

Weighted Functional Data Analysis for the Calibration of a Ground Motion Model in Italy

Teresa Bortolotti^{1*}, Riccardo Peli¹, Giovanni Lanzano², Sara Sgobba², and
Alessandra Menafoglio¹

¹MOX, Department of Mathematics, Politecnico di Milano, Milano, Italy

²Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Milano, Italy

*`teresa.bortolotti@polimi.it`

Abstract

Motivated by the crucial implications of Ground Motion Models 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 of oscillation periods entails embedding the analysis in the context of partially observed functional data and performing data reconstruction. This work proposes a novel methodology that accounts for the fact that parts of the curves are directly observed and other parts are reconstructed, thus characterized by greater uncertainty. The method defines observation-specific functional weights, which enter the estimation process to reduce the impact that the less reliable portions 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 Ground Motion Model considered.

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

Acknowledgment: The Version of Record of this manuscript has been published and is available in Journal of the American Statistical Association, 2024, <https://www.tandfonline.com/doi/full/10.1080/01621459.2023.2300506>.

1 Introduction

In the field of seismic hazard assessment, Ground Motion Models (GMMs, Douglas and Edwards 2016) estimate the expectation of ground motion intensity measures, conditionally on predictors that are descriptive of a given seismic scenario. The intensity of earthquake-induced ground motion is measured at the recording sites by a set of damped harmonic oscillators, each characterized by its natural period of oscillation T (Newmark and Hall, 1982). The seismic response spectrum is a measure of intensity defined as the peak response of the oscillators to the seismic force. Consequently, it can either be seen as a collection of ordinates defined with respect to T , or as a profile along the range of oscillation periods. This gives rise to the threefold possibility of inserting the analysis of ground motion in a scalar (e.g., Bindi et al. 2011; Boore et al. 2014, Lanzano et al. 2019), 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, multivariate approaches inevitably suffer from the curse of dimensionality. By moving the focus from period-specific intensity measures to their continuous profile over the domain of oscillation periods, a functional approach aims to solve the shortcomings of both scalar and multivariate approaches. This work exploits the methodologies of Functional Data Analysis (FDA, Ramsay and Silverman 2005; Horváth and Kokoszka 2012) on seismic data, and provides a functional formulation to the scalar GMM proposed by Lanzano et al. (2019) for the Italian context. To the best of our knowledge, the present work is original in proposing a functional formulation to a GMM for the mean intensity measure. It is worth noting that, consistently with the ergodic scalar GMM by Lanzano et al. (2019), its functional counterpart does not account for systematic event and station nor their spatial correlation. While integrating these effects is an essential step towards the formulation of a non-ergodic mixed-effect functional GMM that properly considers all sources of variability, this aim falls outside the scope of the present work. Additionally, as the functional model sticks to the linear form of Lanzano et al. (2019), it does not handle nonlinear terms.

The peculiarity of the analyzed ground motion data, which coincide with those used in Lanzano et al. (2019), 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 observed only up to a certain period, and not on 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 domain, we are reluctant in erasing them from the dataset, or in reducing the domain of analysis similarly as in Menafoglio et al. (2020). 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 incompletely observed functional data. 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). We exploit these last techniques to reconstruct the missing patterns of the acceleration spectra, in order to preserve the formulation of the functional GMM over the entire domain.

The methodological novelty of this work fits downstream of curves reconstruction. We propose a weighted workflow which accounts for the uncertainty introduced in the reconstruction step. In particular, the proposed framework couples each curve with a weight function, taking value one where original observations are available and decreasing to zero the more the reconstruction of the missing trajectory becomes uncertain. The classical concurrent functional regression and smoothing are modified in order to include weights. A weighted least squares criterion for smoothing is discussed in Ramsay and Silverman (2005), and associates the longitudinal observations of a curve to scalar weights varying over the sampling instants. Differently to what we propose in this work, though, the weights are equal for all curves. This use of weights allows the optimal smoothed curves to be characterized 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 least square criterion for functional regression (Ramsay and Silverman, 2005), conversely, weights vary across observations but are constant over the domain of analysis. To the best of our knowledge, the inclusion of curve-specific functional weights in the estimation step is novel both for concurrent functional regression and for smoothing. By including curve-specific functional weights, only the originally observed data and the most reliable portions of the reconstructed trajectories have an impact on the regression results. The optimal solution to the regression problem is found in the middle between two extreme cases, namely the analysis conducted on the entire reconstructed trajectories and analysis restricted to the parts of the curves that are directly observed. The remaining of the work is organized as follows. The considered ground motion model and the seismological data are introduced in Section 2. The proposed weighted functional methodology is presented in Section 3. Section 4 reports the simulation studies conducted to evaluate the performance of the proposed methodology. In Section 5, we discuss and interpret the results of the analysis performed on the considered seismological data. Concluding remarks and discussions are provided in Section 5.

2 Model and data

2.1 Model

In the context of seismic hazard assessment, GMMs estimate the distribution of an intensity measure (IM) conditionally on seismic variables that are descriptive of the source of the earthquake, the site of the recording station and the path taken by the seismic wave from the epicenter to the station.

Background The model of Lanzano et al. (2019), which we refer to as ITA18, resorts to ordinary least-squares to separately fit 37 scalar 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 $T_j \in \mathcal{T} := [0.01 \text{ s}, 10 \text{ s}]$, $j = 1, \dots, 36$. The median value of the j -th IM is estimated according to the following form:

$$\begin{aligned} \log_{10} IM_j = & a_j \\ & + \underbrace{b_{1j}(M_w - M_{h,j})\mathbb{1}_{(M_w \leq M_{h,j})} + b_{2j}(M_w - M_{h,j})\mathbb{1}_{(M_w \geq M_{h,j})} + f_{1j}SoF_1 + f_{2j}SoF_2}_{F_{M,j}(M_w, SoF)} \\ & + \underbrace{c_{1j}(M_w - M_{\text{ref},j})\log_{10} R_j + c_{2j}\log_{10} R_j + c_{3j}R_j}_{F_{D,j}(M_w, R_j)} \\ & + \underbrace{k_j \log_{10} \frac{V_0}{800}}_{F_{S,j}(V_{S30})} + \epsilon_j, \end{aligned} \quad (1)$$

where a_j is the offset, $F_{M,j}(M_w, SoF)$, $F_{D,j}(M_w, R)$, $F_{S,j}(V_{S30})$ are the source-, path- and site-related terms respectively, and ϵ_j is the remaining error. The source term is specified as a step-wise linear function which changes slope at $M_{h,j}$, namely the hinge magnitude at T_j . Terms f_1 and f_2 are the coefficients related to two dummy variables SoF_1 and SoF_2 associated to the levels of style-of-faulting: strike-slip and thrust faulting. Normal faulting is set as baseline. The path term is given by the summation of three terms, the first two accounting for the geometrical spreading of the waves from the source, and the third accounting for the anelastic attenuation. Parameter $M_{\text{ref},j}$ is the reference magnitude at T_j . The distance R_j represents a correction of the pure Joyner-Boore distance (d_{JB}) – i.e., the closest distance to the surface projection of an extended fault – and is defined as $R_j = \sqrt{d_{JB}^2 + h_j^2}$, where h_j is the period-specific parameter of pseudodepth, measured in kilometers. Lastly, variable V_{S30} in the site-related term is the shear-wave velocity, and $V_0 = V_{S30}$ if $V_{S30} \leq 1500 \text{ m/s}$, $V_0 = 1500 \text{ m/s}$ otherwise. Note that $F_{S,j}$ is assumed independent of V_{S30} for large values of shear-wave velocity (Kamai et al., 2014).

