mathematics

MDPI

Article

COVID-19 Data Imputation by Multiple Function-on-Function Principal Component Regression

Christian Acal †, **Manuel Escabias** †, **Ana M. Aguilera** *,† and **Mariano J. Valderrama** †

Department of Statistics and O.R. and IMAG, University of Granada, 18071 Granada, Spain; chracal@ugr.es (C.A.); escabias@ugr.es (M.E.); valderra@ugr.es (M.J.V.)

* Correspondence: aaguiler@ugr.es

† These authors contributed equally to this work.

Abstract: The aim of this paper is the imputation of missing data of COVID-19 hospitalized and intensive care curves in several Spanish regions. Taking into account that the curves of cases, deceases and recovered people are completely observed, a function-on-function regression model is proposed to estimate the missing values of the functional responses associated with hospitalized and intensive care curves. The estimation of the functional coefficient model in terms of principal components' regression with the completely observed data provides a prediction equation for the imputation of the unobserved data for the response. An application with data from the first wave of COVID-19 in Spain is developed after properly homogenizing, registering and smoothing the data in a common interval so that the observed curves become comparable. Finally, Canonical Correlation Analysis is performed on the functional principal components to interpret the relationship between hospital occupancy rate and illness response variables.

Keywords: functional data analysis; function-on-function regression; functional principal components; B-splines; COVID-19

check for updates

Citation: Acal, C.; Escabias, M.; Aguilera, A.M.; Valderrama, M.J. COVID-19 Data Imputation by Multiple Function-on-Function Principal Component Regression. *Mathematics* **2021**, *9*, 1237. https://doi.org/10.3390/math9111237

Academic Editor: Jin-Ting Zhang

Received: 21 April 2021
Accepted: 24 May 2021
Published: 28 May 2021

1. Introduction

The virus SARS-CoV-2 has been the main global concern ever since its start, at the end of 2019 in China. Its rapid propagation has put all areas of society on alert, not only the field of medicine. Nevertheless, a year and half after the beginning of the pandemic, the virus incidence has not seemed to decrease and the number of deaths continues its upward trend throughout the world. To obtain some idea of extremely negative impact of the pandemic, Coronavirus Disease (COVID-19) has caused a total of 2,780,266 deaths over the planet as of 28 March 2021, according to the real-time database developed by Johns Hopkins University [1]. Another crucial topic derived from the illness is the economic crisis which has devastated all countries. For instance, the unemployment rate is up 5.1% in last three months of 2020 in the UK, according to official data.

In order to combat this terrible situation, there is a great need to understand the development of the pandemic. Knowing its behaviour will enable correct decision making to mitigate the spread of the virus and to restore people's daily lives as soon as possible. To do so, the scientific community is focusing all its efforts on developing new techniques, capable of modelling and predicting the evolution of COVID-19. The main variables of interest that gauge how the epidemiological situation stands in a country are the number of positive, recovered and deceased cases. Another important indicator is the number of people who are hospitalized or in intensive care units. From a mathematical perspective, many authors have already attempted to tackle these variables from different statistical perspectives. A new Bayesian indicator is introduced in [2] to forecast the beginning of a new wave. In [3], semi-empirical models based on the logistic map are considered in order to predict the variables in different phases of the pandemic in Spain. Likewise, Ref. [4] apply SIR models to analyse the trend of the disease over the world and, more specifically, in India. These variables are also addressed from the time series design by considering quasi-Poisson regression and two-piece scale mixture normal distribution when there is a lack of symmetry in the error's distribution in [5,6], respectively. Additionally, Ref. [7] make an exhaustive comparison of five deep learning methods to forecast the number of new cases and recovered cases in Italy, Spain, France, China, the USA and Australia. Regarding the role of the environmental conditions in the evolution of the illness, Ref. [8] study whether the number of cases in China is connected with the daily average temperature and relative humidity through a generalized additive model. On this point, Ref. [9] show how the choice of the spatio-temporal model may affect the relationship between the spread of the virus and certain environmental conditions. Information theory metrics are also used to understand how time series associated with the pandemic are interconnected or causally related to each other [10]. In addition, how the incubation period distribution could vary by age and gender is investigated in [11]. On the other hand, a new family of distributions is introduced in [12] to model daily cases and deaths in Egypt and Saudi Arabia.

Taking the nature of the variables of interest into account, an approach based on Functional Data Analysis (FDA) is proposed in the current paper for data imputation. FDA is a modern branch of statistics that aims to analyse the information coming from curves or functions that evolve over time, space or other continuous arguments. Under this definition, it is clear that the number of COVID-19 positive, recovered, deceased, hospitalized and intensive care cases come from the observation of functional variables. FDA is usually applied in many areas of knowledge, such as Biosciences, Environment, Economy, Chemometrics and Electronics. A detailed review of the most important FDA methodologies, applications and computational aspects can be seen in books [13–17]. In this regard, some works have been developed, focused on revealing complex patterns of COVID-19 illness from an FDA viewpoint. Functional Principal Component Analysis (FPCA) and functional time series approaches based on dynamic FPCA are applied in [18] to explain variability and predict COVID-19 confirmed and death cases in the United States. On the other hand, a new Varimax rotation approach to FPCA is introduced in [19] to better interpret the main modes of variability in COVID-19 confirmed cases in the first wave in Spain. Time-varying FDA methods, to model the cumulative COVID-19 curves of cases by pooling data across countries, are applied in [20]. A multivariate FDA approach was also considered for spatio-temporal prediction of COVID-19 mortality counts in Spain [21].

All statistical models require complete and high-quality data to be able to provide accurate predictions, but, unfortunately, neither of these aspects are normally fulfilled during a pandemic. In the first wave of COVID-19 in Spain, a change in the way of recording data in some Autonomous Communities produced incomplete data in hospitalized and intensive care curves. In this paper, a functional linear regression model is proposed for the imputation of these missing curves, so that complete data are available to estimate the predictive models with guarantees. Although there are many works related to the imputation of multivariate data [22,23], there is a lot to be done in the functional framework. A novel approach for multiple imputation based on functional mixed effects models was proposed by [24] in a longitudinal data context. Different solutions to scalar-on-function regression with missing observations in the response are considered in [25–29]. Additionally, an extension to multiple functional regression imputation that handles both scalar and functional response variables related to EEG data is proposed in [30]. Likewise, different FDA imputation methods under sparse and irregular functional data settings are performed in [31]. The extension of the function-on-function linear regression (FFLR) model [32–34] to the case of multiple functional predictors is proposed in this paper to estimate the curves of hospitalized and intensive care people (functional responses) from the curves of confirmed, deceased and recovered cases (functional predictors).

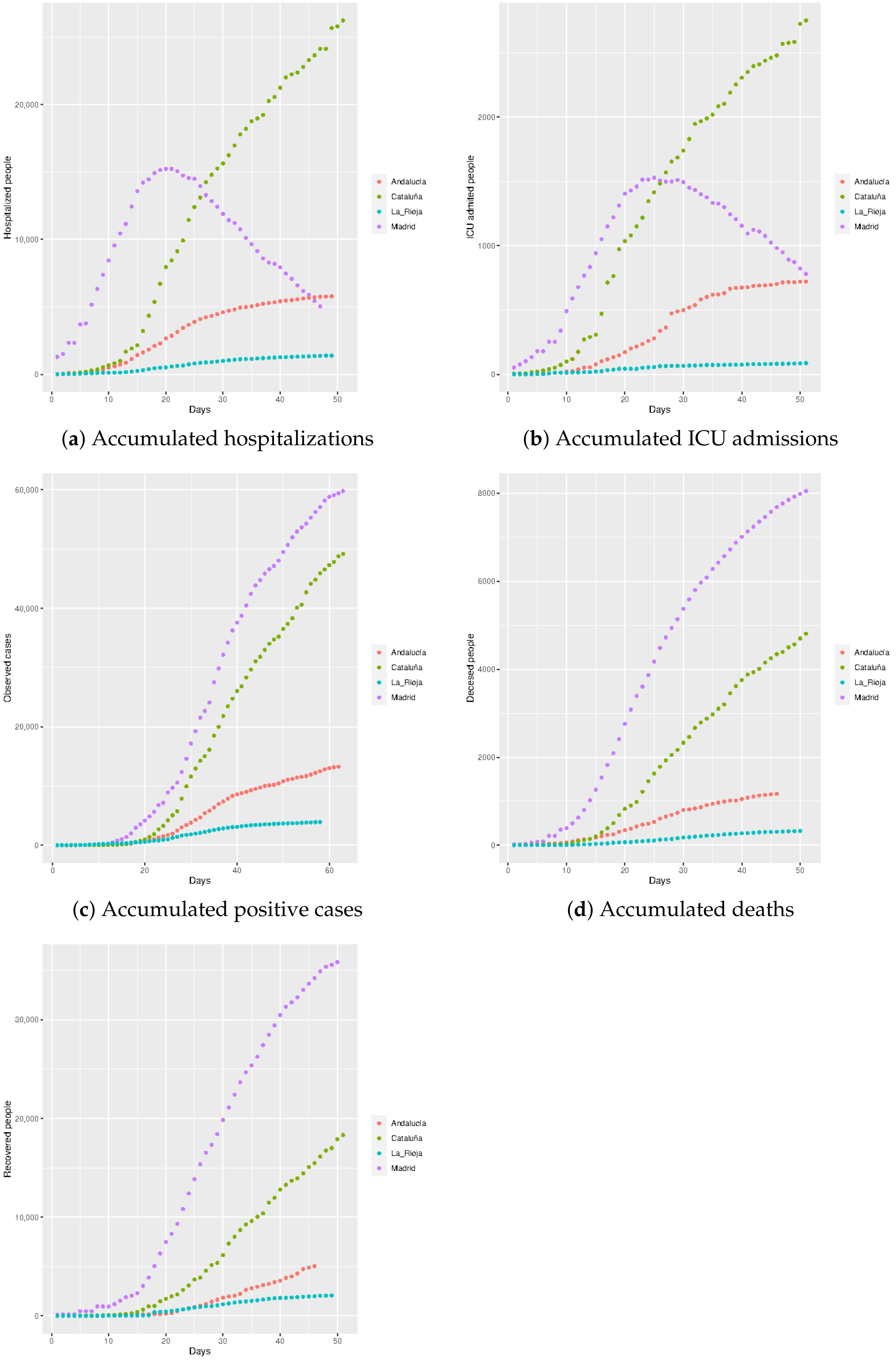
In addition to this introduction, the manuscript scheme consists of a description of the data where the process of the homogenization, registration and smoothing of the sample curves is detailed in order to make them comparable (Section 2). The theoretical framework of multiple function-on-function linear regression and the imputation procedure based on principal components regression appear in Section 3. An application to COVID-19 data in the Spanish Autonomous Communities during the first wave of the pandemic is developed in Section 4. Finally, Section 5 contains a discussion about the results obtained throughout this paper.

