The classical technique of penalized least squares for smoothing presented in Ramsay and Silverman (2005) easily extends to the inclusion of scalar weights, which vary over the sampling instants but are common to all functional observations. This use of weights allow the op-

timal smoothed curve to be characterised by various degrees of regularity over the domain. In the non-parametric context, methods of weighted smoothing splines are employed with an equivalent purpose and belong to the category of spatially adaptive splines (*e.g.* Pintore et al. (2006), Davies and Meise (2008)). In the weighted penalized least square criterion for functional regression (Ramsay and Silverman, 2005), weights vary across observations but are constant over the domain of analysis. The present work extends the techniques of weighted penalized smoothing and weighted penalized functional linear regression discussed in Ramsay and Silverman (2005), to include curve-specific functional weights. Specifically, the proposed framework couples each curve with a weight function, taking value one where original observations are available and decreasing to zero the further the reconstructed trajectory gets from the last recorded value.

2 Model and data

2.1 Model

In the context of seismic hazard assessment, Ground Motion Models estimate the distribution of an intensity measure (IM) of ground motion conditionally on seismic parameters that are descriptive of the *source* of the earthquake, the *site* of registration and the *path* taken by the seismic wave from the epicentre to the recording station. The focus of this work lies in the extension to a functional framework of the scalar GMM proposed in Lanzano et al. (2019), which we refer to as ITA18.

Background In Lanzano et al. (2019), the authors resort to a linear ordinary least-square regression to separately fit 37 models of the IMs, *i.e.* peak ground acceleration (PGA) and the ordinates of elastic acceleration response spectra, SA at 5% damping (Douglas, 2003), each corresponding to a vibration period $T_j \in \mathcal{T} := [0.04, 10\ s],\ j = 1, \ldots, 36$. The median values of such IMs are estimated according to the following functional form:

$$\log_{10} \mathrm{IM} = a + F_M(M_w, SoF) + F_D(M_w, R) + F_S(V_{S30}) + \epsilon, \tag{1}$$

where a is the offset, $F_M(M_w, SoF)$, $F_D(M_w, R)$, $F_S(V_{S30})$ are the *source*-, *path*- and *site*-related terms respectively, and ϵ is the remaining error. The source is specified as a step-wise linear function

$$\begin{aligned} F_M(M_w) &= \begin{cases} b_1(M_w - M_h) & M_w \leq M_h \\ b_2(M_w - M_h) & M_w > M_h \end{cases}, \\ F_M(SoF) &= f_k SoF_k, \end{aligned}$$

in which the straight line changes slope at the hinge magnitude M_h. Terms f_k, for $k = 1, 2, 3$, are the coefficients related to three dummy variables accounting for the *style-of-faulting* (SoF_k: strike-slip, thrust faulting, normal faulting). Coefficient f_3 related to the normal faulting is constrained to zero when the regression is performed.

The path term takes the form

$$F_D(M_w, R) = [c_1(M_w - M_{\text{ref}}) + c_2] \log_{10}(R) + c_3 R,$$

where parameter M_{ref} is the reference magnitude. The two terms of this summation account for the geometrical spreading of the waves from the source and for the anelastic attenuation, respectively. The distance R represents a correction of the pure Joyner-Boore distance – *i.e.* the closest distance to the surface projection of an extended fault – and is defined as $R = \sqrt{d_{JB}^2 + h^2}$, where h is the parameter of pseudodepth measured in kilometres.
Lastly, the site-related term has the form

$$F_S(V_{S30}) = k \log\left(\frac{V_0}{800}\right),$$

where $V_0 = V_{S30}$ if $V_{S30} \leq 1500\ m/s$, $V_0 = 1500\ m/s$ otherwise. According to Kamai et al. (2014), the scaling with V_{S30} is assumed to be linear for values lower than $1500\ m/s$, while for larger values the amplification here is considered as independent on the shear-wave velocity.

Functional embedding of the scalar model Parameters M_h, M_{ref} and h appearing in (1) are known to be dependent on the spectral periods. For this reason, they are typically included in the regression model either as known (Sabetta et al., 2021) or unknown (Lanzano et al., 2019) functions of the vibration period. Since the latter approach is non-trivial when applied to a functional framework, resulting in a non-linear regression model, we assume them to be known functions of the period. In particular, we take advantage of the estimates of M_h, M_{ref} and h obtained period-wise from the preliminary step of non-linear regression discussed in Lanzano et al. (2019). Functions $\mathcal{M}_{\text{ref}}$ and h are defined in the space generated by a cubic B-spline basis, where the optimal coefficients are the result of a step of penalized smoothing. The estimate of M_h made by Lanzano et al. (2019) forces a step-wise behaviour along the period domain, producing jumps in the prediction of the spectrum for scenarios close to the hinge magnitude. In order to solve the discontinuity issues in the predictions, the work of Sabetta et al. (2021) corrects M_h to have a smoother variation in the range of periods $[0.25\ s, 0.7\ s]$. Following this line, we define function $\mathcal{M}_h$ on a basis of quadratic B-spline via a smoothing that penalizes its first derivative. Figure 1a shows how the smoothing of M_h results in a continuous function over the period domain.
We acknowledge that these are modelling choices made *a priori* according to how the issue is typically handled in the literature on this topic. Such choices may be revised in further extensions of the work, the most straightforward of which extends the functional form (1) to a non-linear regression.
A functional definition of the covariates in (1) follows naturally, and eventually leads to the embedding of the scalar model into a fully functional framework:

$$\log_{10} \mathcal{IM} = \alpha + \mathcal{F}_M(M_w, \text{SoF}, \mathcal{M}_h) + \mathcal{F}_D(M_w, d_{JB}, \mathcal{M}_{\text{ref}}, h) + \mathcal{F}_S(V_{S30}) + \mathcal{E}. \quad (2)$$

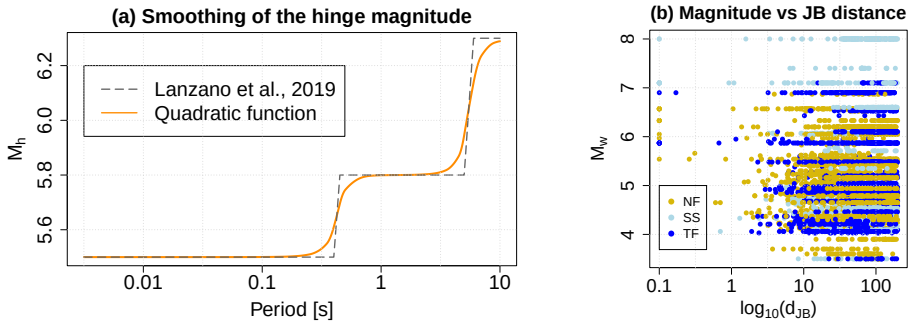

Figure 1: (a) Definition of M_h as the function resulting from the smoothing of its step-function estimate obtained in Lanzano et al. (2019). (b) Scatter plot of magnitude vs Joyner-Boore distance (d_{JB}), coloured by style-of-faulting. The records at $d_{JB} = 0$ are plotted at 0.1 km.

In (2), $\mathcal{IM}$ is a random variable with values in a convenient functional space, α, $\mathcal{F}_M$, $\mathcal{F}_D$ and $\mathcal{F}_S$ are known functions with domain $\mathcal{T}$ and $\mathcal{E}$ is assumed to be generated by a zero mean stochastic process.

2.2 Data

The analysis is carried out on the same dataset used for the calibration of ITA18, which includes 5568 records, relative to 146 earthquakes and 1657 stations. The bulk of the data comes from the ITalian ACcelerometric Archive (ITACA; Russo et al. (2022)), which collects the manually-revised and good quality waveforms recorded by the most important and large seismic networks in Italy. The data included in the ITACA collection were selected according to the following criteria: (1) earthquakes of active shallow crustal regions (only events of tectonic origin with focal depth lower than 30 km) occurred in the period time 1972–2017, (2) minimum moment magnitude (M_w) set to 3.5, (3) Joyner–Boore distance lower than 200 km, and (4) stations with surface instruments and with low or no interactions with nearby structures. The dataset was also enriched with recordings of high-magnitude ($M_w > 6.1$) worldwide events associated to strike-slip and thrust faulting mechanisms. Additional details on the dataset selection are provided in Lanzano et al. (2018). Figure 1b shows the magnitude-distance distribution of the calibration data that supports the reliability of the model calibration in the intervals 3.5-8 and 0.1-200 km for magnitude and distance, respectively.

Domain definition The sampling of the discrete observations of IM is not uniform over $[0, 10\ s]$. Conversely, 26 out of the 37 sampling instants are in the interval $[0, 2\ s]$, while the other 11 points span the remaining of the domain. This motivates us to define the spectrum on the interval $\mathcal{T} = (\log_{10}(0), \log_{10}(10)]$, thus considering the sequence $(\log_{10}(t_1), \ldots, \log_{10}(t_N))$ as the sampling instants. Such modelling choice has the twofold advantage of obtaining a more uniform sampling of the curves over the domain of definition and of better representing the greater seismological interest

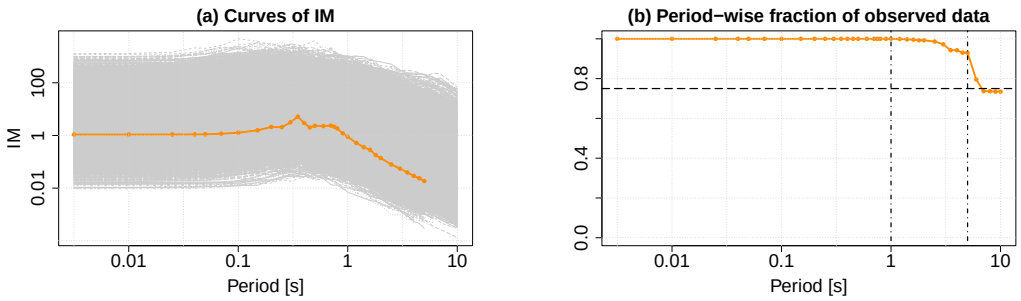

Figure 2: (a) Curves of IM. The orange line represents a curve that is not fully observed up to 10 s. (b) Period-specific fraction of the observed records over the total records. The dashed horizontal line marks 75%. The vertical lines mark the last sampling points where 100% and >90% of curves are observed.

that practitioners have on short rather than long periods. It is worth pointing out that we decide to cut the domain on the left by making the PGA correspond to the ordinate $SA(log(T) = -2.5)$ rather than to $SA(log(T) = -\infty)$ (Bradley (2011), Lanzano et al. (2019)). Doing so, we are reducing the impact that the left tail of the definition domain would else have in the estimation process. Figure 2a displays the longitudinal observations of intensity measure resulting from these modelling choices.

Partially observed response variable RotD50 (Boore, 2010), which is the intensity measure considered as response variable in this work, results from the combination of three mutually orthogonal components of spectral acceleration measured at the recording sites. Accelerometric stations make use of high-pass filters that may differ from site to site and from component to component of spectral acceleration. This implies that some longitudinal observations may not be validly recorded at all registration periods, but only at the lower ones. Figure 2b shows, for each registration period T, the fraction of longitudinal data that are observed at T. We may notice that the percentage of unobserved curves is low and stable up to a period of about 5 s, and that it rapidly increases up to 25% at 10 s. We refer to Sections 4.2 and 5.1 for a report on the strategies adopted to reconstruct the missing trajectories of the curves, from their last valid observation up to $T = 10$ s.

3 Methods

Let $\mathcal{T}$ be an open subset of $\mathbb{R}$ and $w : T \to [0, 1]$ be a bounded non-negative function, which we refer to as *weight*. Now let $f, g \in L^2(\mathcal{T})$ and let w, v be weights associated to f and g respectively. We define the *weighted inner product* in L^2 as

$$\langle f, g \rangle_W = \int_{\mathcal{T}} \sqrt{w(s)} f(s) \sqrt{v(s)} g(s) ds,$$

and the *weighted* L^2 *norm* of f with respect to w as $||f||_W = \sqrt{\langle f, f\rangle_W}$. It is trivial to see that if the L^2 norm of f is finite, then also the weighted L^2 norm of f is finite.