Functional embedding of the scalar model Parameters $M_{h,j}$, $M_{\text{ref},j}$ and h_j 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 it is critical that these parameters are consistent with their well-known physical meaning, and since fitting a nonlinear model – besides being methodologically nontrivial – would not guarantee physical consistency of the estimates, we assume them to be known functions of the period. In particular, we exploit the estimates of $M_{h,j}$, $M_{\text{ref},j}$ and h_j obtained period-wise from the preliminary step of non-linear regression discussed by Lanzano et al. (2019). We then define functions \mathcal{M}_{ref} and h in the space generated by a cubic B-spline basis, where the optimal coefficients result from a step of penalized smoothing. The point estimates of $M_{h,j}$ display a step-wise

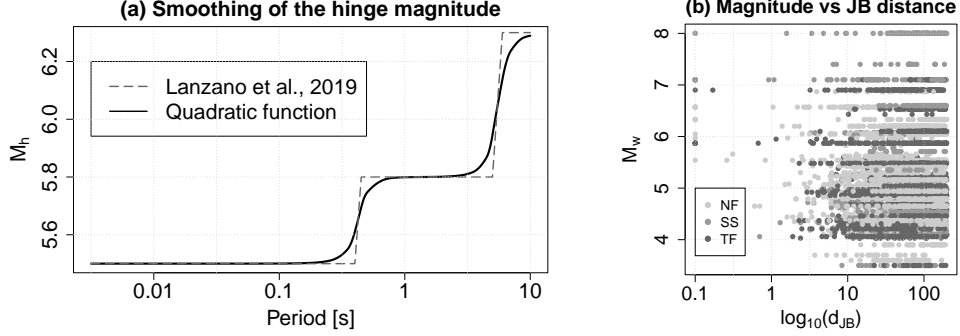


Figure 1: (a) Definition of M_h as the function resulting from the smoothing of its step-function estimate from Lanzano et al. (2019). (b) Scatter plot of magnitude vs Joyner-Boore distance (d_{JB}), colored by style-of-faulting. The records at $d_{JB} = 0$ are plotted at 0.1 km.

behavior along the domain, producing jumps in the prediction of the spectrum for scenarios close to the hinge magnitude. In order to solve these discontinuity issues, Sabetta et al. (2021) corrects $M_{h,j}$ to have a smoother variation in the range of periods $[0.25 \text{ s}, 0.7 \text{ s}]$. Following this line, we define function \mathcal{M}_h on a quadratic B-spline basis via a smoothing of $M_{h,1}, \dots, M_{h,37}$ that penalizes the first derivative. Figure 1a shows function \mathcal{M}_h resulting from this smoothing step. 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:

$$\begin{aligned} \log_{10} \mathcal{IM} = & \alpha + b_1(M_w - \mathcal{M}_h)\mathbb{1}_{(M_w \leq \mathcal{M}_h)} + b_2(M_w - \mathcal{M}_h)\mathbb{1}_{(M_w \geq \mathcal{M}_h)} + f_1 SoF_1 \\ & + f_2 SoF_2 + c_1(M_w - \mathcal{M}_{\text{ref}}) \log_{10} \mathcal{R} + c_2 \log_{10} \mathcal{R} + c_3 \mathcal{R} + k \log_{10} \frac{V_0}{800} + \mathcal{E}. \end{aligned} \quad (2)$$

In (2), \mathcal{IM} is a random variable with values in the space of square integrable functions, α , \mathcal{M}_h , \mathcal{M}_{ref} and \mathcal{R} are known functions with domain \mathcal{T} and \mathcal{E} is assumed to be generated by a zero mean stochastic process – i.e. a random process over the domain of oscillation periods, whose point-wise expected value is a.e. zero, namely $\mathbb{E}[\mathcal{E}(t)] = 0$ a.e. in \mathcal{T} . Coefficients $\alpha, b_1, b_2, f_1, f_2, c_1, c_2, c_3, k$, are the unknown functions that we aim to estimate.

2.2 Data

The analysis is carried out on the same dataset used for the calibration of ITA18, which includes 5607 records, relative to 146 earthquakes and 1657 stations (Lanzano et al., 2022). 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 in ITACA were selected according to the following criteria: (i) earthquakes of active shallow crustal regions (only events of tectonic origin with focal depth lower than 30 km) occurred in the period 1972–2017, (ii) minimum moment magnitude (M_w) set to 3.5, (iii) Joyner–Boore distance lower than 200 km, and (iv) 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 (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 $[0.01\text{ s}, 10\text{ s}]$. Conversely, 26 sampling instants are in the interval $[0.01\text{ s}, 2\text{ s}]$, while 11 points span the remaining of the domain. This motivates us to analyse the spectrum on the logarithm transformation of the periods, thus considering the sequence $(\log_{10}(0.01\text{ s}), \dots, \log_{10}(10\text{ s}))$ as the sampling instants. This has the twofold advantage of obtaining a more uniform sampling of the curves over the domain and of better representing the greater seismological interest that practitioners have on short rather than long periods. As mentioned earlier in this section, we aim to provide a functional formulation of (1) that is valid for the PGA jointly with the interval $[0.01\text{ s}, 10\text{ s}]$. Since the PGA corresponds to the limiting value of SA when the period approaches 0, a natural definition of the domain of analysis would be $[0\text{ s}, 10\text{ s}]$. However, opting for a logarithmic transformation of the periods, and the need of conducting the analysis on a bounded domain, imply reconsidering the choice of the sampling instant which the PGA corresponds to. In absence of specific literature on the matter to refer to, we set the left end of the domain at -2.5 and assign here the value of PGA. Setting the PGA close to -2 (i.e. $\log_{10}(0.01)$) aligns with the fact that the values of PGA and SA(0.01) nearly coincide in the practice (Bradley, 2011; Lanzano et al., 2019), and guarantees a regular sampling of the domain up to its left end. Figure 2a displays the longitudinal observations of the IM profiles resulting from such choice.