2. Data Homogenization, Registration and Smoothing

Spain is organized administratively in autonomous communities (ACs) or territorial governments that have transferred health affairs. This territorial organization consists of 17 ACs plus two autonomous cities (Ceuta and Melilla) located on the African continent, and which have been excluded from the analyses presented here because they have no exclusive competences in the organization of health care assumed by the Spanish government. The 17 ACs are, in alphabetical order: Andalucía, Aragón, Asturias, Islas Baleares, Islas Canarias, Cantabria, Castilla La Mancha, Castilla León, Cataluña, Extremadura, Galicia, Madrid, Murcia, Navarra, País Vasco, La Rioja and Valencia. The population is highly variable between the different ACs. While Madrid, Catalunya and Andalucía have more than six, seven and eight million inhabitants, respectively (6,663,394, 7,675,217 and 8,414,240), La Rioja has approximately three hundred thousand inhabitants (316,798).

The first wave of the COVID-19 pandemic in Spain occurred between 2 February and 27 April 2020. In those early days of the pandemic, Spanish authorities published daily and accumulated data of the evolution of the pandemic in Spain, based on the information communicated by the different ACs. Specifically, the data, published daily, correspond to the following variables: number of confirmed (positive) cases, hospitalized people, people in intensive care units (ICUs), recovered people and deceased persons. The observed data for some of the ACs can be seen in Figure 1. The problem that arose, and gave rise to this work, is that some ACs (Castilla La Mancha, Castilla León, Madrid and Galicia) modified the recording of the data associated with people in ICU and hospitalized people from a specific day (see Figure 2). The mathematical action against COVID-19 of the Spanish Mathematics Committee (http://matematicas.uclm.es/cemat/covid19/) (accessed on: 26 May 2021) called for the development of a meta-predictor (collaborative prediction) based on the predictions from different models/algorithms, contributed by interested researchers, which builds optimized combinations of them, disaggregated by ACs. Therefore, the imputation of the missing hospitalized and ICU data is fundamental to building forecasting models to provide optimal predictions of the evolution of the pandemic through these variables. In order to solve this problem, a functional regression model is proposed in this paper to estimate the expected form of the missing accumulated data of ICU admissions and hospitalizations from the observed accumulated data of cases, deaths and recoveries.

From this point, the time evolution of COVID-19 cases, deceases, recoveries, hospitalizations and ICU admission will be considered as functional variables that will be denoted as $X_1(t)$, $X_2(t)$, $X_3(t)$, $Y_1(t)$ and $Y_2(t)$, respectively (the X-variables will be considered as predictors and the Y-variables as responses in the functional regression models). The observed data are the number of daily cumulative informed values of these five functional variables for the seventeen ACs in Spain from 20 February 2020 to 27 April 2020 (see data source (https://cnecovid.isciii.es/covid19/#documentacion-y-datos) (accessed on: 26 May 2021)). Then, a random sample of curves $\{(x_{ij}(t), y_{ik}(t)) : i = 1\ldots, 17; j = 1,2,3; k = 1,2\}$ which are observed daily are available.



(**a**) Accumulated hospitalizations

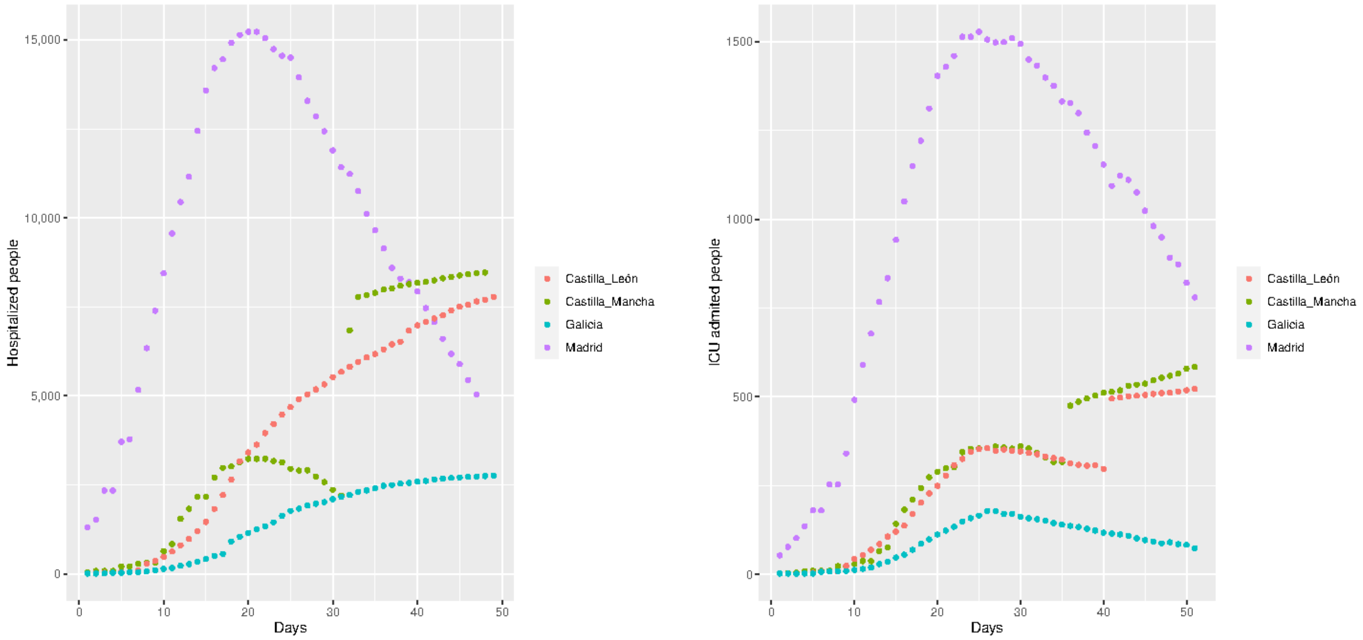
(**b**) Accumulated ICU admissions

(**c**) Accumulated positive cases

(**d**) Accumulated deaths

(**e**) Accumulated recoveries

Figure 1. Discrete daily observations of accumulated positive cases, deaths, hospitalizations, ICU admissions and recoveries in Madrid, Andalucía, Cataluña and La Rioja.



(**a**) Accumulated hospitalizations

(**b**) Accumulated ICU admissions

Figure 2. Discrete daily observations of accumulated hospitalizations and ICU admissions in Madrid, Castilla La Mancha, Castilla León and Galicia.

Before carrying out a functional analysis of the data, it is necessary to complete a data registration, given the absence of uniformity in the publication of the observations. This means that, in the same functional variable, the first day with available data and the number of discrete observations in each AC are different. For example, in Andalucía, the first recorded data of hospitalized persons were from 10 March, in which 32 hospitalized people were registered and, for this variable, there were 49 discrete observations in this AC. On the other hand, in Madrid, the first recorded data of hospitalized persons were from 12 March, in which 1304 hospitalized people were registered, and the number of discrete observations in this AC was 47. However, the curves of positive cases recorded 62 and 63 discrete observations in Andalucía and Madrid, respectively. In addition, the population size of each AC could influence the adjustments of the proposed models, as larger numbers of cases were observed in larger communities, and the different population sizes of the different ACs makes it impossible to compare data between them. In order to avoid both problems, the number of cases per 10,000 inhabitants is considered, and the first observation for each curve corresponds to the day that first exceeds the maximum of the first reported values, discarding the previous ones.