3.1 Weighted smoothing

Let $y_1, \ldots, y_n$ be realizations of independent and identically distributed functional random variables with values on $L^2(\mathcal{T})$, $\mathcal{T} \subset \mathbb{R}$. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be vectors of discrete observations of the curves at the sampling instants $t_1, \ldots, t_N$. For each observation $\boldsymbol{y}_i$, the smoothing technique (Ramsay and Silverman (2005), de Boor (1978)) fits the discrete observations $y_i(t_1), \ldots, y_i(t_N)$ according to the model

$$y_i(t_j) = x_i(t_j) + \epsilon_i(t_j), \quad \forall\, j = 1, \ldots, N.$$

Penalized weighted least-square criterion Define *smoothing error* the square integrable function $\epsilon_i = y_i(t_j) - x_i(t_j)$, and let w_i be the associated weight as specified in the introduction to this section. Then the smoothing is performed resorting to a *penalized weighted least-square* (PWLS) criterion, that solves

$$\hat{x}_i = \underset{x \in H^2(\mathcal{T})}{\operatorname{argmin}} \sum_{j=1}^{T} \left(\sqrt{w_i(t_j)}(y_i(t_j) - x(t_j)) \right)^2 + \zeta ||D^2 x||^2_{L^2(\mathcal{T})}. \tag{3}$$

In (3), the term $||D^2 x||^2_{L^2(\mathcal{T})}$ quantifies how abrupt the changes in the smoothed curve are, by evaluating the L^2 norm of its second derivative, while ζ is the smoothing parameter, which may be tuned via generalized cross-validation. Hereafter, we assume that x belongs to the Sobolev space $H^2(\mathcal{T})$ (*i.e.*, that has first and second derivatives in $L^2(\mathcal{T})$) to guarantee the finiteness of $||D^2 x||^2_{L^2(\mathcal{T})}$. Notice that the error sum of squares in (3) is discounted at each sampling instant t_j by the value of w_i in t_j, w_i playing the role of giving different weight to smoothing errors made at different time instants.

Problem (3) is formulated in an infinite dimensional space. In order to solve it, dimensionality is reduced by projecting the smooth curve x on the space generated by a finite set of basis functions $\{\phi_1, \ldots, \phi_L\}$, so that it may be expressed in the form

$$x_i(t) = \sum_{l=1}^{L} c_{il}\phi_l(t) = \boldsymbol{c}_i^T \boldsymbol{\phi}(t), \tag{4}$$

and hence univocally identified with respect to the basis by the vector of coefficients of the linear combination of the basis functions, $\boldsymbol{c} \in \mathbb{R}^L$. Doing so, it is possible to embed the problem in a finite dimensional space, and to express the smoothing criterion in matricial form. This implies that the smoothing criterion may equivalently be expressed for x or $\boldsymbol{c}$.
First, observe that the penalization term may be re-expressed as

$$||D^2 x||^2_{L^2(\mathcal{T})} = \int_{\mathcal{T}} \left[(D^2 x)(s)\right]^2 ds = \int_{\mathcal{T}} \left[(D^2 \boldsymbol{c}^T \boldsymbol{\phi})(s)\right]^2 ds = \boldsymbol{c}^T P \boldsymbol{c}, \tag{5}$$

where $[P]_{lk} = \langle D^2\phi_l, D^2\phi_k \rangle_{L^2(\mathcal{T})}$.
Secondly, observe that the weighted error sum of squares may be written as

$$\sum_{j=1}^{N} \left(\sqrt{w_i(t_j)}(y_i(t_j) - \boldsymbol{c}^T \boldsymbol{\phi}(t)) \right)^2 = (\boldsymbol{y}_i - \Phi \boldsymbol{c})^T W_i (\boldsymbol{y}_i - \Phi \boldsymbol{c}), \tag{6}$$

where $W_i = \text{diag}(w_i(t_1), \ldots, w_i(t_N))$ is a N-order diagonal matrix, and $\Phi \in \mathbb{R}^{(L \times N)}$ contains the values taken by the basis functions at the sampling instants.
Equation (5) and (6) allow one to formulate the PWLS criterion (3) as the problem of finding the minimum $\boldsymbol{c} \in \mathbb{R}^L$ of the quadratic form

$$(\boldsymbol{y}_i - \Phi \boldsymbol{c})^T W_i (\boldsymbol{y}_i - \Phi \boldsymbol{c}) + \zeta \boldsymbol{c}^T P \boldsymbol{c}. \tag{7}$$

It is immediate to verify that, by taking the derivative of (7) and by setting it to zero, the solution is found in closed form as

$$\hat{\boldsymbol{c}}_i = \left(\Phi^T W_i \Phi + \zeta P \right)^{-1} \Phi^T W_i \boldsymbol{y}_i, \tag{8}$$

so that it is possible to identify a smoothing map $S_\Phi^{w_i} = \left(\Phi^T W_i \Phi + \zeta P \right)^{-1} \Phi^T W_i$ such that

$$\hat{\boldsymbol{c}}_i = S_\Phi^{w_i} \boldsymbol{y}_i. \tag{9}$$

Construction of the smoothing map Let $Y \in \mathbb{R}^{(n \times N)}$ be the matrix containing the values that the n observations take in N sampling points, and let $C \in \mathbb{R}^{(n \times L)}$ be the matrix that collects all the n optimal coefficient vectors, each of which found via equation (9). We are interested in identifying a smoothing map $\boldsymbol{S}_\Phi$ that collectively links matrices Y and C.
The identification of such a map comes from an extension of the case of classical weighted smoothing (Ramsay and Silverman, 2005). The latter accounts for correlations among time instants by applying the same weighting to all raw data, and by identifying a smoothing map S_Φ that is common to all the observations. In this case the mapping of Y into C is easily given by the common S_Φ in the form

$$C = Y S_\Phi^T. \tag{10}$$

In a more general setting than the classical weighted smoothing, the weights system is applied differently to each raw curve, and requires the computation of n curve-specific smoothing maps $S_\Phi^{w_i}$, so that a valid counterpart of equation (10) cannot be identified directly, but needs a little more handling.
Observe that by applying the vec() operator to both sides of (10) and exploiting the properties of the Kronecker product one gets

$$\text{vec}(C) = \text{vec}(Y S_\Phi^T) = (S_\Phi \otimes I)\text{vec}(Y), \tag{11}$$

where I is the n-dimensional identity matrix. The Kronecker product $(S_\Phi \otimes I)$ expands the common mapping S_Φ in a sparse matrix of dimension $(Ln \times Nn)$.

Let now S^{w_i}, $i = 1, \dots, n$, be the n curve-specific smoothing maps, where we omit subscript Φ for clarity of notation. Then one may check that the counterpart of $(S_\Phi \otimes I)$ in this general setting is given by the matrix $\boldsymbol{S}_\Phi$ having the form

$$\boldsymbol{S}_\Phi \coloneqq \begin{pmatrix} S_{11}^1 & 0 & \dots & 0 & S_{12}^1 & 0 & \dots & 0 & & S_{1T}^1 & 0 & \dots & 0 \\ 0 & S_{11}^2 & \dots & 0 & 0 & S_{12}^2 & \dots & 0 & & 0 & S_{1T}^2 & \dots & 0 \\ \vdots & & \ddots & & \vdots & & \ddots & & & \vdots & & \ddots & \\ 0 & \dots & 0 & S_{11}^n & 0 & \dots & 0 & S_{12}^n & & 0 & \dots & 0 & S_{1T}^n \\ & & & & & & & & & & & & \\ & & & & & & & & & & & & \\ S_{L1}^1 & 0 & \dots & 0 & S_{L2}^1 & 0 & \dots & 0 & & S_{LT}^1 & 0 & \dots & 0 \\ 0 & S_{L1}^2 & \dots & 0 & 0 & S_{L2}^2 & \dots & 0 & & 0 & S_{LT}^2 & \dots & 0 \\ \vdots & & \ddots & & \vdots & & \ddots & & & \vdots & & \ddots & \\ 0 & \dots & 0 & S_{L1}^n & 0 & \dots & 0 & S_{L2}^n & & 0 & \dots & 0 & S_{LT}^n \end{pmatrix},$$

so that eventually the following holds:

$$\mathrm{vec}(C) = \boldsymbol{S}_\Phi \mathrm{vec}(Y). \tag{12}$$

3.2 Weighted regression

As our application interest lies in a functional extension of a linear regression with functional covariates, we consider a *functional concurrent linear regression model* with independent functional covariates $x_1(t), \dots, x_q(t)$. In the framework defined in Section 3.1, the model formulates

$$\boldsymbol{y}(t) = X(t)\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t), \qquad t \in \mathcal{T}, \tag{13}$$

where $\boldsymbol{\beta}(t) = (\beta_1(t), \dots, \beta_q(t))^T$ denotes the vector of functional coefficients evaluated in t, $X(t) \in \mathbb{R}^{n \times q}$ is the design matrix at t and $\boldsymbol{y}(t)$ is a n-dimensional vector containing the response functions evaluated in t. The error term is a n-dimensional vector of functions ϵ_i, that are assumed to be independent realizations of a zero-mean stochastic process.

Penalized weighted functional least-square criterion Similarly to what we did for the smoothing, the aim of this section is to extend to a weighted approach the *penalized functional least-square* criterion discussed in Ramsay and Silverman (2005), by exploiting the features of the weighted L^2 norm introduced above. Notice that the systems of weights introduced for smoothing and here for regression could in principle be different, as they could account for different types of uncertainty. This work, however, treats them as equal, and regards the two weighted techniques as a unique, novel procedure.
Let $w_1, \dots, w_n$ be the weights associated to the errors $\epsilon_1, \dots, \epsilon_n$. Then we define the

penalized weighted functional least-square (PWFLS) criterion as the minimization of

$$\begin{aligned}\text{PWFLS} &= \sum_{i=1}^{n} ||\epsilon_i||^2_{L^2(\mathcal{T}),W} + \sum_{j=1}^{q} \lambda_j ||D^2\beta_j||^2_{L^2(\mathcal{T})} \\ &= \sum_{i=1}^{n} \int_{\mathcal{T}} \left(\sqrt{w_i(s)}\epsilon_i(s)\right)^2 ds + \sum_{j=1}^{q} \int_{\mathcal{T}} \lambda_j \left(D^2\beta_j(s)\right)^2 ds, \end{aligned} \tag{14}$$

where $\sum_{j=1}^{q} \lambda_j ||D^2\beta_j||^2_{L^2(\mathcal{T})}$ is a roughness penalty that enters the criterion to regularize and stabilize the estimates of the regression coefficients, and $\lambda_1, \ldots, \lambda_q$ are the coefficient-specific penalization parameters which can be tuned via generalized cross-validation to allow for different degrees of smoothness in the coefficients estimates. Recall that the roughness penalty $\sum_{j=1}^{q} \lambda_j ||D^2\beta_j||^2_{L^2(\mathcal{T})}$ allows one to estimate regression coefficients β_j, which are in principle infinite dimensional, from a finite sample (Horváth and Kokoszka, 2012), counterbalancing the pursuit of a good fitting with the estimation of a coefficient that is regular, stable and able to provide useful insights on the phenomenon under analysis.

By linearity of the integral, the operations of integration and summation in (14) can be interchanged, and consequently one may write

$$\begin{aligned}\text{PWFLS} &= \int_{\mathcal{T}} \sum_{i=1}^{n} \left[\sqrt{w_i(s)}\left(y_i(s) - \boldsymbol{x}_i(s)^T\boldsymbol{\beta}(s)\right)\right]^2 ds + \int_{\mathcal{T}} \sum_{j=1}^{q} \lambda_j \left(D^2\beta_j(s)\right)^2 ds \\ &= \int_{\mathcal{T}} \left[\boldsymbol{y}(s) - X(s)\boldsymbol{\beta}(s)\right]^T W(s) \left[\boldsymbol{y}(s) - X(s)\boldsymbol{\beta}(s)\right] ds + \int_{\mathcal{T}} [L\boldsymbol{\beta}(s)]^T \Lambda [L\boldsymbol{\beta}(s)] ds, \end{aligned} \tag{15}$$

where we set $W(s) = \text{diag}\left(w_1(s), \ldots, w_n(s)\right)$ to be the diagonal matrix of the weights evaluated in s, L to be a linear differential operator taking the second derivative of each regression coefficient, and $\Lambda = \text{diag}\left(\lambda_1, \ldots, \lambda_q(s)\right)^T$ the diagonal matrix of the q penalization parameters.