Partially observed response variable RotD50 (Boore, 2010) – the IM 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

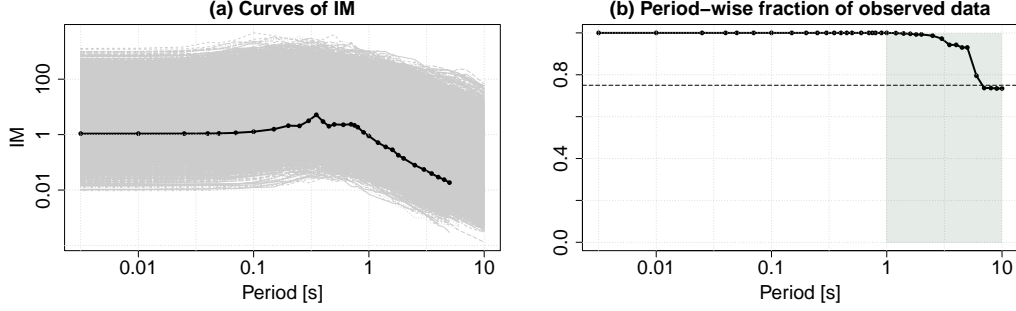


Figure 2: (a) Profiles of IM (grey). The dark line represents a partially observed IM profile. (b) Period-specific fraction of observed records. The dashed horizontal line marks 0.75. The vertical band highlights the partially observed portion of the domain.

curves, from their last valid observation up to $T = 10$ s.

3 Methods

We detail below the proposed curve-specific weighted methodology to get point estimates of functional regression coefficients. Details on the assessment of the uncertainty related to the point estimates are reported in Section 1.3 of the Supplementary Material.

3.1 Weighted Regression

Let y_1, \dots, y_n be realizations of independent and identically distributed functional random variables with values in $L^2(\mathcal{T})$, \mathcal{T} open subset of \mathbb{R} . We consider a functional concurrent linear regression model with independent functional covariates $x_1(t), \dots, x_q(t)$, namely

$$\mathbf{y}(t) = X(t)\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t), \quad t \in \mathcal{T}, \quad (3)$$

where $\boldsymbol{\beta}(t) = (\beta_1(t), \dots, \beta_q(t))^T$ denotes the vector of functional coefficients in $L^2(\mathcal{T})$, evaluated in t , $X(t) \in \mathbb{R}^{n \times q}$ is the design matrix at t and $\mathbf{y}(t)$ is a n -dimensional vector containing the response functions evaluated in t . The error term is a n -dimensional vector of functions $\epsilon_1, \dots, \epsilon_n$, that are assumed to be independent realizations of a zero-mean stochastic process valued in $L^2(\mathcal{T})$.

Penalized weighted functional least-square criterion The aim of this section is to extend to a curve-specific weighted approach the penalized functional least-square criterion discussed in Ramsay and Silverman (2005). We do so by exploiting the features of a weighted L^2 -norm, which we briefly introduce here. Let $w : \mathcal{T} \rightarrow [0, 1]$ be a bounded non-negative function, which we refer to as *weight*, let $f \in L^2(\mathcal{T})$ and w be a weight associated to f . We define the weighted L^2 -norm of f with respect to w as $\|f\|_w = \sqrt{\int_{\mathcal{T}} w(s)f(s)^2 ds}$. 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. Let w_1, \dots, w_n be the weights associated to the errors $\epsilon_1, \dots, \epsilon_n$. We define the penalized weighted functional least-square (PWFLS) criterion as the minimization of

$$\begin{aligned} \text{PWFLS} &= \sum_{i=1}^n \|\epsilon_i\|_{L^2(\mathcal{T}), w_i}^2 + \sum_{j=1}^q \lambda_j \|D^2 \beta_j\|_{L^2(\mathcal{T})}^2 \\ &= \sum_{i=1}^n \int_{\mathcal{T}} w_i(s) \epsilon_i(s)^2 ds + \sum_{j=1}^q \int_{\mathcal{T}} \lambda_j (D^2 \beta_j(s))^2 ds, \end{aligned} \quad (4)$$

where $\sum_{j=1}^q \lambda_j \|D^2 \beta_j\|_{L^2(\mathcal{T})}^2$ is a roughness penalty that enters the criterion to regularize and stabilize the estimates of the regression coefficients, and $\lambda_1, \dots, \lambda_q$ are the positive coefficient-specific penalization parameters which can be tuned to allow for different degrees of smoothness in the coefficients estimates. The roughness penalty $\sum_{j=1}^q \lambda_j \|D^2 \beta_j\|_{L^2(\mathcal{T})}^2$ 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 (4) can be interchanged, and consequently one may write

$$\begin{aligned} &\int_{\mathcal{T}} \sum_{i=1}^n w_i(s) (y_i(s) - \mathbf{x}_i(s)^T \boldsymbol{\beta}(s))^2 ds + \int_{\mathcal{T}} \sum_{j=1}^q \lambda_j (D^2 \beta_j(s))^2 ds \\ &= \int_{\mathcal{T}} [\mathbf{y}(s) - X(s)\boldsymbol{\beta}(s)]^T W(s) [\mathbf{y}(s) - X(s)\boldsymbol{\beta}(s)] ds + \int_{\mathcal{T}} [L\boldsymbol{\beta}(s)]^T \Lambda [L\boldsymbol{\beta}(s)] ds, \end{aligned} \quad (5)$$

where we set $W(s) = \text{diag}(w_1(s), \dots, w_n(s))$ to be the diagonal matrix of the weights evaluated in s , L the linear differential operator taking the second derivative of each regression coefficient, and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_q)^T$ the diagonal matrix of the q penalization parameters. The minimization of (5) passes through a dimensionality reduction of the problem. Each regression coefficient β_j is estimated as an element of the finite dimensional space generated by a convenient basis $\boldsymbol{\theta}_j$, and the part of β_j that remains not captured is assumed to be negligible. Accordingly, this section will consider the regression coefficients to be

$$\beta_j(t) = \sum_{l=1}^{L_j} b_{jl} \theta_l^j(t), \quad j = 1, \dots, q. \quad (6)$$

The argument above is formulated in its most general setting, which considers the bases for each regression coefficient as distinct. Such comprehensiveness is particularly convenient when one has a priori knowledge that the effects entering the regression model have different degrees of roughness, as it allows to flexibly adjust the definition of each coefficient β_j in the space generated by suitable basis functions $\theta_1^j, \dots, \theta_{L_j}^j$. Let $L_{\beta} = \sum_{j=1}^q L_j$. Note that (6) can be compacted in matricial form as $\boldsymbol{\beta}(t) = \Theta(t)\mathbf{b}$, where $\Theta(t) \in \mathbb{R}^{q \times L_{\beta}}$ is the matrix of the point evaluations at t of

the basis functions

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

and \mathbf{b} is the L_β -dimensional vector of the coefficients of the projections of β_1, \dots, β_q on bases $\theta_1, \dots, \theta_q$. A finite dimensional formulation of model (3) follows the considerations made above and reads $\mathbf{y}(t) = X(t)\Theta(t)\mathbf{b} + \epsilon(t)$. By putting this compact form in (5) one gets

$$\int_{\mathcal{T}} [\mathbf{y}(s) - X(s)\Theta(s)\mathbf{b}]^T W(s) [\mathbf{y}(s) - X(s)\Theta(s)\mathbf{b}] ds + \int_{\mathcal{T}} [L\Theta(s)\mathbf{b}]^T \Lambda [L\Theta(s)\mathbf{b}] ds.$$

This quadratic form is the starting point of the calculation, extensively reported in Section 1.1 of the Supplementary Material, that leads to the following equation for vector \mathbf{b} :

$$[J + R] \mathbf{b} = \int \Theta(s)^T X(s)^T W(s) \mathbf{y}(s) ds, \quad (7)$$

where $J := (\int_{\mathcal{T}} \Theta(s)^T X(s)^T W(s) X(s) \Theta(s) ds)$, and $R \in \mathbb{R}^{L_\beta \times L_\beta}$ is the matrix accounting for the penalization term.