After data homogenization, the period of observation and the number of discrete observations of each functional variable for each AC continue to differ. An important constraint of FDA methods is that all sample curves of a functional variable must be observed in the same domain. Classic solutions to this problem are based on the registration of all curves in a common interval (see [13]). In this paper, we propose registering all curves in the interval $[0,1]$ by applying the FDA methodologies to the synchronized curves defined by

$$x^*_{ij}(u) = x_{ij}(T_{ij-start} + u(T - T_{ij-start})), \;\; y^*_{ik}(u) = y_{ik}(T_{ik-start} + u(T - T_{ik-start})) \; \forall u \in [0,1],$$

where $[T_{ij-start}, T]$ and $[T_{ik-start}, T]$ are the observed domains for the $i-$th predictor and the $k-$th response curves, respectively $(i = 1, \ldots, 17; j = 1,2,3; k = 1,2)$. From now on, and by abuse of a notation that helps to simplify the exposition, x_{ij} and y_{ik} will represent the registered curves.

From Discrete Daily Observations to Curves

Although the functional data are sets of curves, their true functional form is unknown and the recorded data are observations of each curve at a finite collection of timepoints. Then, the first step in FDA is to reconstruct the functional form of the curves from the observed discrete data.

There are different approaches to the processing of functional data, among which we can highlight the classic ones based on a basis representation of the curves [13] and the ones based on local–polynomial regression [16]. In this paper, basis expansions of the curves are considered by assuming that each of the functional variables $(X_1, X_2, X_3; Y_1, Y_2)$ generating the sample curves, are smooth stochastic processes with trajectories in the space $L^2([0,1])$ of squared integrable functions in the interval $[0,1]$. In what follows, the basis expansion approach is illustrated for a random sample of a functional variable, defined on a general interval T. In our dataset, this procedure must be performed on each of the five considered functional variables for which the type and dimension of the basis could be different.

Consider a random of sample curves $\{x_i(t) : i = 1, \ldots, n;\ t \in T\}$ from a functional variable X with values in $L^2(T)$, and let us assume that noisy observations x_{ik} are available for each curve at a set of time knots $t_{i1}, t_{i2}, \ldots, t_{im_i} \in T$, that is,

$$\mathrm{x}_{ik} = x_i(t_{ik}) + \epsilon_{ik} \ \ i = 1, \ldots, n;\ k = 1, \ldots, m_i.$$

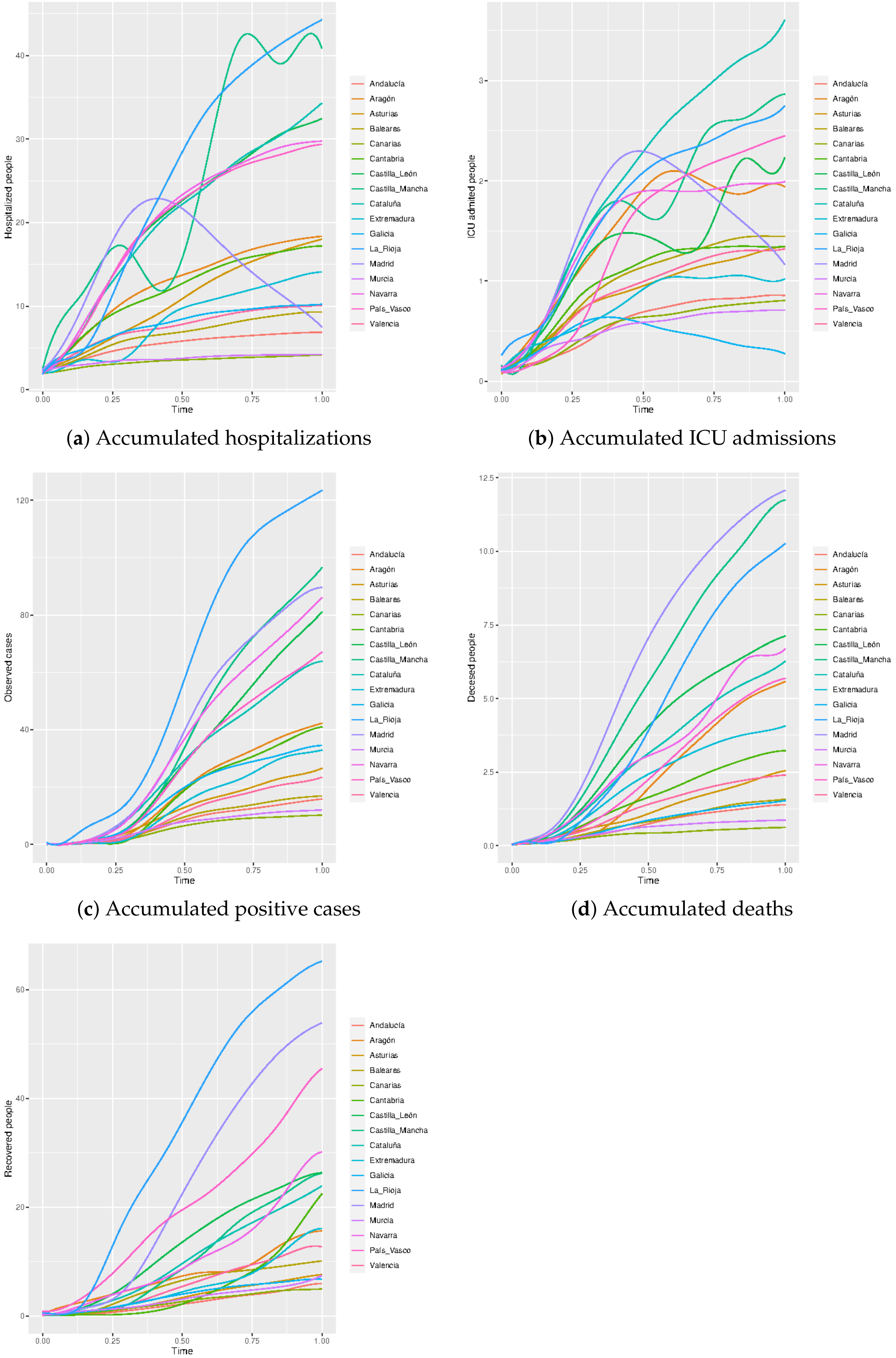
Let us also suppose that the sample curves belong to a finite-dimensional space generated by a basis of functions $\{\phi_1(t), \ldots, \phi_p(t)\}$. Therefore, each curve of the functional data set admits a basis representation in the form

$$x_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t),\ i = 1, \ldots, n. \tag{1}$$

The functional form of each curve is then determined by the vector of its basis coefficients $a_i = (a_{i1}, \ldots, a_{ip})'$, which can be estimated in different ways, with least squares approximation being the most common method, providing the following estimation: $\hat{a}_i = (\Phi_i'\Phi_i)^{-1}\Phi_i' x_i$, where $\Phi_i = (\phi_j(t_{ik}))_{m_i \times p}$, $j = 1, \ldots, p$, $k = 1, \ldots, m_i$.

The type of basis must be selected according to the characteristics of the curves in the functional dataset. The most common basis are B-splines and trigonometric functions (see, for example, [13]). The former generates spaces of spline functions, piecewise polynomial functions that are smoothly joined and have good local behaviour. The latter provides suitable spaces for periodic functions. Many other bases were used in practice, such as bases of wavelets which are more appropriate for curves with discontinuities and sharp spikes. An application of wavelet approximation from sample curves of lupus and stress level was developed in [32]. A robust estimation of the mean function, together with a simultaneous confidence band, based on polynomial spline estimation, is developed in [35].

In this paper, a basis of cubic B-splines of dimension ten with equally spaced knots was used to approximate the five samples of curves of COVID-19 from their daily discrete data. Least squares approximation was performed on each curve to estimate the basis coefficients. The cubic regression splines of all curves considered here can be seen in Figure 3.



(**a**) Accumulated hospitalizations

(**b**) Accumulated ICU admissions

(**c**) Accumulated positive cases

(**d**) Accumulated deaths

(**e**) Accumulated recoveries

Figure 3. Curves of accumulated positive cases, deaths, recoveries, hospitalizations and ICU admissions.

3. Functional Linear Regression Imputation with Missing Values in the Response

Motivated by the imputation of the missing curves of COVID-19 hospitalized and intensive care people, a functional linear regression model with functional response and several functional predictors is proposed in this paper. The general formulation of this multiple function-on-function linear regression (MFFLR) model and its estimation in terms of functional principal components regression are summarized in this section.

3.1. Multiple Function-on-Function Linear Model

The multiple function-on-function linear model allows for the estimation of a functional response Y from a vector of J functional predictor variables denoted by $X = (X_1, \ldots, X_J)'$. Let us consider a random sample from (X, Y) denoted by $\{(x_i, y_i) : i = 1, \ldots, n\}$ with $x_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})'$, and let us assume that all functional variables take values on the Hilbert space $L^2(T)$ of the squared integrable functions on the interval T, with the usual inner product defined by $< f, g >= \int_T f(t)g(t)dt, \forall t \in T$.