The minimization of (15) passes through a dimensionality reduction of the problem. The idea consists in moving from an infinite dimensional setting to a multivariate framework, by projecting the observations in the space spanned by the basis functions, as already done in Section 3.1. The functional coefficients are estimated as elements of finite dimensional spaces generated by suitable basis functions, and the part of the curve that remains not captured is assumed to be negligible and included in the regression error. Accordingly, this section will refer to the response variable

and to the functional coefficients as

$$y_i(t) = \sum_{l=1}^{L_y} c_{il}\phi_l(t), \qquad i = 1, \ldots, n,$$
$$\beta_j(t) = \sum_{l=1}^{L_j} b_{jl}\theta_l^j(t), \qquad j = 1, \ldots, q. \tag{16}$$

The argument above is formulated in its most general setting, which considers the bases for the observations and for each one of the regression coefficients as distinct. Such comprehensiveness is particularly convenient when one has *a priori* knowledge that the effects entering the regression model have different levels of roughness or smoothness, as it allows to flexibly adjust the definition of each coefficient β_j in the space generated by suitable basis functions $\theta_1^j, \ldots, \theta_{L_j}^j$. Note that formulation (16) can be compacted in matricial form as follows

$$\boldsymbol{\beta}(t) = \Theta(t)\boldsymbol{b}, \tag{17}$$

where $\Theta(t)$ is the $q \times L_\beta$ matrix of the point evaluations at t of the basis functions

$$\Theta(t) = \begin{pmatrix} \theta_1^1(t) & \theta_2^1(t) & \ldots & \theta_{L_1}^1(t) & 0 & 0 & \ldots & 0 & & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & \theta_1^2(t) & \theta_2^2(t) & \ldots & \theta_{L_2}^2(t) & & 0 & 0 & \ldots & 0 \\ \vdots & & \ddots & & \vdots & & \ddots & & \cdots & \vdots & & \ddots & \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & & \theta_1^q(t) & \theta_2^q(t) & \ldots & \theta_{L_q}^q(t) \end{pmatrix},$$

and $\boldsymbol{b}$ is the L_β-dimensional vector of the coefficients of the projections of $\{\beta_1, \ldots, \beta_q\}$ on bases $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q$.
A new phrasing of model (13) follows all the considerations made above and takes the form

$$\mathrm{C}\boldsymbol{\phi}(t) = X(t)\Theta(t)\boldsymbol{b} + \mathcal{E}(t). \tag{18}$$

Putting (18) in (15) one gets

$$\int_{\mathcal{T}} \left[\mathrm{C}\boldsymbol{\phi}(s) - X(s)\Theta(s)\boldsymbol{b}\right]^T W(s) \left[\mathrm{C}\boldsymbol{\phi}(s) - X(s)\Theta(s)\boldsymbol{b}\right] ds + \int_{\mathcal{T}} \left[L\Theta(s)\boldsymbol{b}\right]^T \Lambda \left[L\Theta(s)\boldsymbol{b}\right] ds.$$

This quadratic form is the starting point of the calculation, extensively reported in Appendix A.1, that leads to the following equation for vector $\boldsymbol{b}$:

$$\left[J + R\right]\boldsymbol{b} = \int \Theta(s)^T X(s)^T W(s) \mathrm{C}\boldsymbol{\phi}(s) ds, \tag{19}$$

where

$$J \coloneqq \left(\int_{\mathcal{T}} \Theta(s)^T X(s)^T W(s) X(s)^T \Theta(s) ds \right),$$

and R is the $L_\beta \times L_\beta$ matrix accounting for the penalization term.

3.3 Uncertainty assessment

Following the argument reported in Appendix A.2, one finds that the raw observations Y are mapped into the matrix of coefficients $\hat{\boldsymbol{\beta}}$ according to the relation

$$\hat{\boldsymbol{\beta}}(t) = S_\Theta(t) S_\beta S_\Phi \text{vec}(\text{Y}). \tag{20}$$

The linkage expressed by (20) is crucial for the assessment of the variability associated to the point estimates of the regression coefficients. On the one hand, it allows one to quantify the point-wise variability associated to each β_j; on the other, it justifies the use of a bootstrap approach to generate a sample of regression coefficients and estimate their variability simultaneously over the periods domain.

Point-wise variability The point-wise variability of $\hat{\boldsymbol{\beta}}$ comes easily from (20). It suffices to observe that the variance of the observations Y is given by

$$\text{Var}\left(\text{vec}(Y)\right) = \Sigma_e \otimes I_n,$$

where Σ_e is the covariance matrix of the residuals $\hat{\epsilon}_i$ of the regression model. Then one immediately finds that the $q \times q$ covariance matrix of vector $\hat{\boldsymbol{\beta}}(t)$, $\forall t \in \mathcal{T}$, is given by

$$\text{Var}\left(\hat{\boldsymbol{\beta}}(t)\right) = \text{Var}\left(S_\Theta(t) S_\beta S_\Phi \text{vec}(Y)\right) = S_\Theta(t) S_\beta S_\Phi (\Sigma_e \otimes \text{I}_n) S_\Phi^T S_\beta^T S_\Theta(t)^T. \tag{21}$$

Simultaneous variability Equation (21) provides an estimate of the variability that has point-wise validity, meaning that conclusions based on it can only be drawn one-at-a-time. To overcome the intrinsic limitations of such an estimate, the present work makes use of a method based on a bootstrap resampling that quantifies the uncertainty associated to $\hat{\beta}$'s simultaneously over the whole domain.
Bootstrap resampling methods and results of asymptotic validity of the bootstrap methodology, which are guaranteed in the scalar case by the law of large numbers and by the Glivenko-Cantelli theorem, find an extension in the framework of functional data analysis, where the distributional properties of the statistics are typically problematic to handle (Cuevas et al. (2004), Politis and Romano (1994), Cuevas and Fraiman (2004)). In particular, the work of Cuevas and Fraiman (2004) derives a result of bootstrap validity for functional statistics defined on differentiable operators. Observe that the vec() operator and the projection maps $S_\Theta(t), S_\beta$ and S_Φ appearing in (20) satisfy the regularity conditions required by the result of Cuevas and Fraiman (2004), and so does their composition, so that we are justified in the use of a bootstrap method to get a valid estimate of the distribution of the functional coefficients. We thus resort to the following resampling scheme.

1. Estimate $\hat{\boldsymbol{\beta}}(t)$ of the regression model $\boldsymbol{y}(t) = X(t)\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t)$, as from Section 3;

2. Evaluate the residuals $\hat{\boldsymbol{\epsilon}}(t) = \boldsymbol{y}(t) - X(t)\hat{\boldsymbol{\beta}}(t)$;

3. Randomly generate a bootstrap sample $\hat{\boldsymbol{\epsilon}}^*$ from the empirical distribution of the residuals and define the new pairings $\{(y_1^*, \boldsymbol{x}_1), \ldots, (y_n^*, \boldsymbol{x}_n)\}$ as

$$y_i^*(t) = \boldsymbol{x}_i^T(t)\hat{\boldsymbol{\beta}}(t) + \hat{\epsilon}_i^*(t).$$

4. Estimate $\hat{\boldsymbol{\beta}}^*(t)$ of the regression model $\boldsymbol{y}^*(t) = X(t)\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t)$, with $\boldsymbol{y}^*(t) = (y_1^*, ..., y_n^*)^T$,

5. Repeat (3) and (4) for B times, with M sufficiently large.

The empirical distributions of the sample $\{\hat{\boldsymbol{\beta}}^*(t)\}_{m=1}^M$ are visualized via functional boxplots. The amplitude of the fences is considered a reliable estimate of the confidence that one globally has on the true coefficients.

We point out that more compound techniques relying on parametric bootstrap to do simultaneous inference for functional parameters are present in the literature. Degras (2011) considers a simple function-on-scalar regression and proposes a parametric bootstrap that builds simultaneous confidence bands around the estimate of the functional coefficient. Chang et al. (2017) extend this work by proposing a wild bootstrap methodology to handle regression with multiple covariates and errors that are non-normal and heterogeneous. The simulation-based method of Degras (2017) provides theory, method and implementation of simultaneous confidence bands for functional statistics and parameters. Cao et al. (2012) associate a spline estimator for the mean function of dense functional data to a simultaneous confidence band which is asymptotically correct. Employing such methodologies is outside the scope of the present work and may be object of future research, aiming to accurately identify simultaneous confidence bands for coefficient estimates or for other functional statistics.

4 Simulation study

This section is devoted to the validation through a simulation study of the weighted methodology presented above. Synthetic partially observed data are simulated so as to capture the characteristics of variability of the data of the case study. Monte Carlo simulations are employed to test the soundness of the weighted against the non-weighted approach with respect to three aspects: *(i)* the effectiveness in reducing the impact that the method adopted to reconstruct the missing trajectories of the curves have on the final coefficients estimates, *(ii)* the effect that their shape has on the results of the analysis, *(iii)* the property of stabilization of the coefficients estimates when the fraction of partially observed data increases.

4.1 Simulation of partially observed functional data

Data are generated according to model

$$y_i(t) = \beta_0(t) + \beta_1(t)x_{1i} + \beta_2(t)x_{2i}(t) + \epsilon_i(t), \tag{22}$$

where ϵ_i are independent realizations of a zero-mean stochastic process. The inclusion of a scalar and a functional covariate in (22) allows us to test the soundness of the weighted methodology both for a *function-on-scalar* and a *concurrent* linear regression model. The scheme adopted for the simulation of the covariates and the regression coefficients entering (22) is motivated by our intention to capture the main modes of variability of the functional data of the case study that propelled this work. Specifically, the rationale is as follows:

1. Consider a scalar covariate $\boldsymbol{z}_1$ and a functional covariate $\boldsymbol{z}_2(t)$ among those entering the functional form (2), then fit via unweighted least squares the regression model $\log_{10}(SA_i(t)) = \alpha_0(t)+\alpha_1(t)z_{1i}+\alpha_2(t)z_{2i}(t)+\xi_i(t)$ to obtain the estimates $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2$ and $\hat{\xi}_1, \dots, \hat{\xi}_n$,

2. Set $\beta_k = a_k\hat{\alpha}_k$ so that the magnitudes of $\beta_0, \beta_1, \beta_2$ are comparable, thus preventing the effect of a single covariate from being preponderant compared to that of the others,

3. Sample the covariates $x_{1i}, x_{2i}(t)$ and the regression residuals $\epsilon_i(t)$ (as specified below),

4. Set $y_i(t) = \beta_0(t) + \beta_1(t)x_{1i} + \beta_2(t)x_{2i}(t) + \epsilon_i(t)$,

5. Evaluate $(y_i(t_1), \dots, y_i(t_N))$ on the sampling instants $(t_1, \dots, t_N)$,

6. Sample the smoothing error $e_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ for $j = 1, \dots, N$ and set $\tilde{y}_i(t_j) = y_i(t_j) + e_{ij}$.

It is necessary to specify how the sampling of the covariates and of the regression error is done in step 3. The scalar covariate is simply generated from the univariate normal distribution whose mean and variance are the mean and the variance of the original covariate $\boldsymbol{z}_1$, namely $x_{1,i} \sim \mathcal{N}(\mu_1, \sigma_1^2)$ for $i = 1, \dots, n$. The argument for the functional covariate is only slightly more compound. Let $\boldsymbol{b}(t) = (b_1(t), \dots, b_L(t))^T$ be the basis on which $\boldsymbol{z}_2$ is defined, so that $z_{2i}(t) = \boldsymbol{c}_i^T\boldsymbol{b}(t)$. Then the synthetic covariate x_{2i} is defined on the same space of basis functions and is built by sampling new random coefficients $\tilde{\boldsymbol{c}}_i$ from the multivariate normal distribution whose mean vector and covariance operator are the mean vector and the covariance matrix of the $\boldsymbol{c}_1, \dots, \boldsymbol{c}_L$, *i.e.* $\tilde{\boldsymbol{c}}_i \sim \mathcal{N}_L(\bar{\boldsymbol{c}}, \Sigma_{\boldsymbol{c}})$. The rationale behind the generation of the regression residuals ϵ_i is to define them in the space generated by the $K = 2$ harmonics φ_k that account for about 99% of the variability of the $\hat{\xi}_1, \dots, \hat{\xi}_n$ obtained in step 1. Then the new scores s_{i1}, s_{i2} are sampled from the bivariate normal distribution centred in zero and with covariance the diagonal matrix of the first two eigenvalues of the FPC decomposition, namely $s_{i1}, s_{i2} \sim, \mathcal{N}_2(\boldsymbol{0}, \Lambda_2)$. Eventually, the novel regression residuals are built by setting $\epsilon_i(t) = \sum_{k=1,2} s_{ik}\varphi_k(t)$. We point out that as the regression of the original $\log_{10}(SA)$ curves on only two of the seismic covariates leaves much of the variability into the regression error, the estimated residuals $\hat{\xi}_1, \dots, \hat{\xi}_n$ obtained in step 1 are previously scaled of a factor 10 so that they do not impact too much on the synthetic observations.