3.2 Weighted Smoothing

The closed form solution to the regression problem is found under the assumption that the responses are known functional data. In practice, only a finite number of longitudinal points of the curves is observed. It should be noted that (7) is general and valid regardless of whether the discrete longitudinal observations undergo a smoothing procedure or not. Performing a smoothing step has an impact on the calculation of integrals. When no smoothing is performed, the integrals in (7) are estimated using quadrature rules. In the case where the observations are smoothed on a sufficiently rich basis, the longitudinal observations are non-linearly interpolated, the integrals are known, and the calculation traces back to the coefficients of the observations with respect to the basis. This can lead to a great computational advantage in contexts with a large number of observations and sampling points. Concerning the application of this work, the discrete observations of the seismic functional covariates are smoothed as discussed in Section 2. Since the IM profiles are partially observed, we propose an ad hoc weighted smoothing which modulates the impact that the reconstructed observations of a curve have on its smooth estimate.

Penalized weighted smoothing For each curve y_i , let $\mathbf{y}_i = (y_i(t_1), \dots, y_i(t_N))$ be the vector of discrete observations of y_i at the sampling instants $t_1 \in \mathcal{T}, \dots, t_N \in \mathcal{T}$. For each \mathbf{y}_i , the smoothing technique (e.g. Ramsay and Silverman 2005, de Boor 2001) fits the discrete observations $y_i(t_1), \dots, y_i(t_N)$ with a function f_i belonging to a convenient functional space, according to the model $y_i(t_j) = f_i(t_j) + e_{ij}$, $\forall j =$

$1, \dots, N$, where e_{i1}, \dots, e_{iN} are independent mean zero random variables. The penalized weighted least-square criterion finds the function \hat{f}_i that solves

$$\hat{f}_i = \underset{f \in H^2(\mathcal{T})}{\operatorname{argmin}} \sum_{j=1}^T v_{ij} (y_i(t_j) - f(t_j))^2 + \zeta \|D^2 f\|_{L^2(\mathcal{T})}^2. \quad (8)$$

In (8), the sum of squared errors is discounted at each sampling instant t_j by a term $v_{ij} \in [0, 1]$, which plays the role of giving different weight to smoothing errors made at different sampling instants. The penalization term $\|D^2 f\|_{L^2(\mathcal{T})}^2$ quantifies the roughness of f and is added to the least squares to impose a certain degree of smoothing on the optimal curve. The smoothing parameter $\zeta \in (0, \infty)$ controls the impact of the penalization with respect to the least squares, and may be tuned via generalized cross-validation. Notice that the argmin is taken over the Sobolev space $H^2(\mathcal{T})$ in order to guarantee the finiteness of $\|D^2 f\|_{L^2(\mathcal{T})}^2$. The solution to (8) is known to be a cubic spline with knots at the data points t_j (de Boor, 2001), which can be expressed in the form $\hat{f}_i(t) = \sum_{j=1}^r \hat{c}_{ij} \phi_j(t) = \hat{\mathbf{c}}_i^T \boldsymbol{\phi}(t)$, where ϕ_1, \dots, ϕ_r are cubic B-spline basis functions with $r=N+2$, and $\mathbf{c}_i \in \mathbb{R}^r$ uniquely identifies \hat{f}_i with respect to the basis. Problem (8) can then be equivalently expressed in a multivariate form for \mathbf{c}_i , and the solution is found in closed form as $\hat{\mathbf{c}}_i = S_{\Phi}^i \mathbf{y}_i$, where S_{Φ}^i is the curve-specific smoothing matrix that depends on the basis functions, on the weights v_{i1}, \dots, v_{iN} , and on the penalization $\zeta \|D^2 f\|_{L^2(\mathcal{T})}^2$ (Ramsay and Silverman 2005, Section 5.2).

Regression coefficients estimates for smooth responses and covariates In our analysis, the smoothing error is considered negligible and included in the regression error. Then, the i -th smooth response variable reads $\hat{f}_i(t) = \sum_{l=1}^r \hat{c}_{il} \phi_l(t)$, and the vector of smoothed response variables at t can be represented in matricial form as $\hat{\mathbf{f}}(t) = \mathbf{C} \boldsymbol{\phi}(t)$, where $\mathbf{C} \in \mathbb{R}^{n \times r}$ is the matrix of the coefficients $\hat{\mathbf{c}}_1, \dots, \hat{\mathbf{c}}_n$ with respect to the cubic B-spline basis $\boldsymbol{\phi}$. Let the design matrix at t , $\mathbf{X}(t)$, contain the evaluations of the smoothed covariates at t . Then it is trivial to show that the coefficients estimates in the case of smooth responses and covariates is found in closed form as

$$[J + R] \mathbf{b} = \int \Theta(s)^T \mathbf{X}(s)^T \mathbf{W}(s) \mathbf{C} \boldsymbol{\phi}(s) ds, \quad (9)$$

where J and R are defined as in Section 3.1.

4 Simulation Study

This section is devoted to the validation through a simulation study of the proposed weighted methodology. In particular, we aim to assess: (i) the robustness of the estimates with respect to the method adopted for the reconstruction of the missing trajectories, comparing the results obtained with the weighted and the unweighted analysis, (ii) the accuracy of the estimates as the definition of the weights changes,

(iii) the robustness of the estimates with respect to an increase in the fraction of partially observed data.

4.1 Simulation of Partially Observed Functional Data

Functional data are generated according to model $y_i(t) = \beta_0(t) + \beta_1(t)x_{1i} + \beta_2(t)x_{2i}(t) + \epsilon_i(t)$, where ϵ_i are independent realizations of a zero-mean stochastic process. The inclusion of a scalar and a functional covariate 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 leverages the main modes of variability of the response profiles and of some predictors from the case study, to sample functional data from a novel and unknown distribution. Partially observed longitudinal observations are generated by coupling a fraction p of the simulated curves with curve-specific domains, via a randomized right-censoring procedure. Further details on the simulation of partially observed data are in Section 2.1 of the Supplementary Material.