Then, the functional linear model is formulated as

$$y_i(t) = \alpha(t) + \sum_{j=1}^{J} \int_T x_{ij}(s)\beta_j(s,t)ds + \varepsilon_i(t),\ i = 1, \ldots, n, \tag{2}$$

where $\alpha(t)$ is the intercept function, $\beta_j(s,t)$ are the J coefficient functions and $\epsilon_i(t)$ are independent functional errors. Model (2) can be written in matrix form as

$$y_i(t) = \alpha(t) + \int_T x_i(s)'\beta(s,t)ds + \varepsilon_i(t),\ i = 1, \ldots, n,$$

where $x_i(s) = (x_{i1}(s), x_{i2}(s), \ldots, x_{iJ}(s))'$ and $\beta(s,t) = (\beta_1(s,t), \beta_2(s,t), \ldots, \beta_J(s,t))'$.

This expression considers that all functional variables are defined in the same interval T, but this is not a restriction and the model can be easily generalized for different domains in each of the functional variables. The estimation of this model is an ill-posed problem that is usually solved by least-squares penalized approaches and basis expansion of functional parameters and/or sample curves [13]. Some of the basis expansion approaches reduce the model to a multivariate linear model for the matrix of response basis coefficients in terms of the matrix of predictors' basis coefficients. The main problem is that this multivariate model is affected by high multicollinearity, which causes inaccurate estimation of the parameters. Despite the good predictive ability of the model, this fact makes its interpretation more difficult. The most-studied solutions avoid the need for cross-validation to estimate the penalty parameter by reducing the problem to linear regression on uncorrelated predictor variables. Approaches based on functional PCA [36–41] and functional Partial Least Squares (PLS) [42–47] were widely studied in the literature in the context of different functional regression models.

In this paper, a principal components' regression approach is considered. This can be seen as an extension of the principal components' prediction models developed in [48,49] to predict a functional variable in a future interval of time from its evolution in the past. In the so-called PCP models, the functional response and the functional predictor correspond to the same functional variable, but were observed in different time periods. In the present approach, truncated principal component decompositions of the functional response and the functional predictors turn the functional linear model into a multivariate linear model in terms of a reduced set of response and predictor principal components.

3.2. Functional Principal Component Regression

Let us consider the principal component decompositions of both the response and the predictor functional variables, given by

$$x_{ij}(t) = \overline{x}_j(t) + \sum_{l=1}^{n-1} \xi_{il}^{x_j} f_l^{x_j}(t), \quad y_i(t) = \overline{y}(t) + \sum_{l=1}^{n-1} \xi_{il}^{y} f_l^{y}(t), \tag{3}$$

where the principal components scores are given by

$$\xi_{il}^{x_j} =< x_{ij} - \bar{x}_j, f_l^{x_j} >= \int_T (x_{ij}(t) - \bar{x}_j(t)) f_l^{x_j}(t) dt, \quad \xi_{il}^{y} =< y_i - \bar{y}, f_l^{y} >= \int_T (y_i(t) - \bar{y}(t)) f_l^{y}(t) dt, \tag{4}$$

with the weight functions $f_l^{x_j}$ and f_l^y being the eigenfunctions of the sample covariance operators of $x_{ij}(t)$ and $y_i(t)$, respectively. The principal components scores are centered uncorrelated scalar variables with maximum variance given by the eigenvalues associated with their weight functions: $Var(\xi_{il}^{x_j}) = \lambda_l^{x_j}$, $Var(\xi_{il}^{y}) = \lambda_l^{y}$.

Theoretical and asymptotic properties of FPCA for Hilbert-valued random functions were studied in [50–54]. In the case of a basis expansion for each functional variable (see Equation (1)), each functional PCA is equivalent to the multivariate PCA of the matrix $A\Psi^{1/2}$, with $A = (a_{ij})$ being the $n \times p$ matrix of basis coefficients and Ψ being the $p \times p$ matrix of inner products between basis functions, $\Psi = (\Psi_{ij}) =< \phi_i, \phi_j >, \ i, j = 1, ..., p$. The vector of basis coefficients of the the $l-$th PC weight function $f_l(t)$ is given by $b_l = \Psi^{-1/2} v_l$, where v_l is the $l-$th eigenvector of the sample covariance matrix of $A\Psi^{1/2}$ (see [55] for a detailed study).

The principal component decompositions given in Equation (3) turn the MFFLR Model (2) into a linear regression model for each PC of the functional response Y on all PCs of the functional predictors

$$\xi_{ik}^{y} = \sum_{j=1}^{J} \sum_{l=1}^{n-1} b_{kl}^{x_j} \xi_{il}^{x_j} + \varepsilon_{ik}, \quad i = 1, \ldots, n; \ k = 1, \ldots, n-1, \tag{5}$$

with the functional coefficients given by $\beta_j(s,t) = \sum_{k=1}^{n-1} \sum_{l=1}^{n-1} b_{kl}^{x_j} f_k^{x_j}(s) f_l^{y}(t)$.

By truncating each principal component decomposition, the following principal component multiple function-on-function linear regression (PC-MFFLR) model for the functional response is obtained

$$\hat{y}_i(s) = \bar{y}(s) + \sum_{k=1}^{K} \hat{\xi}_{ik}^{y} f_k^{y}(s) = \bar{y}(s) + \sum_{k=1}^{K} \left(\sum_{j=1}^{J} \sum_{l \in L_{kj}} \hat{b}_{kl}^{x_j} \xi_{il}^{x_j} \right) f_k^{y}(s), \tag{6}$$

with $\hat{b}_{kl}^{x_j}$ being the linear least-squared estimation of the regression coefficients b_{kl}.

Different selection model approaches were developed to select the optimum PCs of each predictor variable (subsets L_{kj}), to be considered in Model (6) when it comes to estimating the first J PCs of the response variable. It is well known that PCs are ordered according to their explained variability and that the most explanatory components of the predictor variable might not be the most correlated with the response variable. In the case of the simple function-on-function linear model with only one predictor, a procedure that selects pairs of response/predictor PCs based on both explained variability and correlation was developed in [32]. A supervised version of FPCA that estimates the PCs by considering the correlation of the functional predictor and response variable was developed for the scalar-on-function regression model [56]. The usual selection models procedures based on stepwise and best subset regression, combined with cross-validation, can be adapted to this functional regression context.

3.3. Imputation of Missing Response Curves

Let us consider that all the predictor variables X_j are completely observed and only the response variable Y has missing values. Let us assume, without loss of generality, that in the sample, the first n values of the response are observed and the last m values are missing. That means that there are n complete observed curves for all variables and m incomplete observations that are missing values for the response.

In order to estimate the missing response curves, the parameters b_{kl} in Model (5) are estimated with the complete n sample curves of response and predictors. Then, the missing response curves $\{y_i^{miss}(s) : i = n+1, \ldots, n+m\}$ are estimated by computing the principal component scores of predictors $\{\xi_{il}^{x_j} : i = n+1, \ldots, n+m, l = 1, \ldots, n-1\}$ given by the Expression (4), and substituting them in the Equation (6). Then, the estimated PC-MFFLR model can be used to predict new response values Y on a test sample and to provide an accurate interpretation of the relationship between the predictor and the response variables.

If the objective is to predict the response variable in a future interval, a regression model of type (6) could be estimated to predict the response variable $Y(s)$ in the future interval of amplitude k, denoted by $[T, T+k]$, in terms of the predictor variables $(X_1(t), \ldots, X_J(t), Y(t))$ in the past interval of time $[0, T]$. In the case of the COVID-19 data, the parameter k must be selected by taking the average number of days it takes for a person to develop severe symptoms and need to be admitted to the hospital into account.

4. Covid-19 Application Results

Let us remember that the main aim of this paper is the imputation of hospitalized and intensive care curves for those ACs with missing data. To do this, multiple function-on-function linear regression approaches are developed here. In addition, a canonical correlation analysis (CCA) is performed to interpret the relationship between variables related with hospital occupation (hospitalized and intensive care people) and illness response (positive, deceased and recovered people). The computational results were obtained with the free software R ('fda' and 'yacca' R-packages for FPCA and CCA, respectively).

4.1. Data Imputation

The imputation problem is solved by applying a multiple function-on-function linear regression for each of the responses $Y_1(t)$ (hospitalized) and $Y_2(t)$ (intensive care) from the three functional predictors $X_1(t)$ (sick), $X_2(t)$ (deceased) and $X_3(t)$ (recovered). Both functional regression models are estimated from the data of the thirteen ACs with complete data (training sample). Then, the predictions for the four ACs with missing data (Castilla La Mancha, Castilla León, Galicia and Madrid) are used for data imputation.

The first step is the estimation of the functional PCs for each of the five functional predictors. As a result, the first PC explained almost all the variability in the five predictors $(99.32\%, 98.73\%, 97.97\%, 98.59\%, 96.37\%$ for X_1, X_2, X_3, Y_1, Y_2, respectively). Figures 4 and 5 show the weight functions associated to each first PC, and the perturbations of the sample mean curves obtained, by adding and subtracting a multiple of them. In order to obtain weight functions and PC scores which are much easier to interpret, two new functional Varimax rotation approaches were introduced in [19] with application for COVID-19 confirmed people.

After obtaining these functional principal components' analyses, we consider a training sample composed of all the ACs except Castilla La Mancha, Castilla León, Galicia and Madrid, which will be considered as the prediction sample.