Simulation of the observation domain We want each simulated curve to be observed over a curve-specific domain. To this end, curve y_i is coupled with two independent random variables $U_i \sim \text{Unif}([1.5, 3.5])$ and $P_i \sim \text{Be}(p)$, and the corresponding instant of right censoring T_i is defined as

$$T_i = \begin{cases} 3.5, & P_i = 0 \\ U_i, & P_i = 1 \end{cases}.$$

This sampling approach ensures that a fraction p of data are partially observed with censoring instant T_i in the interval $[1.5, 3.5]$. Parameter p is set to 0.4 in the first two sets of simulations. The third set, reported in Appendix B.1, tests the impact of the choice of p on the results.

Definition of the weights The definition of a functional weight should reflect the reliability that we have on a functional datum along the domain. The full reliability associated to the observed values of a curve is represented by a weight set to 1. As we move more and more away from the last observed value, the reliability on the extrapolated values is corrected to become continuously smaller. A logistic function is a convenient choice to achieve a decrease in confidence from 1 to small values.
Suppose that the i-th functional datum is observed up to an instant $t = T_i$. Then we define

$$w_i(t) = \begin{cases} 1\ , & t \leq T_i \\ \frac{1}{1+e^{(t-\mu_i)\alpha_i}} + c_i\ , & t > T_i \end{cases}, \tag{23}$$

where we set $\alpha_i = a\sigma_{T_i}$, σ_{T_i} being the empirical standard deviation of the observed curves at T_i and $a > 0$ a hyperparameter. The corrective term $c_i = 1 - \frac{1}{1+e^{(T_i-\mu_i)\alpha_i}}$ guarantees continuity of the weight at T_i. The location parameter μ_i identifies the inflection point, which is curve-specific and set equal to half the length of the missing domain. Notice that, once μ_i is fixed, the shape of the weight is completely determined by the scale parameter α_i, which controls the rate of decay of the logistic function. The larger α_i, the more abrupt is the decrease of the weight to 0. Here, large values of α_i are associated to a great variability of the complete curves in T_i, meaning that if a record is censored at a period characterized by large variability of the observed curves, then the confidence associated to the reconstruction quickly falls to 0. Hyperparameter a is intended to regulate the impact of σ_{T_i} on the decay. The joint effect of a and the correction c_i results in a weight that takes larger values and shows more gradual decay when σ_{T_i} is small, and that decreases rapidly to zero when σ_{T_i} is large.
The first and the third batteries of simulations maintain a fixed and equal to 10, while the second set assesses the impact of the choice of the setting of a on the results. Figure 3 shows a simulated partially observed curve, its reconstruction with the method proposed by Kraus (2015) and the associated logistic weight.

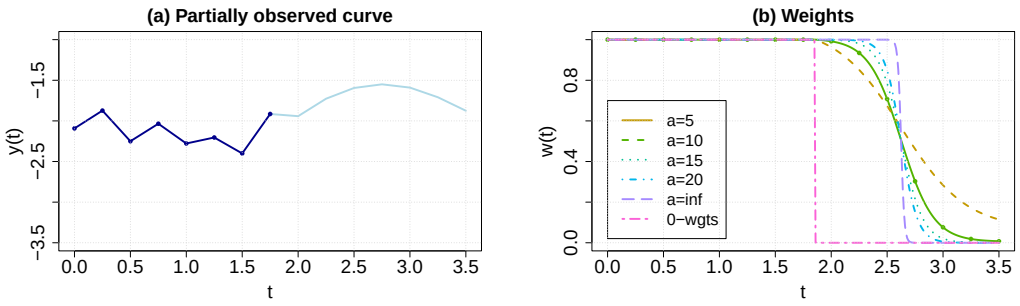

Figure 3: (a) Partially observed curve (blue) and its reconstruction (light blue). (b) Logistic weight associated to the partially observed curve, where the hyperparameter for the decay is set $a = 10$.

4.2 Validation of the weighted analysis

The performance of the weighted against the non-weighted analysis is assessed in terms of the Mean Squared Error (MSE) and the variance of the estimators $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$, defined, for $j = 0, 1, 2$, as

$$\begin{aligned} \text{MSE}(\hat{\beta}_j) &= \mathbb{E}\left[\left|\left|\hat{\beta}_j - \beta_j\right|\right|_2^2\right], \\ \text{Var}(\hat{\beta}_j) &= \mathbb{E}\left[\left|\left|\hat{\beta}_j - \mathbb{E}\left[\hat{\beta}_j\right]\right|\right|_2^2\right]. \end{aligned} \tag{24}$$

In the practice, MSE and variance of each $\hat{\beta}_j$ are extracted from the empirical distributions of the $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$, obtained via Monte Carlo simulation with B=100 repetitions, by approximating the population means in (24) with their finite sample counterparts.
For $b = 1, \ldots, B$, the simulation does: *(i)* generate a sample of fully and partially observed functional data, *(ii)* reconstruct the right-censored curves, *(iii)* define the logistic weights, *(iv)* smooth the discrete observations, *(v)* estimate the regression coefficients $(\hat{\beta}_0^b, \hat{\beta}_1^b, \hat{\beta}_2^b)$. Depending on the approach considered, steps *(iv)* and *(v)* are carried out including or non including the weights in the estimation criteria.
The comparison is carried out in three series of simulations. First, we evaluate the impact that the reconstruction methodology has on the performance of the entire analysis, for both the weighted and unweighted approaches. Secondly, the rate of decay of the logistic weights is varied within a set of values, to check whether there is an optimal definition of the weight system in the range between two extreme options: opt for an unweighted analysis or force the weights to zero right after the last valid observation instant of a curve. Finally, we vary the fraction of partially observed curves over the total sample size to check if the application of the weighted method positively impacts the stability and the accuracy of the estimators in scenarios where the missing information increases.

Robustness to the reconstruction methods As mentioned in the introduction to this work, the reconstruction of partially observed functional data can be

performed exploiting different methodologies. Here three reconstruction strategies are taken into account. For the sake of clarity, they are referred to with acronyms and their working principle is briefly recalled below[1].

Kraus: Reconstruction of the missing trajectory made by a Hilbert-Schmidt operator, estimated via a functional linear ridge regression as it is accurately reported in Kraus (2015). The penalization parameter entering the ridge regression is selected via generalized cross-validation at each step of the Monte Carlo simulation.

KL-PC: Functional completion made by a reconstruction operator, which estimates the principal components of the curve over the entire domain. Then the missing part of the trajectory is reconstructed resorting to the best basis property as the truncated sum of the first K principal components (Kneip and Liebl, 2020). The number K of principal components entering the sum is selected via generalized cross-validation.

KL-AL: It refers to the same procedure as KL-PC, but operates a preliminary step of non-parametric smoothing on the observed parts of the curves. Since in general we are not guaranteed continuity at the boundary of the non-parametric estimate with the reconstructed trajectory, the reconstruction is corrected in order to recover continuity (thus the term ALign).

When comparing the weighted and non-weighted approaches over different reconstruction methods, we expect the inclusion of the weights to reduce the differences among their performances, since they enter the estimation criteria for the smoothing and the regression coefficients by diminishing the impact of the reconstructed trajectories on the resulting estimates. The results displayed by the boxplots in Figure 4 are in agreement with this intuition. The empirical distributions of the squared L^2 distances of $\hat{\beta}_j^b$ from its true value are closer to each other in the weighted analysis than in the unweighted analysis, for each of the three regression coefficients. Additionally, we observe that the weighted methodology is effective in lowering both the variance and the bias (Table 1) of the point estimators, meaning that the analysis benefits in terms of stabilization of the estimators and of estimation accuracy.
We mention that, since method KL-AL is the best performing in both the weighted and non-weighted approaches, it is adopted as reconstruction method for the following two batteries of simulations.

Impact of the weights definition This series of simulations is intended to assess the extent to which the shape of the weights affects the results of the analysis. Their profile is modified by varying the rate of decay of the logistic function, namely by moving the hyperparameter a that enters (23) within the set of values $\{5, 10, 15, 20, \infty\}$. Notice that $a = \infty$ corresponds to the limit condition at which the

[1]The implementation of all three methods considered is available in the `R package` ReconstPoFD, which can be installed from the GitHub account of Dominik Liebl: https://github.com/lidom/ReconstPoFD.

Table 1: Comparison of bias$^2(\beta_j)$ among the adopted reconstruction methods.

Coefficient	Kraus: wgt	KL-PC: wgt	KL-AL: wgt	Kraus	KL-PC	KL-AL
β_0	0.002	0.003	0.002	0.011	0.030	0.011
β_1	0.005	0.005	0.003	0.024	0.028	0.017
β_2	0.000	0.001	0.001	0.004	0.009	0.004

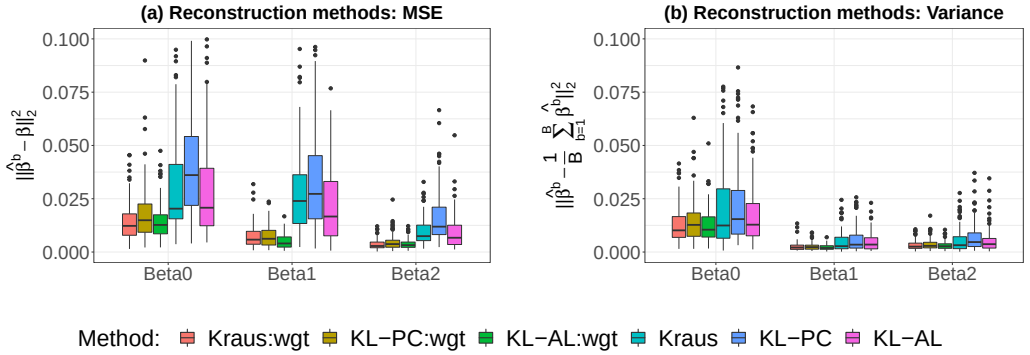

Figure 4: Boxplots of the empirical population of $\left|\left|\hat{\beta}_j^b - \beta_j\right|\right|_2^2$ (left) and of $\left|\left|\hat{\beta}_j^b - \overline{\hat{\beta}_j^b}\right|\right|_2^2$ (right), for every reconstruction method considered and for every regression coefficient β_0, β_1 and β_2. The finite sample means of the empirical populations estimate the population MSE and variance of the coefficients estimators.

weight is a step function, taking value one up to the middle of the missing domain and falling to zero right after that instant. In the computations, this condition is obtained by setting $a = 100$. Other than $a = \infty$, two other limit conditions are considered, which are the non-weighted case – *i.e.* the weights have constant value 1 – and the case denoted as 0-weights, where the weights are step functions falling to a small positive value (set to 10^{-6} rather than to 0 for computational reasons) at the censoring instant. Results are reported in Figure 5. As we may notice from the figure on the right, the boxplots relative to all three coefficients estimators present a minimum of the estimated MSE in correspondence of $a = 10$. The minimum variance is observed at $a = 5$ for β_0 and β_2, and at $a = 10$ for β_1. Both the unweighted case and the case 0-weights, although not corresponding to an increase in the variance, exhibit large bias, which manifests in a strong increase of the MSE. The predictive effectiveness of the model corresponding to weights with $a = 10$ is confirmed by a Leave-One-Out cross-validation (LOOCV) on a realization of $n = 100$ synthetic data. Figure 6 displays the empirical distributions of the L^2 norm of the functional prediction error made by the models on the true curves, assessed via LOOCV. Indeed, we notice that the minimum of the estimated MSE corresponds to $a = 10$. These results jointly show that there is a trade-off between the two alternatives of associating full reliability to the curve as a whole and of completely neglecting the information of the missing trajectory. Specifically, the boxplots suggest that a solution to the trade-off lies in the use of a well-calibrated system of weights that conveniently modulates the importance associated to a functional datum along the

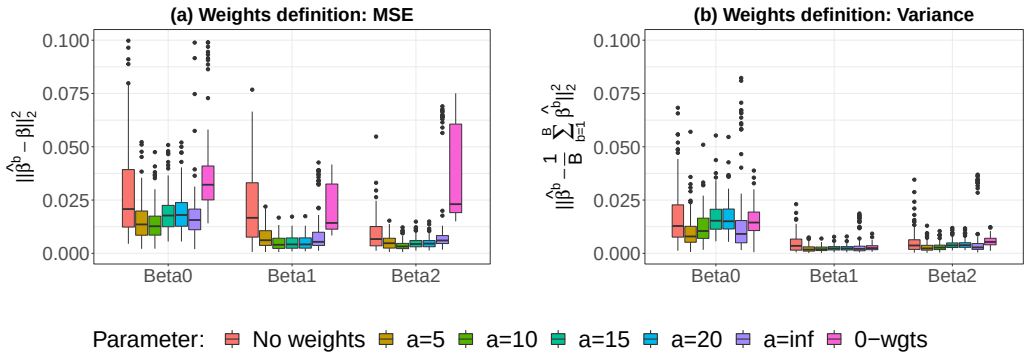

Figure 5: Boxplots of the empirical population of $\left|\left|\hat{\beta}_j^b - \beta_j\right|\right|_2^2$ (left) and of $\left|\left|\hat{\beta}_j^b - \overline{\hat{\beta}_j^b}\right|\right|_2^2$ (right), for every regression coefficient β_0, β_1 and β_2 and for different values of the hyperparameter a entering the weights definition. Larger values of a correspond to greater rates of decay of the weights; $a = \infty$ corresponds to the limit case in which the weights are forced to be zero right after the last observed instant.