Definition of the weights We first point out that the systems of weights introduced for weighted smoothing and weighted 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 combination of the two techniques as a single, weighted procedure. Accordingly, each smoothing weight v_{ij} is directly derived from the definition of the i -th functional weight w_i , introduced in Section 3.1, and set equal to $v_{ij} = w_i(t_j)$. 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. The more the reconstructed value is uncertain, the smaller the weight should be. We consider two possible systems of weights. Logistic weights are a convenient choice to achieve a decrease in confidence from 1 to small values, and provide a clear interpretation in terms of downweighting the reconstructed trajectories. In their definition, parameter a controls the rate of decay of the logistic functions. An alternative definition of the weights descends directly from the correlation between the observed and unobserved parts of the trajectories, and relies on the quantification of the reduction in the uncertainty of the missing trajectory, achieved with the reconstruction. We refer to Section 2.2 of the Supplementary Material for a detailed argument on how both logistic and reconstruction-driven weights are defined. Figure 3 shows a simulated partially observed curve, its reconstruction with the method proposed by Kneip and Liebl (2020), and examples of associated functional weights.

4.2 Validation of the Weighted Analysis

The performances of the weighted and the unweighted analysis are compared in terms of mean squared error and variance of the estimators $\hat{\beta}_j$, $j = 0, 1, 2$, defined as $\text{MSE}(\hat{\beta}_j) = \mathbb{E} \left[\|\hat{\beta}_j - \beta_j\|_2^2 \right]$ and $\text{Var}(\hat{\beta}_j) = \mathbb{E} \left[\|\hat{\beta}_j - \mathbb{E}[\hat{\beta}_j]\|_2^2 \right]$. In the practice, MSE and variance of each $\hat{\beta}_j$ are extracted from the empirical distributions

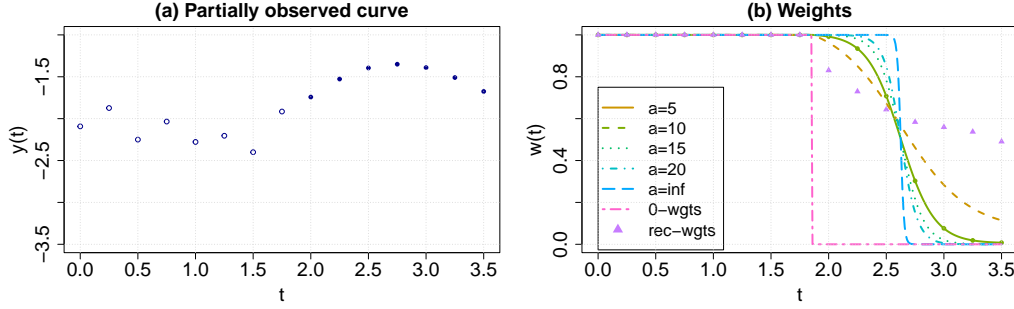


Figure 3: (a) Observed (empty dots) and reconstructed (full dots) longitudinal values of a simulated partially observed curve. (b) Examples of logistic weights associated to the curve, with a varying within the set $\{5, 10, 15, 20, \infty\}$, and of reconstruction-driven weights (rec-wgts). Label 0-wgts corresponds to a step function taking value 1 up to the last observed point and falling to 0 in the remaining part of the domain. In this example, the reconstruction-driven weight is obtained for the method of Kneip and Liebl (2020).

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 with their finite sample counterparts. For $b = 1, \dots, B$, the simulation does: (i) generate a sample of fully and partially observed functional data, (ii) reconstruct the right-censored curves, (iii) define the system of weights, (iv) smooth the discrete observations, (v) obtain the estimates $(\hat{\beta}_0^b, \hat{\beta}_1^b, \hat{\beta}_2^b)$. 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 sets of simulations. First, we evaluate the impact of the reconstruction methodology on the estimates, for both the weighted and unweighted approaches. Second, we assess how alternative definitions of the weighting system affect the resulting estimates. Finally, we vary the fraction p of partially observed data to assess whether the weighted method improves the stability and the accuracy of the estimators as the missing information increases. The results of the last analysis are reported in Section 2.3 in the Supplementary Material. The first and the third sets of simulations are done considering logistic weights with $a = 10$. In the first two sets of simulations, p is set to 0.4.

Robustness to the reconstruction methods We consider three different methodologies for the reconstruction of partially observed functional data. For the sake of clarity, they are referred to with acronyms and their working principle is briefly recalled¹. Acronym Kraus refers to the reconstruction of the missing trajectory made by a Hilbert-Schmidt operator, estimated via a functional linear ridge regression (Kraus, 2015). Acronym KL-PC refers to a functional completion made by a reconstruction operator, which estimates the principal components of the curve over the

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

Table 1: Comparison of $\text{bias}^2(\beta_j)$ among the adopted reconstruction methods. Term *wgt* indicates that the reconstruction method is coupled with the weighted analysis. When the term *wgt* is not present, it indicates that the reconstruction is followed by the classical unweighted procedures of smoothing and concurrent functional regression.

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

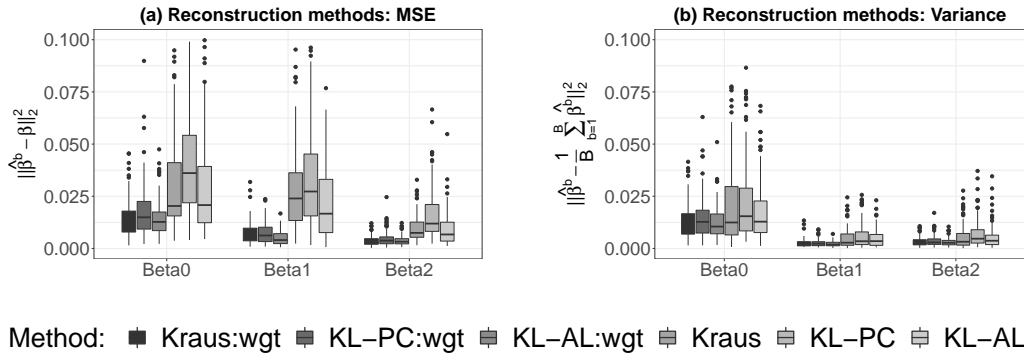


Figure 4: Boxplots of the empirical population of $\|\hat{\beta}_j^b - \beta_j\|_2^2$ (left) and of $\|\hat{\beta}_j^b - \overline{\hat{\beta}_j^b}\|_2^2$ (right), for every reconstruction method considered and for each $\beta_j, j = 0, 1, 2$.

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). Acronym KL-AL refers to the same procedure as KL-PC, but operates a preliminary step of non-parametric smoothing on the observed fragments, which guarantees continuity at the boundary of observed and missing domain (thus the term ALign) (Kneip and Liebl, 2020). When comparing the weighted and unweighted approaches over different reconstruction methods, we expect the inclusion of the weights to reduce the differences in the resulting estimates, since weights enter the estimation criteria by lowering the impact of the reconstructed observations on the estimates. The results displayed 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 regression coefficient. Additionally, we observe that the weighted methodology effectively lowers the variance and bias (Table 1) of the point estimators, meaning that the analysis benefits in terms of stabilization of the estimators and of estimation accuracy. Since method KL-AL is the best performing in both the weighted and unweighted approaches, it is adopted as reconstruction method in the remaining two sets of simulations.