Taking into account that the first component of $X_1(t)$, $X_2(t)$ and $X_3(t)$ were revealed to be highly and significantly correlated with the first components of $Y_1(t)$ and $Y_2(t)$; meanwhile, the other cross-correlations between PCs were not significant, and the function-on function linear regression models were reduced to the following linear models for the first PC of the response in terms of the first PC of each of the predictors

$$\hat{\xi}_{i1}^{y_k} = \gamma_0 + \xi_{i1}^{x_1}\gamma_1^{y_k} + \xi_{i1}^{x_2}\gamma_2^{y_k} + \xi_{i1}^{x_3}\gamma_3^{y_k} + \varepsilon_i^{y_k}, \quad k = 1,2;\ i = 1, \ldots, 17.$$

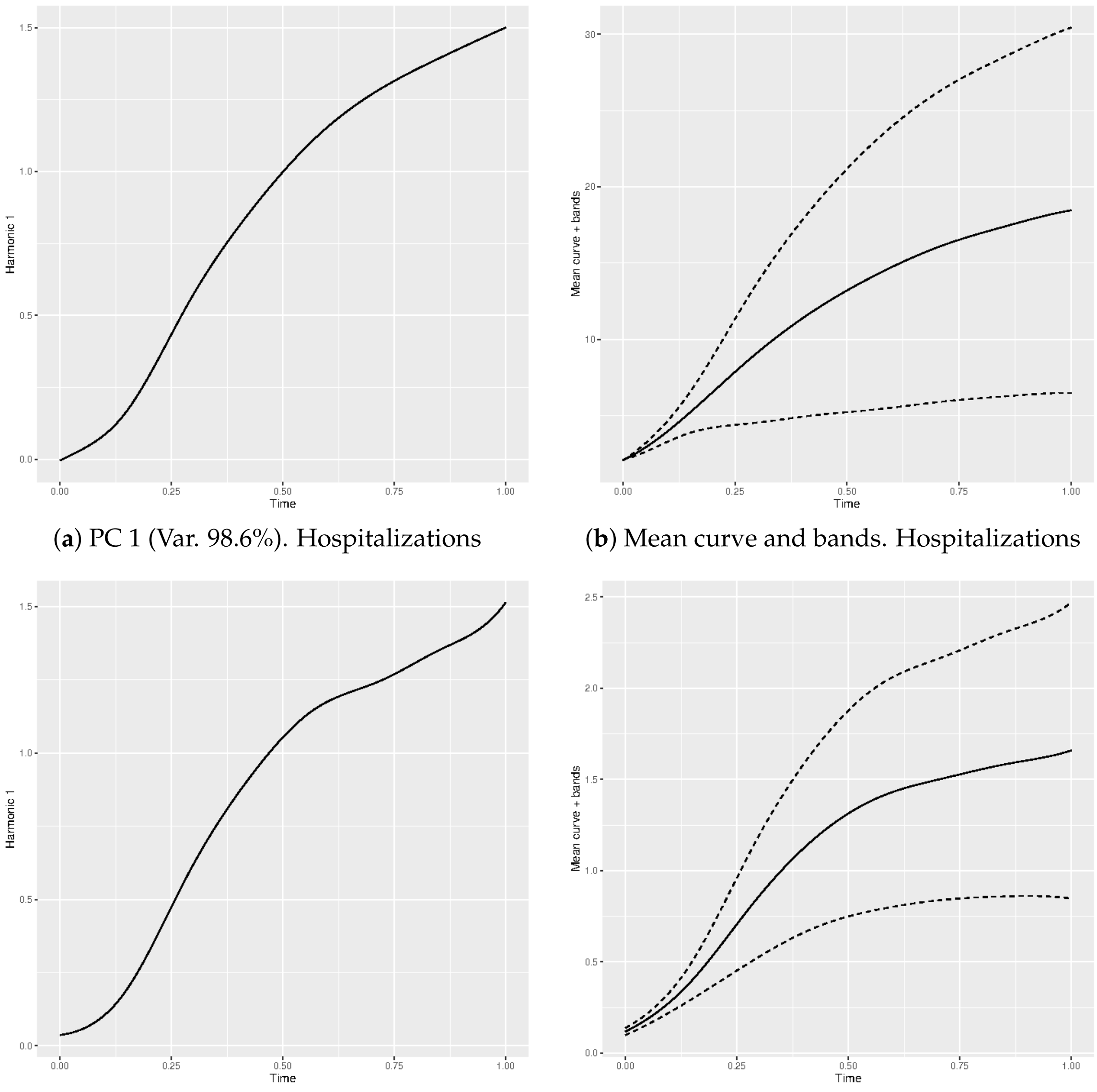
These models allow for the accurate estimation of the first component of $Y_1(t)$ and $Y_2(t)$ from the first components of $X_1(t)$, $X_2(t)$ and $X_3(t)$ with a determination coefficient of $R^2 = 0.9249$ and $R^2 = 0.7443$, respectively. Finally, the Karhunen–Loève expansion, in terms of the predictor principal components, provides the following prediction equation for $Y_1(t)$ and $Y_2(t)$

$$\hat{y}_{ik}(t) = \overline{y}_k(t) + \widehat{\xi}_{i1}^{y_k} f_1^{y_k}(t),\ k = 1, 2;\ i = 1, \ldots, 17. \tag{7}$$

In order to evaluate the prediction ability of these models, the square root of the mean squared errors between observed and predicted curves are calculated by the expression

$$RMSE(y_{ik}) = \left(\int_0^1 (y_{ik}(t) - \widehat{y}_{ik}(t))^2 dt \right)^{\frac{1}{2}}\ k = 1, 2;\ i = 1, \ldots, 13.$$

These results can be seen in Table 1, where it can be observed that the predictions for ICU admission curves are more accurate.



(**a**) PC 1 (Var. 98.6%). Hospitalizations

(**b**) Mean curve and bands. Hospitalizations

(**c**) PC 1 (Var. 96.4%). ICU admisions

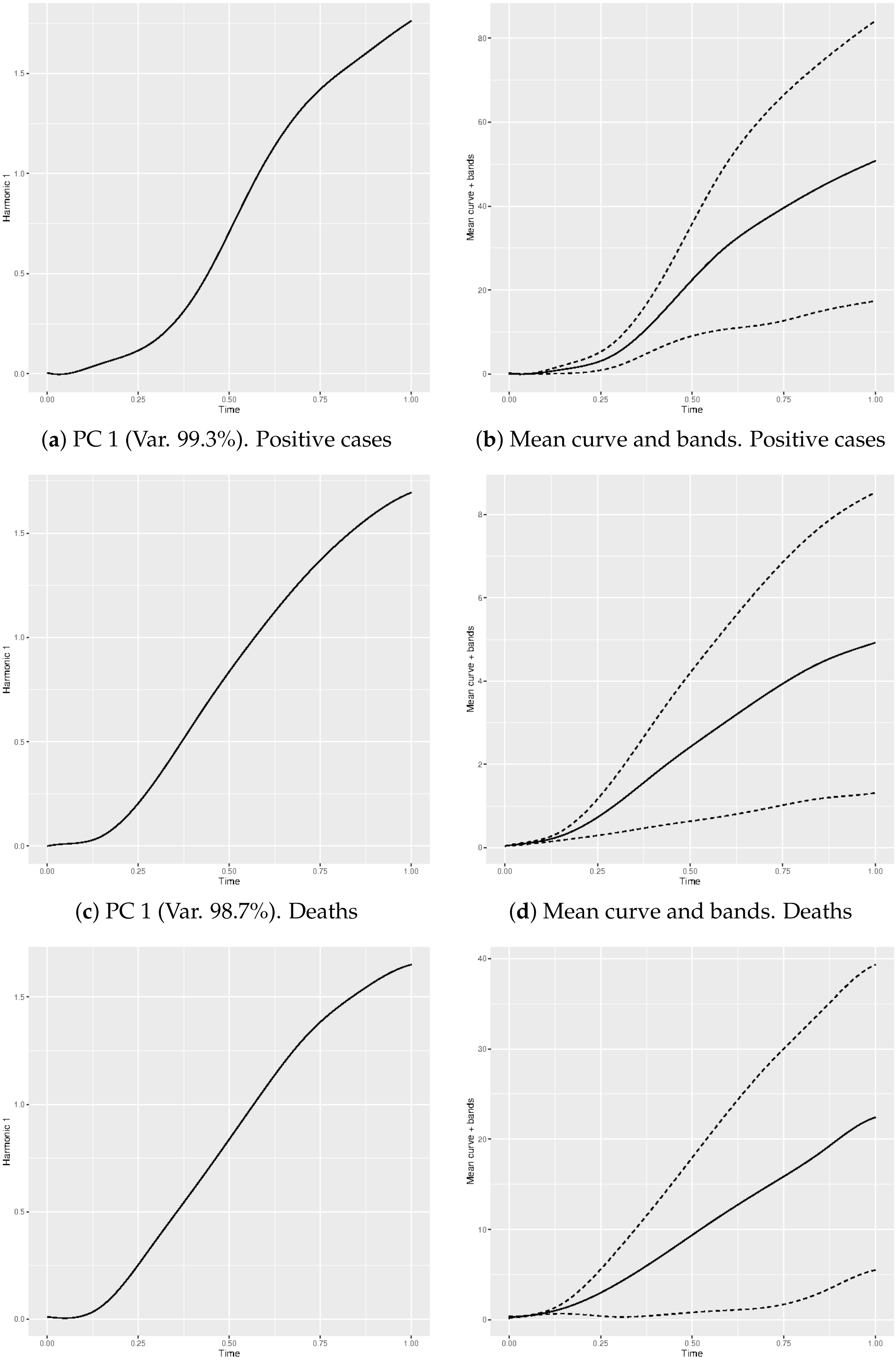
(**d**) Mean curve and bands. ICU admisions