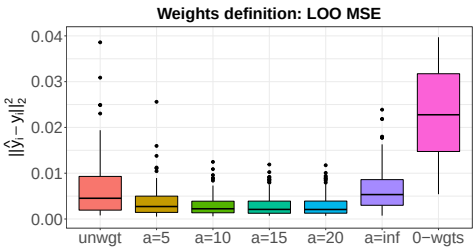

Figure 6: Boxplots of the empirical population of $\left|\left|\hat{y}_i^b - y_i\right|\right|_2^2$, for different values of the hyperparameter a entering the weights definition, obtained via LOO cross-validation.

observed and unobserved segments of the domain.

5 Case study

This section is devoted to the assessment of the efficacy of the weighted functional methodology when it is employed in the fitting of the functional ground motion model in equation (2). The performance of the weighted approach is compared to that of the ordinary least squares adopted for the fitting of the ITA18 scalar model (1) at 37 periods of observation of the acceleration spectrum. We refer to Appendix C.1 for the analysis of correlation between the covariates, which reveals the presence of collinearity between the predictor variables in (2) and, possibly, the ineffectiveness of the estimation procedure in separating the individual effects of the predictors on the response. Although collinearity could in principle be fixed resorting to techniques of model reduction developed in the FDA context (*e.g.* Horváth and Kokoszka (2012), Ramsay and Silverman (2005), Mehrotra and Maity (2022)), the physical interpretability of the regressors motivates the choice of keeping the

functional form (2) unchanged, as it allows to discuss on the results in seismological terms and eases the comparison with ITA18 and similar state-of-the-art ground motion models (*e.g* Bindi et al. (2014), Boore et al. (2014), Kotha et al. (2022)). Yet, note that regularization is performed through the introduction of the penalization term discussed in Section 3.2. A penalization in the fitting criterion hence not only permits to obtain estimates of the functional coefficients from a finite number of observations, but also controls the side effects of collinearity by reducing the variability associated to the estimates. This prompts us to pay special attention, along with the choice of the weights and the reconstruction method, to the selection of penalty hyperparameters that enter the estimation process. Accordingly, the calibration of the functional model occurs in three steps, that select *(i)* the penalty parameters, *(ii)* the parameter a entering the definition of the weights, and *(iii)* the method for the imputation of the partially observed curves.

5.1 Model calibration

The three steps of calibration of the functional ITA18 model exploit estimates of the prediction mean squared errors obtained via a cross-validation procedure. As model (2) works under the ergodic assumption (Anderson and Brune, 1999), we adopt a partitioning strategy that forces data in the training and test to be related to independent events, hence reducing the underestimation of the MSE.

Calibration of the penalization parameters The calibration of the penalization parameters is conducted on the dataset restricted to the fully observed curves and resorting to an unweighted analysis. This implies working under the reasonable assumption that the features of regularity of the regression coefficients can be inferred from the fully observed curves alone, which still account for 75% of the data. Since there is no particular reason to believe that every functional coefficient should be characterised by the same level of regularization, a distinct penalization parameter is selected for each coefficient. The strategy adopted for the calibration of $\lambda_1, \ldots, \lambda_9$ then stems from the computational burden of performing a grid search in a 9-dimensional space. We opt for a greedy approach that iteratively moves one parameter in a range of values, while maintaining the others as fixed, and sets it to the value corresponding to the minimum prediction error estimated through a cross-validation. We refer to Appendix C.2 for a more detailed description of the approach and the obtained results.

Choice of parameter a in weights definition Here, the different models are compared by means of a global measure of inaccuracy in the prediction of the observed values of the curves. Inaccuracy is quantified as the sum of the squared distances from the true and the fitted discrete ordinates of a curve, namely

$$\hat{\epsilon}_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} \left(y_{ij} - \hat{y}_i(t_j) \right)^2 , \tag{25}$$

Table 2: Empirical MSE and the associated variability for all possible values of a.

	unweighted	$a = 5$	$a = 10$	$a = 15$	$a = 20$	$a = \infty$	0-weights
MSE	0.1194	0.1193	0.1192	0.1192	0.1191	**0.1189**	0.1193
σ	0.0227	0.0227	0.0227	0.0227	0.0227	**0.0222**	0.0226

Table 3: Empirical MSE and σ for extrapolation and KL-AL.

	extrapolation	KL-AL
MSE	**0.1189**	0.1190
σ	**0.0220**	0.0222

where N_i is the number of observed ordinates of the functional datum y_i and is included in (25) as a normalization factor. We evaluate the expectation of quantity (25) as the sample mean resulting from a 10-fold cross-validation procedure. We seek for the optimal parameter a within the range of values tested in the simulation study, and contextually assess whether the weighted analysis performs better than the unweighted analysis. The results displayed in Table 2 do not reveal any significant difference in the predictive performances of the methodologies. Nonetheless, the MSE exhibits a decreasing trend from the unweighted analysis (*i.e.* weights equal 1 everywhere) up to case $a = \infty$, and then increases in the case of 0 weights. This motivates the choice of $a = \infty$ as the system of weights to be introduced in the analysis.

Selection of the reconstruction method The last step of calibration consists in the selection of the method adopted for the imputation of the missing trajectories of the acceleration spectra. The optimal reconstruction method resulting from the simulation study, which we refer to with the acronym KL-AL (see Section 4.2), is tested against a naive reconstruction suggested by the profiles of RotD50 displayed in Figure 2a. The idea is to linearly extrapolate each incomplete curve from its last observed value up to $T = 10\ s$, with a slope that captures the descending trend exhibited by the complete records in the right end of the domain. The slope of the extrapolating line is set equal to the mean over all complete records of the slopes of the straight lines that interpolate the complete records at $\bar{T}$ and at $T = 10\ s$. Again, the comparison is carried out by means of the expectation of quantity (25), which is estimated as the sample mean resulting from a 10-fold cross-validation. The result is reported in Table 3. We notice that the 10-fold cross-validation does not highlight any significant difference in the predictive performances of the two methods. Since the point minimum of the MSE is exhibited by the extrapolation method, we are lead to adopt it to perform data reconstruction in the analyses that follow.

5.2 Comparison with scalar ITA18

Figure 7 shows the estimates of the regression coefficients b_1, b_2, c_1, c_2, c_3 and k, each one associated to the functional boxplot of a bootstrap sample of dimension $B = 1000$, generated from its empirical distribution according to the procedure detailed in Section 3.3. We see the scatter of the sample around the point functional estimate as measure of its simultaneous variability over the domain. The smaller the scatter, the lower is the uncertainty associated to the estimate. All functional coefficients estimates generally follow the trends of the scalar estimates while displaying a more regular behaviour. Coefficients b_1 and b_2, plotted in Figure 7a and 7b, capture the linear dependence of ground motion on low magnitudes and high magnitudes respectively. Both have a positive impact on spectral acceleration that grows in the interval $[0, 1\ s]$ and then remains more or less constant until $T = 10\ s$. Differently from the scalar estimate, coefficient b_1 takes high constant values at long periods. Notice that the detachment between the estimates accentuates where the fraction of missing values increases. Here the functional weighted approach impacts the results, with respect to the scalar analysis that neglects the unobserved curves. Figure 7c and 7d display the coefficients related to the geometric attenuation of ground motion with distance, namely c_1 and c_2. At all periods, c_2 captures the linear decay of the spectral acceleration with d_{JB}. Coefficient c_1 complements c_2 in capturing the magnitude dependence of geometric spreading due to finiteness of large magnitude ruptures. As expected, c_1 takes positive values to simulate the more gradual decay in near-source distances from large ruptures (Kotha et al., 2022). The functional estimate for c_1, however, moves away from the scalar estimate at long periods. We notice that lower values of c_1 at long periods are compensated by higher values of b_1, confirming the difficulty of the model to separate the single effects of the predictors on the response. We point out that the scalar least squares performed in Lanzano et al. (2019) do not operate any form of regularization to deal with collinearity. In Figure 7e, c_3 accounts for the exponential decay of ground motion with distance, that is the anelastic attenuation. As we may see from the graph, anelastic attenuation affects ground motion at short periods, and its effect vanishes at longer periods. Positive values taken by c_3 at the right end of the domain could be an issue, as they would indicate an unphysical exponential increase with distance. Note however that the uncertainty associated with the positive estimates of c_3, as evidenced by the functional boxplot in Figure 7c, suggests that these estimates may not be significantly different from zero. A further account of the significance of the coefficients will be the scope of future work (see also Section 3.3). Finally, coefficient k in Figure 7f accounts for the negative scaling of ground motion with the shear-wave velocity. A common issue with this coefficient lies in its instabilities at short periods, where it may get very close to zero or even be positive, conversely to what is observed at all other periods. In our case, the instability is not pronounced and k remains significantly negative for all T. A brief comment on estimates f_1 and f_2 is left in the Appendix C.3.

Figure 8a shows a comparison of the point-wise mean squared errors, resulting from a 10-fold cross-validation. The functional model operates a regularization in the

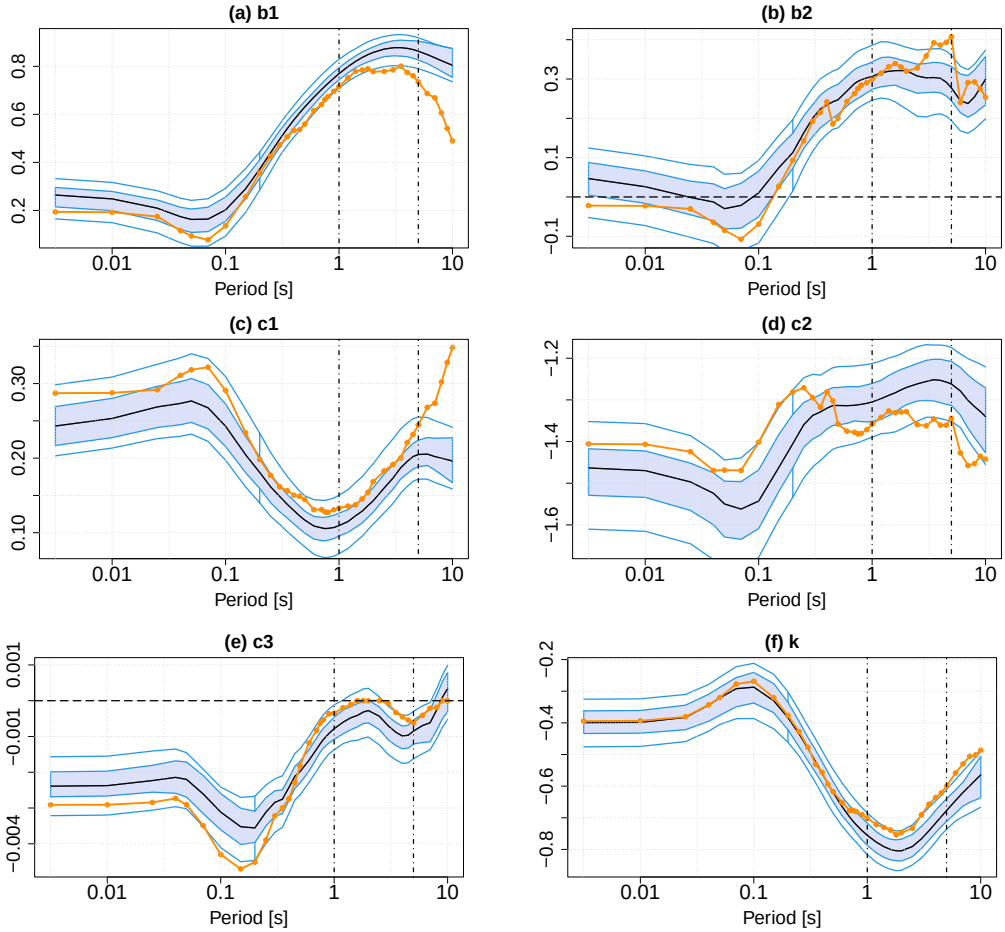

Figure 7: Functional boxplots of the estimated functional regression coefficients and comparison with the estimates of ITA18 (orange). The black continuous lines represent the point estimates of the coefficients. The azure bands are the fences of the functional boxplots. The dashed horizontal line marks zero. The vertical lines mark the points which correspond to the last instants where 100% and $> 90\%$ of the curves are observed.