Impact of the weights definition We consider reconstruction-driven and logistic weights. Different profiles of logistic weights are proposed by varying the rate of decay a within the set of values $\{5, 10, 15, 20, \infty\}$. Notice that $a = \infty$ corresponds to the limit condition at which the weight is a step function, taking value one up to the middle of the missing domain and falling to zero right after that point. In the computations, this condition is obtained by setting $a = 100$. Two other limit conditions are considered, namely the unweighted case – i.e., the weights have constant value 1 and the model reduces to the unweighted analysis – and the case denoted as 0-weights, where the weights are step functions falling to a small positive value (set to 10^{-7} rather than to 0 for computational reasons) at the censoring instant. This last case corresponds to the scenario where missing values are not included in the analysis. Results are reported in Figure 5. The minimum estimated MSE for all coefficients is in correspondence of logistic weights with $a = 10$. The minimum variance is 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 employment of reconstruction-driven weights decreases both the MSE and the variance of the estimates, with respect to the unweighted analysis and the analysis made solely on the observed parts of the curves. Also, this reduction is comparable to the one achieved by the logistic weights. The predictive performance of the methodology, for alternative definitions of the weights, is assessed via a leave-one-out cross-validation (LOO CV) 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. 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 extremes of associating full reliability to the curve as a whole and of completely neglecting the information of the missing trajectory. Specifically, the results suggest that a solution to the trade-off lies in the use of a finely-tuned system of weights that conveniently modulates the influence of a functional datum along the observed and missing domains.

5 Case Study

This section is devoted to the fitting of the functional GMM in equation (2). The performance of the weighted functional methodology is compared to that of the functional ordinary least squares (Section 5.1), and 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 (Section 5.2). We refer to Section 3.1 of the Supplementary Material 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), the physical interpretability of the regressors motivates the choice of keeping the functional form (2) unchanged, as

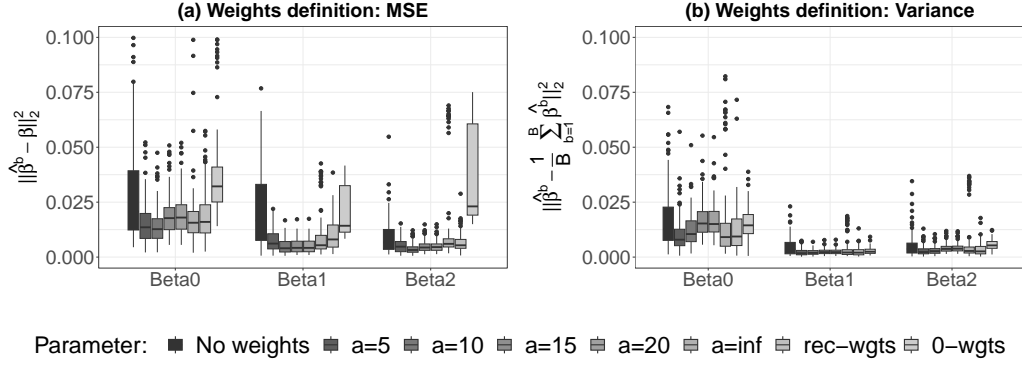


Figure 5: Boxplots of the empirical population of $\|\hat{\beta}_j^b - \beta_j\|_2^2$ (left) and of $\|\hat{\beta}_j^b - \bar{\hat{\beta}}_j^b\|_2^2$ (right), for each $\beta_j, j = 0, 1, 2$, and for the logistic and the reconstruction-driven definitions of the weighting system. For the logistic weights, larger values of a correspond to greater rates of decay of the weights; $a = \infty$ corresponds to a step function, taking value one up to the middle of the missing domain, and zero right after that point.

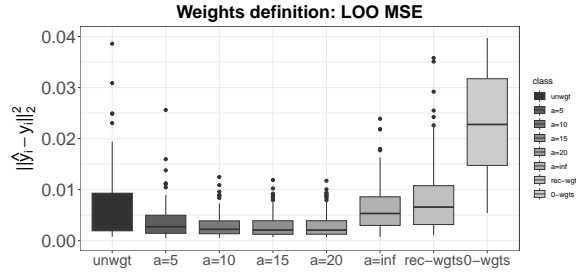


Figure 6: Boxplots of the empirical population of $\|\hat{y}_i^b - y_i\|_2^2$, obtained via LOO CV, for the logistic and the reconstruction-driven definitions of the weighting system. In the definition of the logistic weights, larger values of a correspond to greater rates of decay of the weights; $a = \infty$ corresponds to a step function, taking value one up to the middle of the missing domain, and zero right after that point.

it allows discussing on the results in seismological terms and eases the comparison with ITA18 and similar state-of-the-art GMMs (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 3.1. A penalization in the fitting criterion 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 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 reconstruction method.

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

	unweighted	$a = 5$	$a = 10$	$a = 15$	$a = 20$	$a = \infty$	0-weights
pMSE	0.1191	0.1190	0.1190	0.1189	0.1189	0.1188	0.1192
σ	0.0164	0.0164	0.0164	0.0164	0.0164	0.0164	0.0164

5.1 Model Calibration

The three steps of calibration of the functional ITA18 model are made by means of a global measure of mean inaccuracy in the prediction of the observed values of the curves. Specifically, inaccuracy in the prediction of a curve y_i is quantified as the sum of the squared distances between the true and the fitted discrete ordinates of y_i , namely $\hat{\epsilon}_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} (y_{ij} - \hat{y}_i(t_j))^2$, where N_i is the number of observed ordinates of y_i and is included as a normalization factor. We hereafter refer to the expected value of $\hat{\epsilon}_i^2$ as point-wise mean squared error (pMSE). The empirical counterpart of the pMSE is obtained via an event-wise cross-validation procedure, meaning that, at each iteration of the cross-validation, the data used for training and testing are forced to be related to independent events. As model (2) works under the ergodic assumption (Anderson and Brune, 1999), such partitioning strategy is adopted to reduce the underestimation of the pMSE.

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 characterized by the same level of regularization, a distinct penalization parameter is selected for each coefficient. The strategy adopted for the calibration of $\lambda_1, \dots, \lambda_9$ then stems from the computational burden of performing a grid search in a 9-dimensional space. We opt for an evolutionary algorithm for parameter selection, inspired by a population based training and discussed in Centofanti et al. (2023). We refer to Section 3.2 of the Supplementary Material for a detailed description of the approach and the results.

Choice of parameter a in weights definition 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 empirical pMSE 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. We recall that $a = \infty$ corresponds to giving full weight to a reconstructed trajectory up to half of the unseen domain, and 0 to the remaining part.

Selection of the reconstruction method The KL-AL method, resulting as the optimal reconstruction method from Section 4.2, is tested against a naive reconstruction suggested by the IM profiles 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 curves 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 interpolating lines from \bar{T} to $T = 10$ s. The extrapolation method results in a pMSE of 0.11880 with $\sigma = 0.01639$, while the KL-AL method is associated to a pMSE of 0.11881 with $\sigma = 0.01641$. While the validation procedure does not highlight any significant difference in the predictive performances of the two methods, we are led to extrapolate the curves in the analyses that follow, since it corresponds to the minimum pMSE.