Figure 4. First weight PC functions (**left**), sample mean curves and the perturbations for each functional response (**right**) $\bar{y}_k \pm 2\sqrt{\lambda_1^{y_k} f_1^{y_k}}; k = 1, 2.$

Some of the observed and estimated training curves can be seen in Figures 6 and 7 next to confidence bands for the predicted curves. These confidence bands are obtained by pointwise confidence intervals, computed for each fixed timepoint t_p as follows

$$\widehat{y}_{ik}(t_p) \pm 2 \times \widehat{SE}(\widehat{y}_{ik}(t_p)),\ k = 1, 2,$$

where $\widehat{SE}(\widehat{y}_{ik}(t_p)) = \widehat{SE}(\widehat{\xi}^{y_k}_{ik}) f^{y_k}_1(t_p),\ k = 1, 2$ with $\widehat{SE}(\widehat{\xi}^{y_k}_{ik})$ being the standard error of the PC prediction given by the corresponding multiple linear regression fit.



(**a**) PC 1 (Var. 99.3%). Positive cases

(**b**) Mean curve and bands. Positive cases

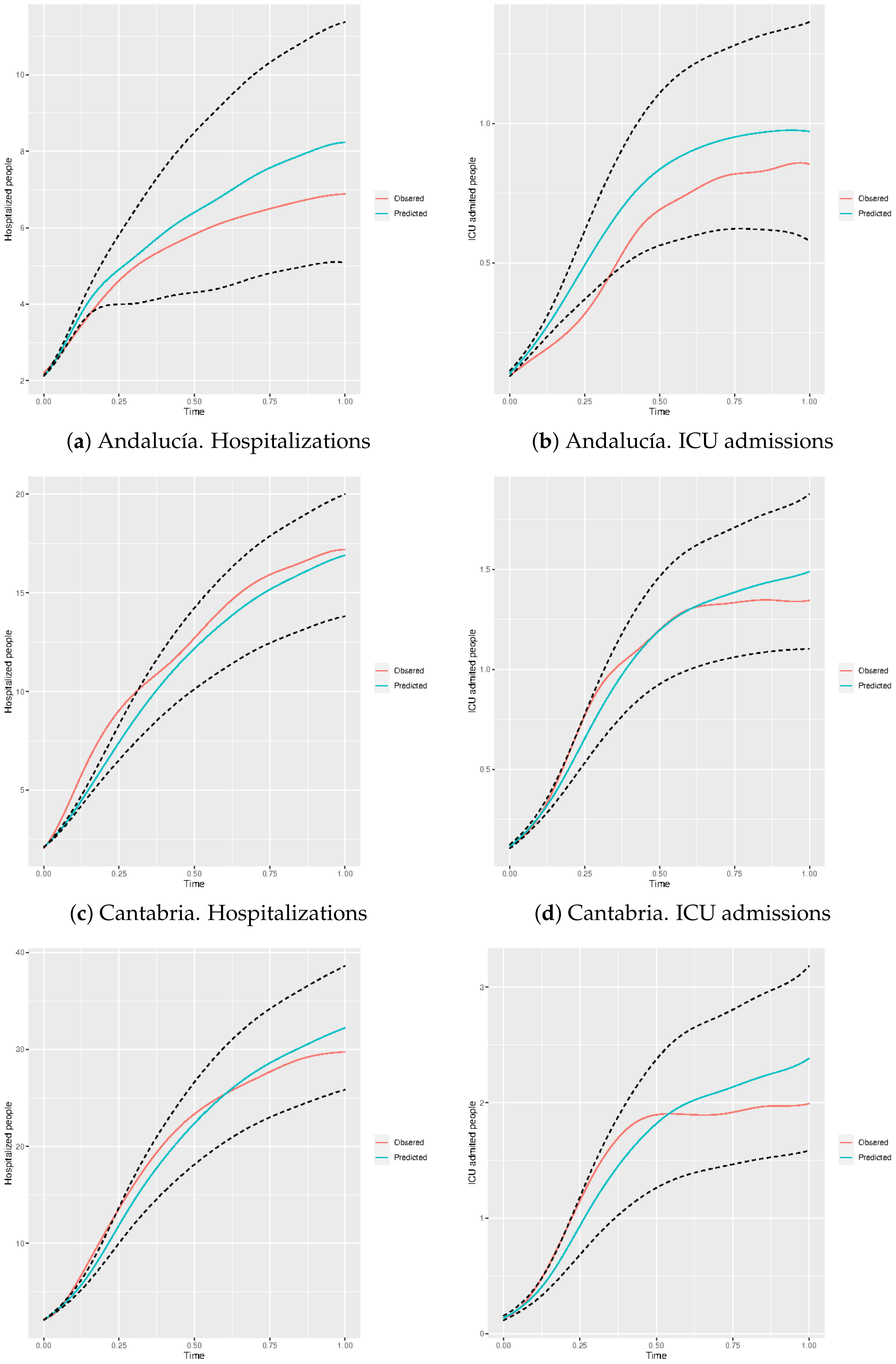
(**c**) PC 1 (Var. 98.7%). Deaths

(**d**) Mean curve and bands. Deaths

(**e**) PC 1 (Var. 98.0%). Recoveries

(**f**) Mean curve and bands. Recoveries

Figure 5. First weight PC functions (a, c and e), sample mean curves and the perturbations (b, d and f) for each functional predictor $\bar{x}_j \pm 2\sqrt{\lambda_1^{x_j}} f_1^{x_j}$, $j = 1, 2, 3$.



(**a**) Andalucía. Hospitalizations

(**b**) Andalucía. ICU admissions

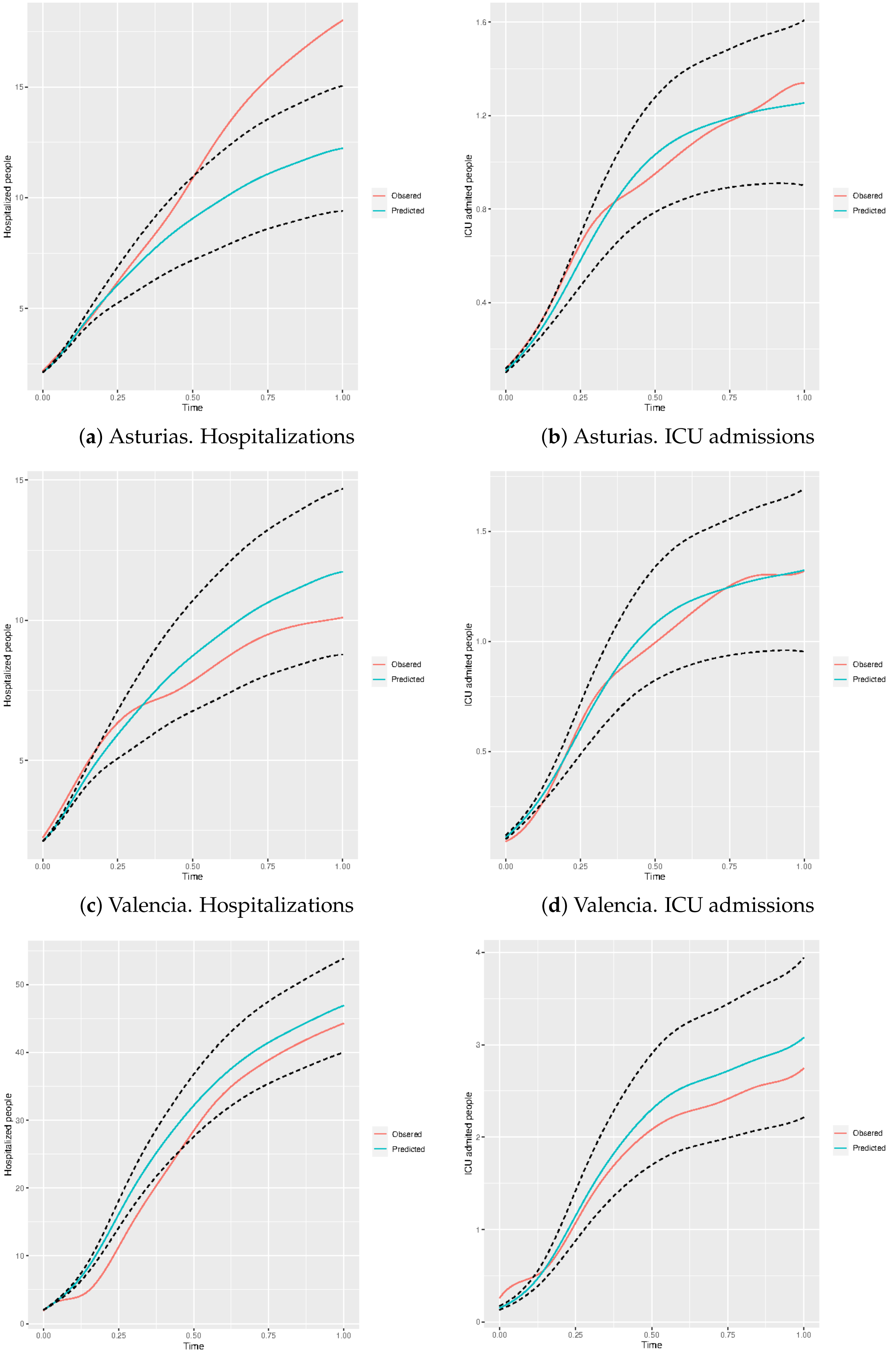
(**c**) Cantabria. Hospitalizations

(**d**) Cantabria. ICU admissions

(**e**) Navarra. Hospitalizations

(**f**) Navarra. ICU admissions

Figure 6. Observed and predicted curves (with pointwise confidence bands) of hospitalizations (left) and ICU admissions (right) in some of the training ACs: Andalucía, Cantabria and Navarra.