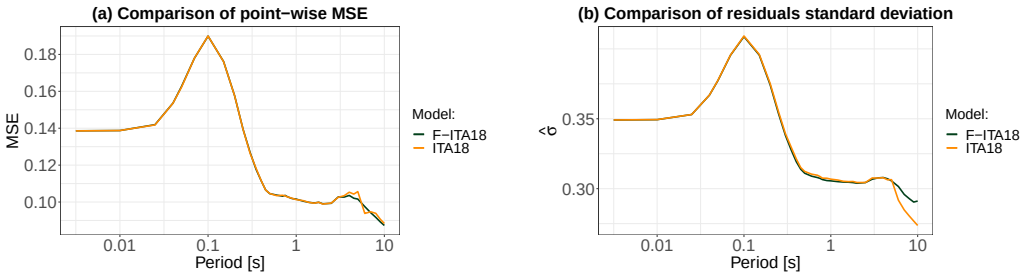

Figure 8: Comparison of the model performance between Scalar (orange) and Functional (green) ITA18, in terms of: (a) Point-wise Mean Squared Error, (b) Point-wise residual standard deviation $\hat{\sigma}$.

estimation process, which also reflects on the shape of the MSE along periods. We note that regularization of the MSE with respect to the scalar approach occurs at the right end of the domain, where the scalar least squares are performed only on the fully observed curves. However, prediction performances in the fully observed domain appear similar, which is likely due to the functional shape of the weight function ($a = \infty$).
Figure 8b reports the comparison between the estimated point-wise residuals standard deviations of the functional and of the scalar models. Since the number of observed curves $n = 5568$ is large with respect to that of the sampling instants $N = 37$ (Ramsay and Silverman (2005), Section 4.6.2), we estimate the point-wise residual relying on the diagonal elements of

$$\hat{\Sigma} = \frac{1}{n-q-1}\hat{\mathcal{E}}^T\hat{\mathcal{E}},$$

where q is the number of covariates entering (2). While the two models share the same trend of $\hat{\sigma}$ at short and medium periods, scalar ITA18 exhibits lower values of residuals standard deviation at long periods, precisely where the effect of the data censoring becomes more relevant.

6 Discussion and conclusions

The present work proposes a novel approach to the analysis of partially observed functional data, maintaining a thorough focus on the application context that motivates the work. The proposed methodology extends the classical penalized smoothing and penalized concurrent regression to the inclusion of weights, which enter the analysis by reducing the impact that the reconstructed parts of partially observed curves have on the final estimates. The soundness of the weighted analysis is tested in a simulation study, which highlights the effectiveness of the method in *(i)* reducing the variance and the mean-square error of the coefficients estimators with respect to the unweighted analysis, *(ii)* improving the predictive performances of the analysis, *(iii)* mitigating the impact of the adopted reconstruction methods on the resulting estimates.

The adoption of the weighted functional methodology for the analysis of GMMs introduces multiple innovative features in the context of ground motion prediction equations. By embedding the analysis in the context of partially observed functional data and by reconstructing the missing ordinates of intensity measures, the method circumvents a massive loss of information, while preserving the functional analysis over the entire range of vibration periods of interest. The functional embedding naturally takes account of the cross-correlation between the spectral ordinates and provides continuous estimates over the considered range of periods. Besides, the method is shown to operate an intrinsic smoothing and stabilization of the coefficients estimates and of the spectral predictions.
Future developments of the weighted method go in multiple directions. Suitable definitions of weights may be driven by correlations between different time instants, *e.g.* by exploiting the estimate of the covariance operator. The estimation of the variability associated to the coefficients estimates, which is currently done including the uncertainties related to the weighted smoothing and weighted regression, may be refined to account also for the uncertainty that propagates from reconstruction. A more rigorous analysis of simultaneous uncertainty, additionally, might stem from one of the methods mentioned in Section 3.3 and allow one to draw statistically-based conclusions on the significance of the functional coefficients estimates. Concerning the applicability of the approach, one may consider to employ the methodology in the more general context of fully observed curves that exhibit various degree of uncertainty over the domain.
Further extensions of our ergodic functional ground motion model move towards the construction of nonergodic and functional seismic-shaking maps. On the one hand, period-continuous systematic corrective terms may be estimated with a functional mixed-effect model that account for site- and event-related random effects. This latter model, too, may be generalized to work with partially observed data and to the inclusion of functional weights. On the other hand, the estimate for the median intensity measure provided in this work naturally combines with the functional geostatistical model for the residuals proposed in Menafoglio et al. (2020), as they jointly set up a convenient tool for the construction of ground motion maps in a fully functional context.

References

Anderson, J. G. and Brune, J. N. (1999) Probabilistic Seismic Hazard Analysis without the Ergodic Assumption. *Seismological Research Letters*, **70**, 19–28.

Bindi, D., Massa, M., Ameri, G., Pacor, F., Puglia, R. and Augliera, P. (2014) Pan-european ground-motion prediction equations for the average horizontal component of pga, pgv, and 5%-damped psa at spectral periods up to 3.0 s using the resorce dataset. *Bulletin of Earthquake Engineering*, **12**.

Bindi, D., Pacor, F., Luzi, L., Puglia, R., Massa, M., Ameri, G. and Paolucci, R. (2011) Ground motion prediction equations derived from the italian strong motion database. *Bulletin of Earthquake Engineering*, **9**, 1899–1920.

Bommer, J. J., Douglas, J. and Strasser, F. O. (2003) Style-of-faulting in ground-motion prediction equations. *Bulletin of Earthquake Engineering*, **1**, 171–203.

de Boor, C. (1978) *A Practical Guide to Splines*. New York: Springer Verlag.

Boore, D. M. (2010) Orientation-independent, nongeometric-mean measures of seismic intensity from two horizontal components of ground motion. *Bulletin of the Seismological Society of America*, **100**, 1830–1835.

Boore, D. M., Stewart, J. P., Seyhan, E. and Atkinson, G. M. (2014) Nga-west2 equations for predicting pga, pgv, and 5% damped psa for shallow crustal eartquakes. *Earthquake Spectra*, **30**, 1057–1085.

Bradley, B. A. (2011) Empirical correlation of pga, spectral accelerations and spectrum intensities from active shallow crustal earthquakes. *Earthquake Engineering Structural Dynamics*, **40**, 1707–1721.

Cao, G., Yang, L. and Todem, D. (2012) Simultaneous inference for the mean function based on dense functional data. *Journal of Nonparametric Statistics*, **24**, 359–377.

Chang, C., Lin, X. and Ogden, R. T. (2017) Simultaneous confidence bands for functional regression models. *Journal of Statistical Planning and Inference*, **188**, 67 – 81.

Cuevas, A., Febrero, M. and Fraiman, R. (2004) An anova test for functional data. *Computational Statistics and Data Analysis*, **47**, 111–122.

Cuevas, A. and Fraiman, R. (2004) On the bootstrap methodology for functional data. *COMPSTAT 2004 — Proceedings in Computational Statistics*, 127–135.

Davies, P. and Meise, M. (2008) Approximating data with weighted smoothing splines. *Journal of Nonparametric Statistics*, **20**, 207–228.

Degras, D. A. (2011) Simultaneous confidence bands for functional regression models. *Statistica Sinica*, **21**, 1735–1765.

— (2017) Simultaneous confidence bands for the mean of functional data. *WIREs Computational Statistics*, **9**, e1397.

Douglas, J. (2003) Earthquake ground motion estimation using strong-motion records: a review of equations for the estimation of peak ground acceleration and response spectral ordinates. *Earth-Science Reviews*, **61**, 43–104.

Douglas, J. and Edwards, B. (2016) Recent and future developments in earthquake ground motion estimation. *Earth-Science Reviews*, **160**, 203–219.

Horváth, L. and Kokoszka, P. (2012) *Inference for Functional Data with Applications*, chap. 2. Springer.

Huang, C. and Galasso, C. (2019) Ground-motion intensity measure correlations observed in italian strong-motion records. *Earthquake Engineering & Structural Dynamics*, **48**, 1634–1660.

Kamai, R., Abrahamson, N. A. and Silva, W. J. (2014) Nonlinear horizontal site amplification for constraining the nga-west2 gmpes. *Earthquake Spectra*, **30**, 1223–1240.

Kneip, A. and Liebl, D. (2020) On the optimal reconstruction of partially observed functional data. *The Annals of Statistics*, **48**, 1692–1717.

Kotha, S. R., Bindi, D. and Cotton, F. (2016) Partially non-ergodic region specific gmpe for europe and middle-east. *Bulletin of Earthquake Engineering*, **14**, 1245–1263.

Kotha, S. R., Weatherill, G., Bindi, D. and Cotton, F. (2022) Near-source magnitude scaling of spectral accelerations: Analysis and update of kotha et al. (2020) model. *Bulletin of Earthquake Engineering*, **20**, 1343–1370.

Kraus, D. (2015) Components and completion of partially observed functional data. *Journal of the Royal Statistical Society*, **77**, 777–801.

Kraus, D. and Stefanucci, M. (2018) Classification of functional fragments by regularized linear classifiers with domain selection. *Biometrika*, **106**, 161–180.

Lanzano, G., Luzi, L., Pacor, F., Felicetta, C., Puglia, R., Sgobba, S. and D'Amico, M. (2019) A revised ground-motion prediction model for shallow crustal earthquakes in italy. *Bulletin of the Seismological Society of America*, **109**, 525–540.

Lanzano, G., Sgobba, S., Caramenti, L. and Menafoglio, A. (2021) Ground-motion model for crustal events in italy by applying the multisource geographically weighted regression (ms-gwr) method. *Bulletin of the Seismological Society of America*, **111**, 3297–3313.

Lanzano, G., Sgobba, S., Luzi, L., Puglia, R., Pacor, F., Felicetta, C., D'Amico, M., Cotton, F. and Bindi, D. (2018) The pan-european engineering strong motion (esm) flatfile: compilation criteria and data statistics. *Bulletin of Earthquake Engineering*, **17**, 561–582.

Mehrotra, S. and Maity, A. (2022) Simultaneous variable selection, clustering, and smoothing in function-on-scalar regression. *The Canadian Journal of Statistics*, **50**, 180–199.

Menafoglio, A., Sgobba, S., Lanzano, G. and Pacor, F. (2020) Simulation of seismic ground motion fields via object-oriented spatial statistics with an application in northern italy. *Stochastic Environmental Research and Risk Assessment*, **34**, 1607–1627.

Newmark, N. and Hall, W. (1982) Earthquake spectra and design. *Earthquake Engineering Research Institute, Oakland, California, U.S.A.*

Pintore, A., Speckman, P. and Holmes, C. C. (2006) Spatially adaptive smoothing splines. *Biometrika*, **93**, 113–125.

Politis, D. N. and Romano, J. P. (1994) Limit theorems for weakly dependent hilbert space valued random variables with application to the stationary bootstrap. *Statistica Sinica*, **4**, 461–476.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis*. Springer.

Russo, E., Felicetta, C., D'Amico, M. C., Sgobba, S., Lanzano, G., Mascandola, C., Pacor, F. and Luzi, L. (2022) Italian accelerometric archive (itaca), version 3.2. URL: https://itaca.mi.ingv.it/ItacaNet_32/.

Sabetta, F., Pugliese, A., Fiorentino, F., Lanzano, G. and Luzi, L. (2021) Simulation of non-stationary stochastic ground motions based on recent italian earthquakes. *Bulletin of Earthquake Engineering*, **19**, 3287–3315.

Stefanucci, M., Sangalli, L. and Brutti, P. (2018) Pca-based discrimination of partially observed functional data, with an application to aneurisk65 data set. *Statistica Neerlandica*, **72**, 246–264.

Worden, C. B., Thompson, E. M., Baker, J. W., A., B. B., Luco, N. and Wald, D. J. (2018) Spatial and spectral interpolation of ground-motion intensity measure observations. *Bulletin of the Seismological Society of America*, **108**, 866–875.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100**, 577–590.