5.2 Comparison with Scalar ITA18

Figure 7 shows the estimated regression coefficients, each one associated to the functional boxplot of a bootstrap sample of dimension $B = 1000$. The use of a bootstrap approach is justified by the argument reported in Section 1.3 of the Supplementary Material. We see the scatter of the sample around the point functional estimate as a 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 behavior. Coefficients b_1 and b_2 , in Figure 7a and 7b, capture the linear dependence of ground motion on low 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. At long periods, coefficient b_1 takes higher values than the scalar estimate. 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 producing the ITA18 estimates 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 a nonphysical

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. 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 to Section 3.3 of the Supplementary Material.

The predictive performances of the two methods are compared by means of the residual standard deviation, estimated via an event-wise 10-fold cross-validation. Since at each iteration of the cross-validation procedure and at each point of the domain the number of observed curves in the test set is much larger than that of the sampling instants, we estimate the point-wise standard deviation of the functional residuals at T_i as $\hat{\sigma}_i^2 = \frac{1}{N_i^b - q - 1} \sum_{j=1}^{N_i} (y_{ij} - \hat{y}_i(t_j))^2$, where q is the number of predictors entering (2) and N_i^b indicates the number of test curves observed at T_i at iteration b (Ramsay and Silverman (2005), Section 4.6.2). Notice that, up to a multiplicative factor, the resulting estimate is given by the square root of the pMSE. The result of the comparison is displayed in Figure 8. The two profiles of the estimated residual standard deviations almost coincide over the whole domain, only slightly differing at the right end of the domain, i.e. where the weights impact the analysis and the functional model is associated with a more regular point-wise standard deviation. This confirms that the employment of a weighted functional model does not result in a loss of predictive performance compared to the scalar model.

For the sake of completeness of the comparison, we report the computational costs of both the scalar and the functional model. We conducted the analysis on a Apple M1 processor with 8 cores and 8GB RAM. The three steps of the proposed weighted functional analysis – i.e. curves reconstruction, weighted smoothing and weighted functional regression – take less than 2 minutes overall. Estimating the residual standard deviation in event-wise 10-fold cross-validation takes 3 minutes. Similar computational times are required to fit the scalar model and compute its residual standard deviation at all periods. The greatest computational cost for the functional model occurs in the calibration phase, when the penalization parameters, the reconstruction method and the weighting system are selected in event-wise 10-fold cross-validation. Each of these jobs, though, takes no longer than 3 hours and need to be executed only once. The scalar model does not need calibration.

6 Discussion

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 smooth-

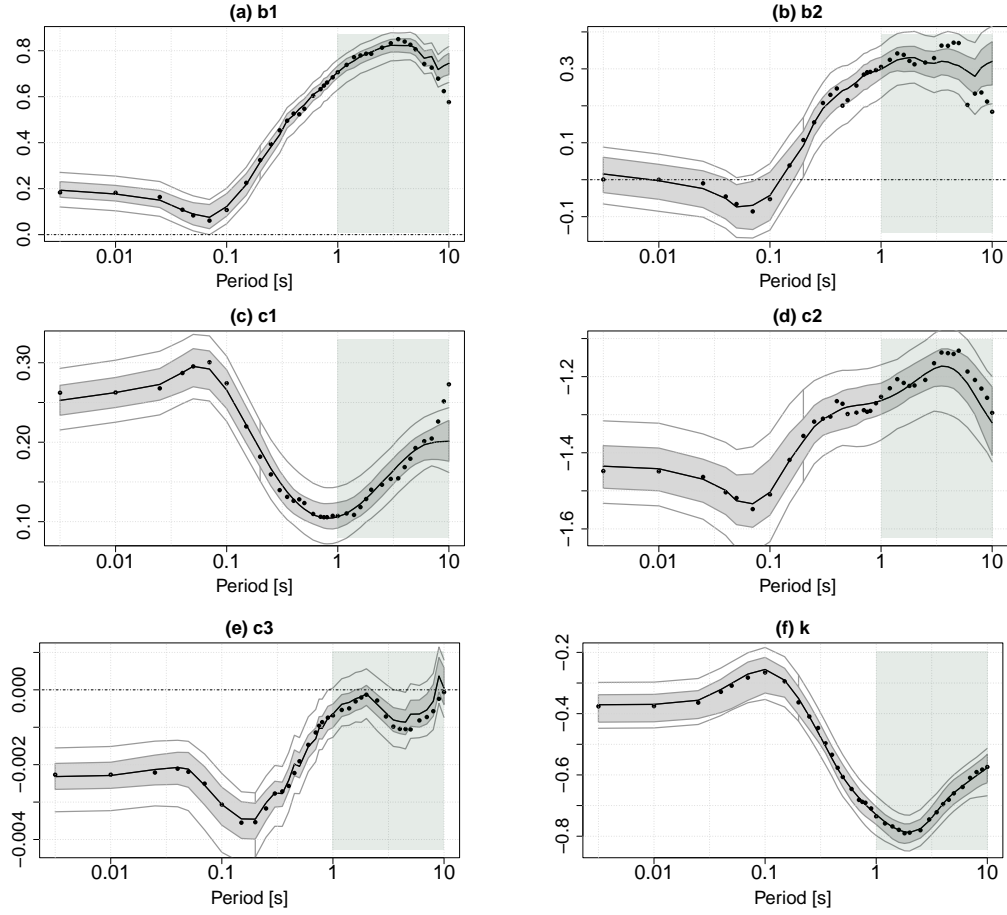


Figure 7: Functional boxplots of the estimated functional regression coefficients and comparison with the ITA18 estimates (black dots). The black lines represent the point estimates of the coefficients. The gray bands are the envelope of the 50% most central functions, with respect to the Modified Band Depth. The outer grey lines denote the fences given by the envelope of functions contained inside the central region, when it is inflated by a factor 1.5. The vertical band highlights the partially observed portion of the domain.

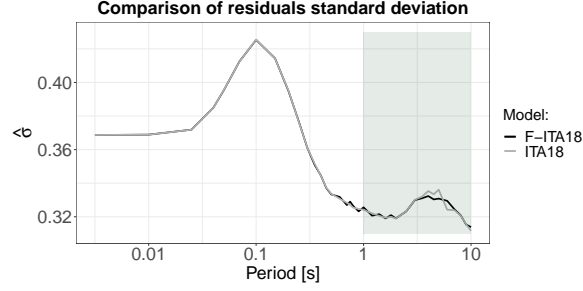


Figure 8: Comparison of the model performance between the functional model and ITA18, in terms of point-wise residual standard deviation $\hat{\sigma}$. The vertical band highlights the partially observed portion of the domain.