(**a**) Asturias. Hospitalizations

(**b**) Asturias. ICU admissions

(**c**) Valencia. Hospitalizations

(**d**) Valencia. ICU admissions

(**e**) La Rioja. Hospitalizations

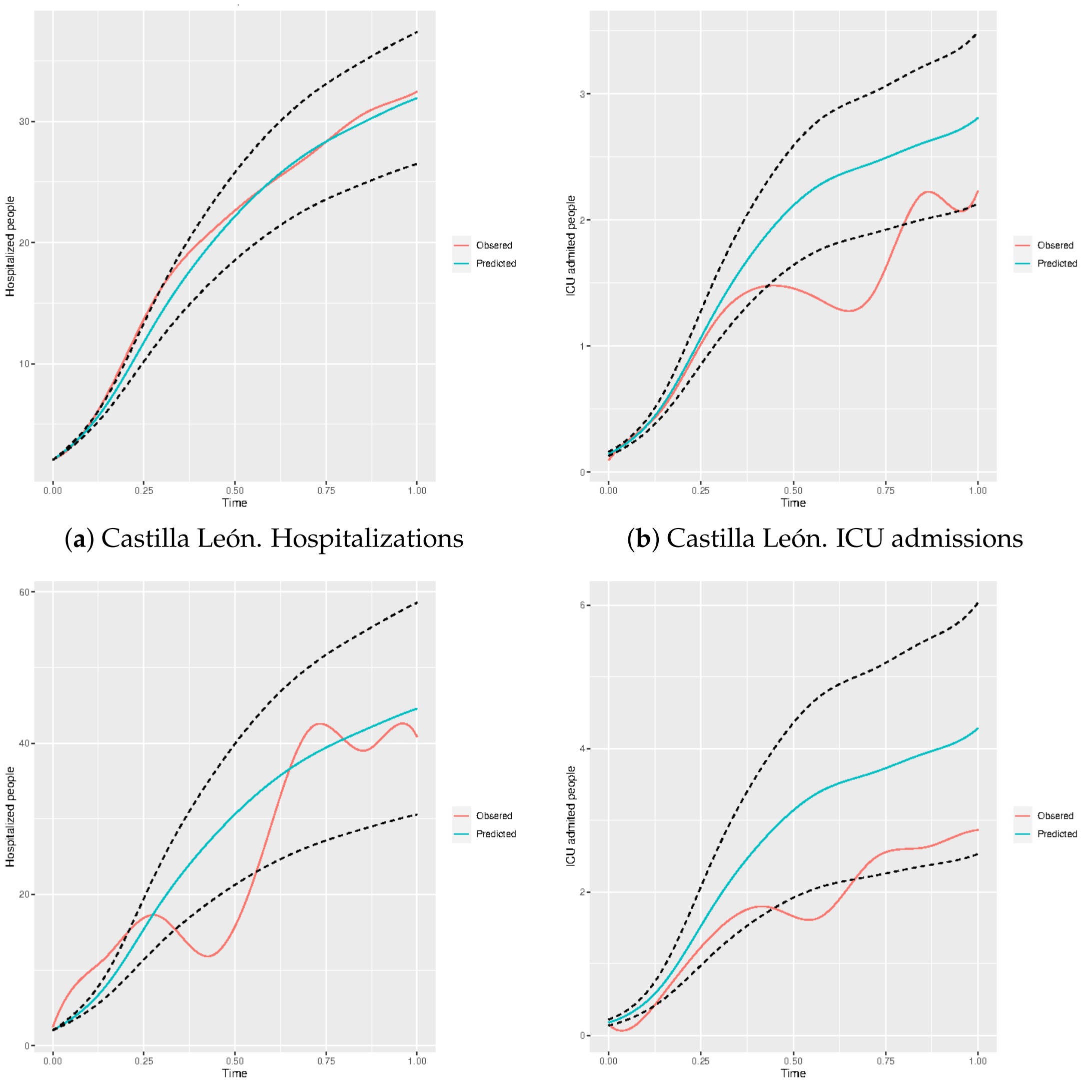
(**f**) La Rioja. ICU admissions

Figure 7. Observed and predicted curves (with pointwise confidence bands) of hospitalizations (**left**) and ICU admissions (**right**) in some of the training ACs: Asturias, Valencia and La Rioja.

Table 1. Root mean squared prediction errors for hospitalizations (y_{i1}) and ICU admissions (y_{i2}) curves in the different training ACs.

AC	$RMSE\ (y_{i1})$	$RMSE\ (y_{i2})$
Andalucía	0.77577948	0.13645363
Aragón	1.51075014	0.16559390
Asturias	3.05564828	0.05135305
Islas Baleares	0.47397162	0.30168996
Islas Canarias	1.17563681	0.04168788
Cantabria	0.91993038	0.06896749
Catalunya	3.45656121	0.62176527
Valencia	0.92591220	0.04144271
Extremadura	3.30961380	0.59974926
Murcia	1.62014752	0.11596922
Navarra	1.26109742	0.20248242
País vasco	4.46884752	0.30765768
La Rioja	3.37692798	0.22644853

Finally, the expected curves provided by the regression models in Equation (7) for hospitalizations and ICU admissions in the badly recorded ACs, next to their confidence bands and observed curves, are drawn in Figures 8 and 9.