A Methods

A.1 Penalized Weighted Functional Least Squares

In the following, t as integration parameter is omitted for clarity of notation.
Let $W(t) = \text{diag}(\boldsymbol{w}(t))$.

$$
\begin{aligned}
\text{WFLS} &= \int (\text{C}\boldsymbol{\phi} - X\Theta\boldsymbol{b})^\intercal W(\text{C}\boldsymbol{\phi} - X\Theta\boldsymbol{b}) \\
&= \int (\text{C}\boldsymbol{\phi})^\intercal W(\text{C}\boldsymbol{\phi}) + \int (X\Theta\boldsymbol{b})^\intercal W(X\Theta\boldsymbol{b}) - \int (X\Theta\boldsymbol{b})^\intercal W(\text{C}\boldsymbol{\phi}) - \int (\text{C}\boldsymbol{\phi})^\intercal W(X\Theta\boldsymbol{b}) \\
&= \int \text{tr}\left[(W\text{C}\boldsymbol{\phi})(\text{C}\boldsymbol{\phi})^\intercal\right] + \int \text{tr}\left[(WX\Theta\boldsymbol{b})(X\Theta\boldsymbol{b})^\intercal\right] - \int \text{tr}\left[(W\text{C}\boldsymbol{\phi})(X\Theta\boldsymbol{b})^\intercal\right] - \int \text{tr}\left[(X\Theta\boldsymbol{b})(W\text{C}\boldsymbol{\phi})^\intercal\right] \\
&= \int \text{tr}\left[W\text{C}\boldsymbol{\phi}\boldsymbol{\phi}^\intercal\text{C}^\intercal\right] + \int \text{tr}\left[WX\Theta\boldsymbol{b}\boldsymbol{b}^\intercal\Theta^\intercal X^\intercal\right] - \int \text{tr}\left[(X\Theta\boldsymbol{b})^\intercal(W\text{C}\boldsymbol{\phi})\right] - \int \text{tr}\left[(X\Theta\boldsymbol{b})(W\text{C}\boldsymbol{\phi})^\intercal\right] \\
&= \int \text{tr}\left[\text{C}^\intercal W\text{C}\boldsymbol{\phi}\boldsymbol{\phi}^\intercal\right] + \int \text{tr}\left[X^\intercal WX\Theta\boldsymbol{b}\boldsymbol{b}^\intercal\Theta^\intercal\right] - 2\int \text{tr}\left[\boldsymbol{b}^\intercal\Theta^\intercal X^\intercal W\text{C}\boldsymbol{\phi}\right] \\
&= \int \text{tr}\left[\text{C}^\intercal W\text{C}\boldsymbol{\phi}\boldsymbol{\phi}^\intercal\right] + \int \text{tr}\left[\boldsymbol{b}^\intercal\Theta^\intercal X^\intercal WX\Theta\boldsymbol{b}\right] - 2\int \text{tr}\left[\boldsymbol{b}^\intercal\Theta^\intercal X^\intercal W\text{C}\boldsymbol{\phi}\right].
\end{aligned}
$$

The operations of integration and summation implied by the trace may be interchanged, so that the previous can be reformulated as

$$\text{WFLS} = \text{tr}\left[\int \text{C}^\intercal W \text{C} \boldsymbol{\phi}\boldsymbol{\phi}^\intercal\right] + \text{tr}\left[\int \boldsymbol{b}^\intercal \Theta^\intercal X^\intercal W X \Theta \boldsymbol{b}\right] - 2\text{tr}\left[\boldsymbol{b}^\intercal \int \Theta^\intercal X^\intercal W \text{C} \boldsymbol{\phi}\right].$$

In order to minimize this quantity, we have to take its derivative with respect to $\boldsymbol{b}$. The first term does not depend on $\boldsymbol{b}$ and hence disappears. The derivative of the third is equal to

$$-2\int \Theta^\intercal X^\intercal W \text{C} \boldsymbol{\phi}$$

and is easily obtained by recalling that the derivative of $\text{tr}(\text{B}^\intercal \text{A})$ with respect to B is A. For the derivation of the term in the middle it suffices to recall the following general rule for the derivation of the trace

$$\nabla_A \text{tr}(ABA^\intercal C) = CAB + C^\intercal AB^\intercal$$

Then,

$$\begin{aligned}\nabla_{\boldsymbol{b}} \text{tr}\left[\int \boldsymbol{b}^\intercal \Theta^\intercal X^\intercal W X \Theta \boldsymbol{b}\right] &= \int \nabla_{\boldsymbol{b}} \text{tr}\left[\boldsymbol{b}\boldsymbol{b}^\intercal \Theta^\intercal X^\intercal W X \Theta\right] \\ &= \left(\int \Theta^\intercal X^\intercal W X \Theta\right)\boldsymbol{b} + \left(\int \Theta^\intercal X^\intercal W X \Theta\right)\boldsymbol{b} \\ &= 2\left(\int \Theta^\intercal X^\intercal W X \Theta\right)\boldsymbol{b}.\end{aligned}$$

We operate similarly for the penalization term

$$\int \left[L\left(\Theta\boldsymbol{b}\right)\right]^\intercal \Lambda \left[L\left(\Theta\boldsymbol{b}\right)\right] = \int \text{tr}\left[\Lambda L(\Theta\boldsymbol{b}) L(\Theta\boldsymbol{b})\right]^\intercal = \int \text{tr}\left[\Lambda L(\Theta)\boldsymbol{b}\boldsymbol{b}^\intercal \left[L(\Theta)\right]^\intercal\right]$$
$$= \int \text{tr}\left[\boldsymbol{b}\boldsymbol{b}^\intercal \left[L(\Theta)\right]^\intercal \Lambda L(\Theta)\right] = \text{tr}\left[\boldsymbol{b}\boldsymbol{b}^\intercal \int \left[L(\Theta)\right]^\intercal \Lambda L(\Theta)\right] = \text{tr}\left[\boldsymbol{b}\boldsymbol{b}^\intercal R\right],$$

where we define R as the $L_\beta \times L_\beta$-dimensional matrix

$$R = \begin{pmatrix} R^1 & 0 & \dots & 0 \\ 0 & R^2 & \dots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \dots & R^q \end{pmatrix},$$

and $[R^j]_{mn} = \lambda_j \langle \left(\theta_m^j\right)'', \left(\theta_n^j\right)'' \rangle_{L^2(\mathcal{T})}$.
Taking advantage again of the properties of the derivative of the trace we get

$$\nabla_{\boldsymbol{b}} \left(\boldsymbol{b}\boldsymbol{b}^\intercal R\right) = 2R\boldsymbol{b}.$$

Summing up, we find that $\boldsymbol{b}$ satisfies

$$2\left(\int \Theta^\intercal X^\intercal W X \Theta\right)\boldsymbol{b} + 2R\boldsymbol{b} - 2\int \Theta^\intercal X^\intercal W \mathrm{C}\boldsymbol{\phi} = 0,$$

which simplifies into

$$\left[\left(\int \Theta^\intercal X^\intercal W X \Theta\right) + R\right]\boldsymbol{b} = \int \Theta^\intercal X^\intercal W \mathrm{C}\boldsymbol{\phi}.$$

A.2 Y-to-$\hat{\boldsymbol{\beta}}$ Map

This subsection is devoted to the construction of a map that acts as linkage between the raw observations Y and the estimate $\hat{\boldsymbol{\beta}}$. To this aim, it is useful to view the overall mapping as the composition of: *(i)* the *smoothing map* that associates the observations to the smooth functions, *(ii)* the *regression map* that connects the smooth functions to the vector of coefficients $\boldsymbol{b}$, *(iii)* the *basis expansion map* that couples the estimated coefficients with the basis functions for the β's.

The construction of the *smoothing map* is already given by (12).
The *regression map* connecting vec(C) into $\boldsymbol{b}$ is found by exploiting the properties of the vec() operator and of Kronecker product for manipulating the term $\mathrm{C}\phi(s)$ in equation (19). It is easy to show that the latter may be written in the form

$$\boldsymbol{b} = [J + R]^{-1}\left(\int \Theta(s)^T X(s)^T W(s)\left[\boldsymbol{\phi}(s)^T \otimes I_n\right] ds\right)\mathrm{vec}(\mathrm{C}).$$

Now it is straightforward to identify the mapping as the $(L_\beta \times Ln)$-dimensional matrix

$$S_\beta \coloneqq [J + R]^{-1}\left(\int \Theta(s)^T X(s)^T W(s)\left[\boldsymbol{\phi}(s)^T \otimes I_n\right] ds\right).$$

The *basis expansion map* carries out the linear combination of the basis functions that uniquely define the vector of functions $\hat{\boldsymbol{\beta}}$ from the estimated coefficients. For every $t \in \mathcal{T}$, the map is directly derived from (17) and coincides with matrix $\Theta(t)$:

$$S_\Theta(t) \coloneqq \Theta(t).$$

Finally, the complete mapping of Y into the vector of functions $\hat{\boldsymbol{\beta}}$ for every $t \in \mathcal{T}$ is given by the composition of all three mappings identified above and may be expressed in matricial form as

$$\mathrm{Map}_t \coloneqq S_\Theta(t) S_\beta S_\Phi,$$

so that we obtain the relation

$$\hat{\boldsymbol{\beta}}(t) = S_\Theta(t) S_\beta S_\Phi \mathrm{vec}(\mathrm{Y}).$$

B Simulation study

B.1 Variation of the fraction of partially observed data

With this set of simulations the intention is to show how the increase of missing information in the data impacts the results of the analysis, and how a weighted approach manages to stabilize the regression coefficients estimators. To this aim, the fraction of partially observed data p entering the data simulation process is varied within the set $\{0.1, 0.2, 0.4, 0.7\}$. The weighted and the non-weighted approaches are employed in Monte Carlo simulations, and the empirical distribution properties of the corresponding estimators $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ are examined. Results are shown in Figure 9. As expected, the estimation error made on each regression coefficient grows with the number of curves that are not fully observed, both in the weighted and in the non-weighted analysis. This is in accordance with the fact that the convergence of the reconstruction methods follows the convergence of the estimators of the mean and of the covariance operator for partially observed functional data (Kraus (2015), Kneip and Liebl (2020)). These are obtained point-wise by using for each t only the available curves at t (Kraus, 2015) or near t (Kneip and Liebl (2020), Yao et al. (2005)), and for each pair (t, s) only the complete pairs of curve values at s and t or in their neighbourhoods. If the number of valid observations on which the point estimators are built reduces, then the resulting estimates become increasingly biased towards the data on which they are calculated, and their distribution presents higher variability around the mean. Consequently, the uncertainty around the reconstructed trajectories increases and propagates into the empirical distribution of the point estimates of the regression coefficients. Figure 9 reveals how the weights help to stabilize the estimates by reducing their variance and their MSE. This result is endorsed by Figure 10, where the boxplots are built from the empirical distributions of the squared norms $\left|\left|\hat{\beta}_j^b - \overline{\hat{\beta}_j^b}\right|\right|^2_{L^2([0,1.75])}$ and $\left|\left|\hat{\beta}_j^b - \overline{\hat{\beta}_j^b}\right|\right|^2_{L^2([1.75,3.5])}$, evaluated separately on the two halves of the domain. Doing so, we are able to discern the stabilizing effect of the weights specifically to the segment $[1.75s, 3.5s]$, *i.e.* where a fraction p of the curves is right-censored and undergoes reconstruction. We observe in fact that the majority of the contribution to the estimators variability comes from the right half of the domain, which is also where the weights regulating effect is restricted to.

C Case study

C.1 Collinearity analysis in the covariates

The presence of collinearity between observed functional covariates x_{i1} and x_{i2} can be inspected resorting to the finite sample counterpart of the cross-covariance function of the random predictors

$$\mathrm{Cov}\left(\boldsymbol{x}(t_l), \boldsymbol{y}(s_m)\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left[x_i(t_l) - \bar{x}(t_l)\right]\left[y_i(s_m) - \bar{y}(s_m)\right],$$

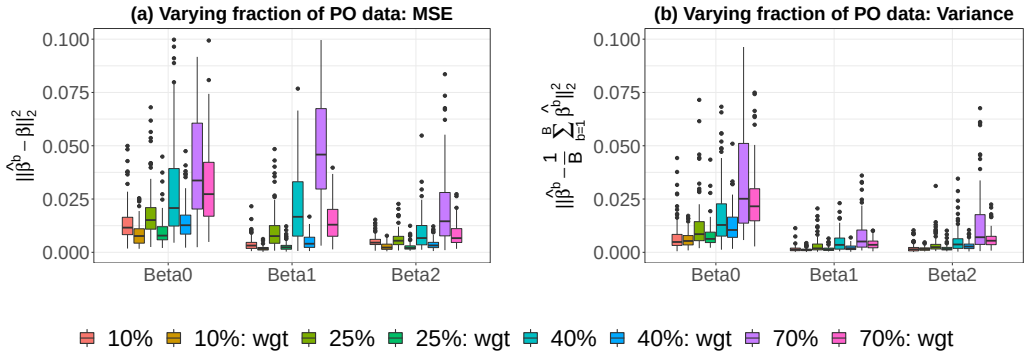

Figure 9: Boxplots of the empirical population of $||\hat{\beta}_j^b - \beta_j||_2^2$ (left) and of $||\hat{\beta}_j^b - \overline{\hat{\beta}_j^b}\,||_2^2$ (right), for every regression coefficient β_0, β_1 and β_2 and for different fractions p of partially observed curves.