ing and penalized concurrent regression to the inclusion of weights, which enter the analysis by reducing the impact that the reconstructed trajectories have on the final estimates. 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. These results lead us to conclude that the optimal solution to the functional regression problem in the presence of partially observed data is found in an intermediate solution between considering only the restriction of the domain where the curves are observed, and considering the whole domain over which the curves are reconstructed. This is in agreement with the conclusions of Stefanucci et al. (2018), where a PCA-based classification is performed on the principal component scores of partially observed functional data, and the optimal classifier is found for the scores computed on an intermediate domain extension between the common and the full domain. The adoption of the weighted functional methodology introduces multiple innovative features in the context of ground motion analysis. 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 handles the cross-correlation between the spectral ordinates and provides continuous estimates over the considered range of periods. Besides, the method operates an intrinsic smoothing and stabilization of the coefficients estimates and of the spectral predictions. Future developments of the weighted methodology go in multiple directions. The definition of reconstruction-driven weights may be derived from alternative statistics summarizing the covariance structure of the reconstruction error, other than the scaled trace of the covariance function considered in this work. Since reconstruction-driven weights, unlike logistic weights, do not impose constraints on the pattern of missing data, they might be further employed in scenarios where the pattern is fragmented across the domain rather than continuous. Additional future work might test the effectiveness of the weighted smoothing step in scenarios where the response profile exhibits a rougher behaviour, and the regularization effect is thus more pronounced. In this context, it

would also become relevant to extend and adapt the weighted analysis to consider not only the uncertainty related to reconstruction, but also the one resulting from smoothing.

Further developments of the functional GMM go toward the formulation of a flexible, non-ergodic mixed-effect functional model, by addressing its existing limitations. Firstly, the fitting technique may be generalized to handle a more flexible functional form including non-linear terms. Secondly, period-continuous systematic corrective terms may be estimated with a functional mixed-effect model that accounts for station- and event-related random effects. This latter model may then be generalized to work with partially observed data and to the inclusion of functional weights. Finally, the proposed functional model for the median IM naturally combines with the functional geostatistical model for the residuals proposed in Menafoglio et al. (2020), as they jointly set up a tool for the estimation of seismic-shaking maps in a fully functional context.

Supplementary Material

Supplement: Analytical arguments, additional simulation and data analysis further supporting our conclusions. (pdf file)

Replication: Codes and data to reproduce simulation and results are in <https://github.com/tbortolotti/WFDA.git>.

Acknowledgments

Teresa Bortolotti and Alessandra Menafoglio acknowledge the support by MUR, grant Dipartimento di Eccellenza 2023–2027.

Disclosure Statement

The authors report there are no competing interests to declare.

References

- Anderson, J. G. and J. N. Brune (1999). Probabilistic seismic hazard analysis without the ergodic assumption. *Seismological Research Letters* 70(1), 19–28.
- Bindi, D., M. Massa, G. Ameri, F. Pacor, R. Puglia, and P. Augliera (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., F. Pacor, L. Luzi, R. Puglia, M. Massa, G. Ameri, and R. Paolucci (2011). Ground motion prediction equations derived from the Italian strong motion database. *Bulletin of Earthquake Engineering* 9(6), 1899–1920.
- 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(4), 1830–1835.

- Boore, D. M., J. P. Stewart, E. Seyhan, and G. M. Atkinson (2014). NGA-West2 equations for predicting PGA, PGV, and 5% damped PSA for shallow crustal earthquakes. *Earthquake Spectra* 30(3), 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(15), 1707–1721.
- Centofanti, F., A. Lepore, A. Menafoglio, B. Palumbo, and S. Vantini (2023). Adaptive smoothing spline estimator for the function-on-function linear regression model. *Computational Statistics* 38, 191–216.
- Davies, P. and M. Meise (2008). Approximating data with weighted smoothing splines. *Journal of Nonparametric Statistics* 20(3), 207–228.
- de Boor, C. (2001). *A Practical Guide to Splines*. Revised Edition, Springer, New York. (Original Edition 1978).
- 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(1), 43–104.
- Douglas, J. and B. Edwards (2016). Recent and future developments in earthquake ground motion estimation. *Earth-Science Reviews* 160, 203–219.
- Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications*, Chapter 2. Springer.
- Huang, C. and C. Galasso (2019). Ground-motion intensity measure correlations observed in italian strong-motion records. *Earthquake Engineering & Structural Dynamics* 48(15), 1634–1660.
- Kamai, R., N. A. Abrahamson, and W. J. Silva (2014). Nonlinear horizontal site amplification for constraining the NGA-West2 GMPEs. *Earthquake Spectra* 30(3), 1223–1240.
- Kneip, A. and D. Liebl (2020). On the optimal reconstruction of partially observed functional data. *The Annals of Statistics* 48(3), 1692–1717.
- Kotha, S. R., G. Weatherill, D. Bindi, and F. Cotton (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(4), 777–801.
- Kraus, D. and M. Stefanucci (2018). Classification of functional fragments by regularized linear classifiers with domain selection. *Biometrika* 106(1), 161–180.

- Lanzano, G., L. Luzi, F. Pacor, C. Felicetta, R. Puglia, S. Sgobba, and M. D’Amico (2019). A revised ground-motion prediction model for shallow crustal earthquakes in Italy. *Bulletin of the Seismological Society of America* 109(2), 525–540.
- Lanzano, G., F. Ramadan, L. Luzi, S. Sgobba, C. Felicetta, F. Pacor, M. D’Amico, R. Puglia, and E. Russo (2022). Parametric table of the ITA18 GMM for PGA, PGV and spectral acceleration ordinates. https://doi.org/10.13127/ita18/sa_flatfile/.
- Lanzano, G., S. Sgobba, L. Luzi, R. Puglia, F. Pacor, C. Felicetta, M. D’Amico, F. Cotton, and D. Bindi (2018). The pan-European Engineering Strong Motion (ESM) flatfile: compilation criteria and data statistics. *Bulletin of Earthquake Engineering* 17, 561–582.
- Menafoglio, A., S. Sgobba, G. Lanzano, and F. Pacor (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 W. Hall (1982). *Earthquake Spectra and Design*. Earthquake Engineering Research Institute, Oakland, California, U.S.A.
- Pintore, A., P. Speckman, and C. C. Holmes (2006). Spatially adaptive smoothing splines. *Biometrika* 93(1), 113–125.
- Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. Springer.
- Russo, E., C. Felicetta, M. C. D’Amico, S. Sgobba, G. Lanzano, C. Mascandola, F. Pacor, and L. Luzi (2022). Italian ACcelerometric Archive (ITACA), version 3.2. https://itaca.mi.ingv.it/ItacaNet_32/.
- Sabetta, F., A. Pugliese, F. Fiorentino, G. Lanzano, and L. Luzi (2021). Simulation of non-stationary stochastic ground motions based on recent Italian earthquakes. *Bulletin of Earthquake Engineering* 19(9), 3287–3315.
- Stefanucci, M., L. Sangalli, and P. Brutti (2018). PCA-based discrimination of partially observed functional data, with an application to AneuRisk65 data set. *Statistica Neerlandica* 72(3), 246–264.
- Worden, C. B., E. M. Thompson, J. W. Baker, B. B. A., N. Luco, and D. J. Wald (2018). Spatial and spectral interpolation of ground-motion intensity measure observations. *Bulletin of the Seismological Society of America* 108(2), 866–875.
- Yao, F., H.-G. Müller, and J.-L. Wang (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 100(470), 577–590.