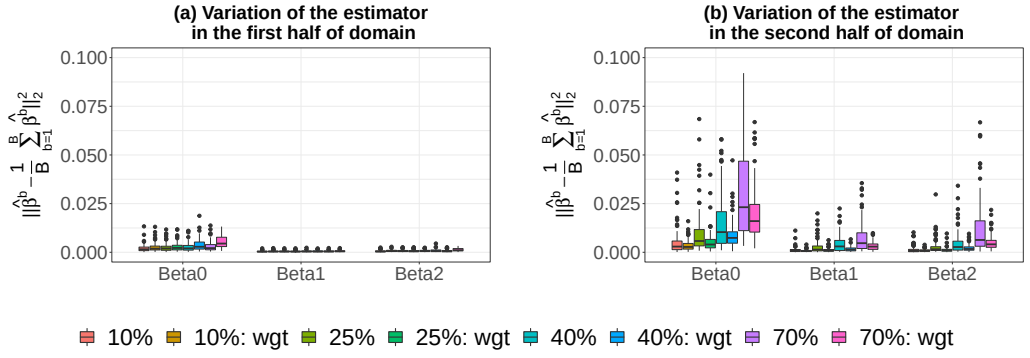

Figure 10: Boxplots of $||\hat{\beta}_j^b - \overline{\hat{\beta}_j^b}\,||_2^2$, where the norms are evaluated separately on the first and on the second half of the domain.

or equivalently of the cross-correlation function

$$\text{Corr}\left(\boldsymbol{x}(t_l), \boldsymbol{y}(s_m)\right) = \frac{1}{n-1}\sum_{i=1}^{n} \frac{\left[x_i(t_l) - \bar{x}(t_l)\right]\left[y_i(s_m) - \bar{y}(s_m)\right]}{\hat{\sigma}_x \hat{\sigma}_y}.$$

The contour plots in Figure 11 display the contour lines of the cross-correlation functions for some notable pairs of covariates entering (2). The domain of definition of the functions is plotted along the horizontal and the vertical axes, which are labelled with the name of the considered functional covariates. A deep investigation of the patterns of variation in the cross-correlation functions falls outside the scope of this paper. Rather, we solely acknowledge that each one of the considered pairs show positive correlation over the bivariate domain. The top-left panel in the figure reveals the presence of almost perfect positive correlation between the magnitude-independent geometric attenuation and the anelastic attenuation across all periods. This is not surprising, as both terms depend exclusively on the correction R of the Joyner-Boore distance (see Section 2). High positive cross-correlations persist also between the covariates accounting for low and high magnitudes and the magnitude-dependent geometric attenuation (bottom-left and bottom-right, respectively). Lastly, the top-right contour plot shows that the terms accounting for low and high magnitudes have levels of cross-correlation that approximately stay within values 0.4-0.7 across all periods. This is a direct consequence of the definition of the *source* term in (1) as a monotone increasing step-wise linear function, which alternately sets to zero the high (low) magnitude term depending on whether the moment magnitude is below (above) the hinge magnitude.

C.2 Calibration of the penalization parameters

The measure of prediction inaccuracy used for comparison is the L^2 norm of the prediction errors

$$\hat{e}_i^2 = \left|\left|y_i^s - \hat{y}_i\right|\right|_2^2, \tag{26}$$

where $\hat{y}_i$ is the i-th predicted curve and y_i^s the smoothed curve obtained via unweighted penalized least squares. It is convenient to evaluate the error committed on the smoothed curves because the comparison is made between different regression criteria and does not evaluate the overall analysis. The average of (26) is estimated via a 5-fold cross-validation, for the combinations of the penalty parameters $\lambda_1, \ldots, \lambda_9$ tested with the greedy procedure detailed below.

The penalization parameters are initialized to value $\bar{\lambda} = 10^{-2}$, which guarantees us a certain level of regularity. We take first intercept a, and let λ_1 vary in the set $\{-6, -5, \ldots, 1\}$, while maintaining the others fixed to $\bar{\lambda}$. $_1$ is set equal to the value corresponding to the lowest MSE resulting from a 5-fold CV. Then, we iterate the procedure for all the other coefficients. Eventually we obtain a list $(\lambda_1, \ldots, \lambda_9)$ which corresponds to the minimum of this greedy search. The second line of Table 4 collects the values that result from this procedure. Since coefficients a and c_2 are associated to anomalous levels of roughness, and since the estimate of c_1 resulting

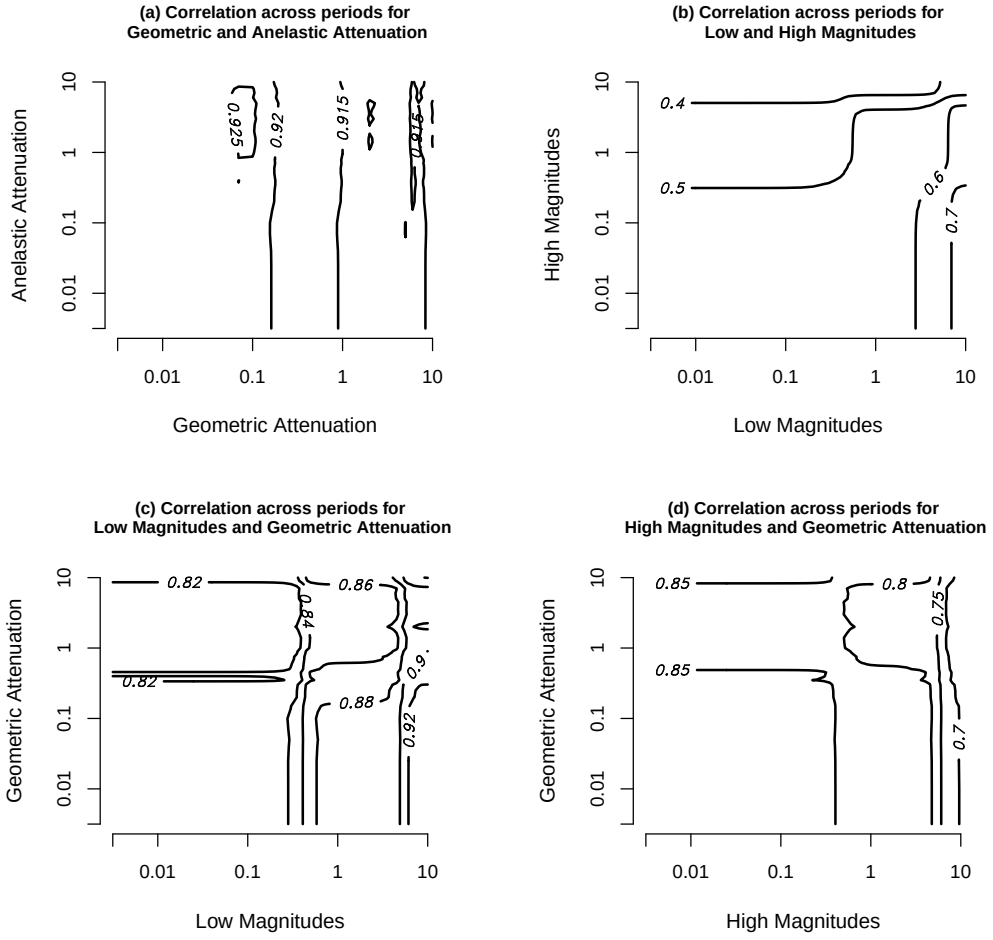

Figure 11: Contour plots of the cross-correlation functions over period for pairs of functional covariates: (a) magnitude-independent geometric attenuation and anelastic attenuation, (b) low magnitudes and high magnitudes, (c) low magnitudes and magnitude-dependent geometric attenuation, (d) high magnitudes and magnitude-dependent geometric attenuation.

Table 4: Optimal penalization coefficients entering the PWFLS criterion for regression.

a	b_1	b_2	f_1	f_2	c_1	c_2	c_3	k
10^{-6}	0.1	0.001	0.01	0.01	0.01	10^{-6}	10	0.01
0.001	0.1	0.001	0.01	0.01	0.1	0.01	10	0.01

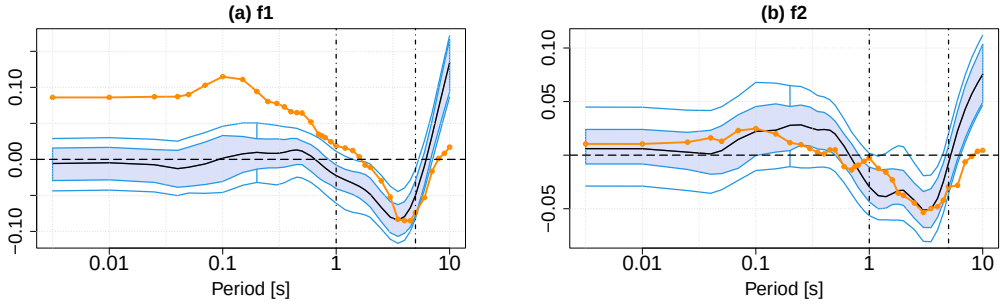

Figure 12: Functional boxplots of the estimated functional regression coefficients f_1 and f_2, and comparison with the estimates of ITA18 (orange). The black continuous lines represent the point estimates of the coefficients. The azure bands are the fences of the functional boxplots. The dashed line marks zero. The vertical lines mark the points which correspond to the last instants where 100% and $> 90\%$ of the curves are observed.

from this choice of parameters exhibits unphysical noise at long periods, the penalty parameters that correspond to these coefficients are forced to take higher values. It is worth pointing out that such modelling choice implies an increase in MSE of the order of one thousandth standard deviation. The third line of Table 4 collects the values of penalization parameters used in all subsequent analyses.

C.3 Comparison of the coefficients estimates

We do not comment on coefficients f_1 and f_2 for two main reasons. Firstly, the faulting mechanism is known to have little impact on the standard deviation of a GMPE (Bommer et al. (2003), Lanzano et al. (2019)), and to be included in the functional form for purposes of seismic hazard assessment, rather than to get a better performance of the regression model. Secondly, coefficient f_2 is known to be dependent on the region where the event occurs (Lanzano et al., 2021). Consequently, the ergodic model expressed in (1) is not expected to capture the effects of the thrust-faulting, to which f_2 is associated. For the sake of completeness, the functional estimates for f_1 and f_2 are displayed in Figure 12.

MOX Technical Reports, last issues

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

30/2022 Bonetti S.; Botti M.; Antonietti P.F.
Discontinuous Galerkin approximation of the fully-coupled thermo-poroelastic problem

29/2022 Fumagalli, I.; Polidori, R.; Renzi, F.; Fusini, L.; Quarteroni, A.; Pontone, G.; Vergara, C.
Fluid-structure interaction analysis of transcatheter aortic valve implantation

28/2022 Ciarletta, P.; Pozzi, G.; Riccobelli, D.
The Föppl–von Kármán equations of elastic plates with initial stress

26/2022 Orlando, G.
A filtering monotonization approach for DG discretizations of hyperbolic problems

25/2022 Cavinato, L; Gozzi, N.; Sollini, M; Kirienko, M; Carlo-Stella, C; Rusconi, C; Chiti, A; Ieva, F.
Perspective transfer model building via imaging-based rules extraction from retrospective cancer subtyping in Hodgkin Lymphoma

27/2022 Lazzari J., Asnaghi R., Clementi L., Santambrogio M. D.
Math Skills: a New Look from Functional Data Analysis

24/2022 Cappozzo, A.; McCrory, C.; Robinson, O.; Freni Sterrantino, A.; Sacerdote, C.; Krogh, V.; Par
A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events

23/2022 Masci, C.; Ieva, F.; Paganoni, A.M.
A multinomial mixed-effects model with discrete random effects for modelling dependence across response categories

22/2022 Regazzoni, F.; Pagani, S.; Quarteroni, A.
Universal Solution Manifold Networks (USM-Nets): non-intrusive mesh-free surrogate models for problems in variable domains

20/2022 Clementi, L.; Gregorio, C; Savarè, L.; Ieva, F; Santambrogio, M.D.; Sangalli, L.M.
A Functional Data Analysis Approach to Left Ventricular Remodeling Assessment