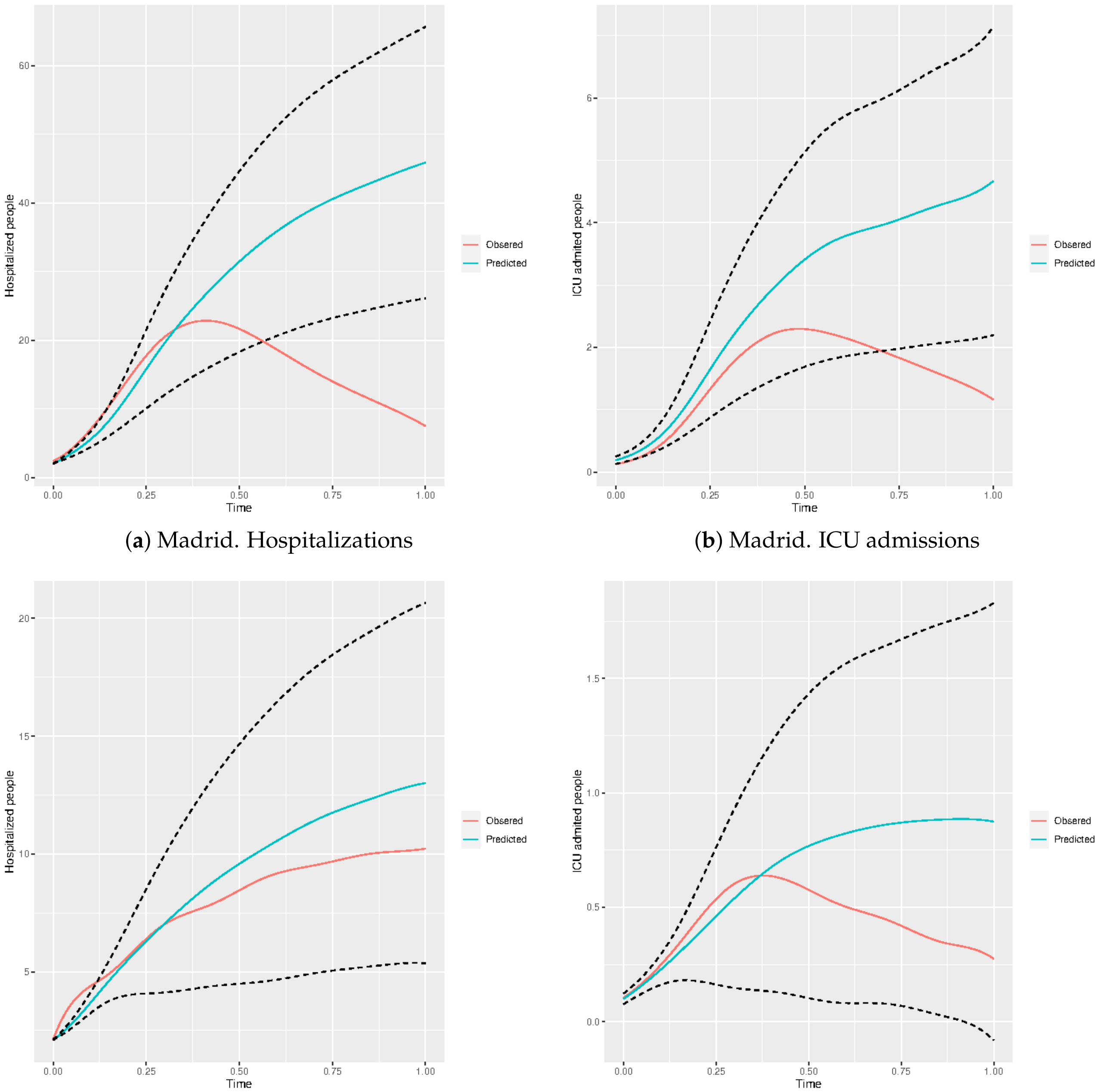
(**a**) Castilla León. Hospitalizations

(**b**) Castilla León. ICU admissions

(**c**) Castilla La Mancha. Hospitalizations

(**d**) Castilla La Mancha. ICU admissions

Figure 8. Observed and predicted curves (with pointwise confidence bands) of hospitalizations (left) and ICU admissions (right) in Castilla León and Castilla La Mancha.



(**a**) Madrid. Hospitalizations

(**b**) Madrid. ICU admissions

(**c**) Galicia. Hospitalizations

(**d**) Galicia. ICU admissions

Figure 9. Observed and predicted curves (with pointwise confidence bands) of hospitalizations (**left**) and ICU admissions (**right**) in Madrid and Galicia.

The prediction of the missing curves using the PC-MFFLR considered models provides a pointwise estimation of hospitalizations and ICU admissions that corrects the inaccurate reported data. These pointwise predictions, next to their anomalous values for the first and last days of the first wave of COVID-19 in Castilla La Mancha, Castilla León, Madrid and Galicia, can be seen in Table 2.

Table 2. Pointwise imputation of hospitalizations and ICU admissions for the first and last days of the first COVID-19 wave in Castilla La Mancha, Castilla León, Madrid and Galicia.

Hospitalizations								
	Castilla La Mancha		**Castilla León**		**Galicia**		**Madrid**	
Time	**Obs**	**Pred**	**Obs**	**Pred**	**Obs**	**Pred**	**Obs**	**Pred**
1	635	413	476	495	557	570	1518	1351
2	838	565	629	630	906	676	2337	1779
3	1547	735	798	784	1043	809	2337	2247
4	1826	932	977	961	1147	961	3710	2772
5	2162	1164	1197	1163	1250	1120	3778	3371
6	2162	1436	1457	1394	1338	1276	5168	4059
7	2707	1758	1823	1656	1447	1424	6338	4853
8	2977	2124	2214	1948	1630	1564	7388	5768
9	3018	2520	2648	2259	1767	1698	8441	6794
...	...	...	...	...	...	...	...	...
...	8173	8200	7080	6970	2609	3171	8191	28,159
...	8199	8317	7174	7064	2652	3222	7930	28,482
...	8243	8430	7264	7155	2674	3270	7464	28,802
...	8304	8542	7397	7246	2694	3316	7077	29,120
...	8342	8654	7506	7336	2707	3362	6601	29,434
...	8385	8763	7555	7424	2722	3407	6183	29,740
...	8417	8868	7653	7508	2735	3449	5892	30,037
...	8444	8969	7703	7586	2746	3484	5441	30,320
...	8464	9062	7777	7658	2758	3511	5039	30,587

ICU Admissions								
	Castilla La Mancha		**Castilla León**		**Galicia**		**Madrid**	
Time	**Obs**	**Pred**	**Obs**	**Pred**	**Obs**	**Pred**	**Obs**	**Pred**
1	23	37	24	35	29	27	77	127
2	23	45	43	44	35	35	102	152
3	29	56	54	54	47	44	135	184
4	37	70	69	67	55	53	180	224
5	37	88	85	83	69	63	180	273
6	65	110	106	102	86	74	253	332
7	76	136	120	124	98	85	253	404
8	142	167	137	150	112	96	340	488
9	182	202	170	178	123	107	491	587
...	...	...	...	...	...	...	...	...
...	531	784	501	615	108	237	1111	2826
...	534	793	503	622	101	237	1076	2853
...	537	801	505	628	96	238	1024	2878
...	546	809	508	634	92	239	981	2903
...	553	817	510	639	87	239	949	2930
...	559	826	511	645	90	239	892	2962
...	565	838	515	653	85	239	873	3001
...	579	852	518	662	83	238	821	3050
...	584	871	522	674	73	236	780	3111

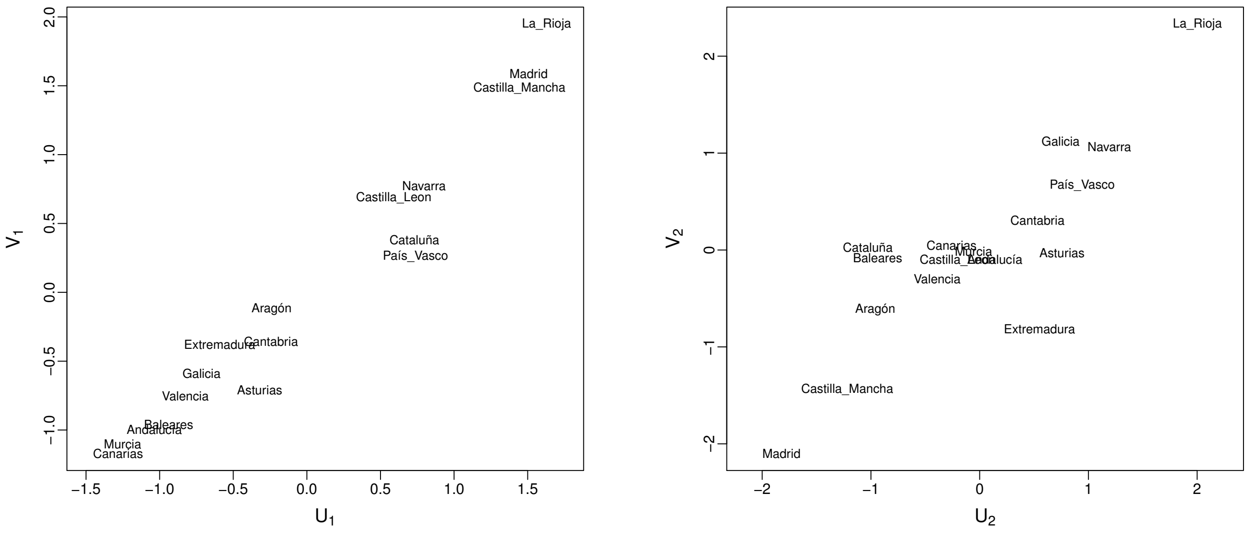
The obtained predictions can be considered as an imputation of the real behaviour of these curves in the observation period if the mode of data communication did not change. Thus, in Castilla La Mancha on 27 April 8464, people were reported as hospitalized and the imputation provided by the model is 9062 cases; in Castilla León, 7777 versus 7658; in Galicia, 2758 versus 3511; in Madrid, 5039 versus 30,587. For ICU admissions, the differences between the registered and imputed cases are, again, evident. It can be seen that, in Castilla La Mancha on 27 April, 584 people were registered as admitted in ICU and the model gives an estimation of 871; in Castilla León, 522 versus 674; in Galicia, 73 versus 236; in Madrid, 780 versus 3111.

4.2. Canonical Correlation Analysis

Once the missing data were imputed and complete curves are available for the 17 ACs, the relationship between the variables related to the number of people admitted to hospitals (hospitalized and ICU people) and the ones affected by the disease (sick, deceased and recovered) can be studied. Canonical Correlation Analysis (CCA) was applied on the two sets of first principal components associated with these functional variables to explore this relationship without necessarily distinguishing between independent and dependent variables. The analysis makes sense because the correlations between PCs in the two groups are very high, suggesting that the variables are not linearly independent.

In agreement with the above, the first principal component of each functional variable is selected to carry out the analysis. Thus, the dataset consists of a sample of the seventeen Spanish ACs in an attempt to determine which factors influence in the hospital occupancy rate. The two groups of variables are, on the one hand, Hospital occupancy rate (HOR) formed by the first PC of hospitalized people and of ICU people $(\hat{\xi}_1^{y_1}, \hat{\xi}_1^{y_2})$, and, on the other hand, Illness response (IR) comprised by the first PC of positive people, deceased people and recovered people $(\hat{\xi}_1^{x_1}, \hat{\xi}_1^{x_2}, \hat{\xi}_1^{x_3})$. The estimates of the squared canonical correlations between the two canonical variables for each pair appear in Table 3, next to the outcomes associated with the Barlett's test for testing the null hypothesis that the two canonical variate pairs are uncorrelated. As a result, it can be concluded that both canonical pairs are significantly correlated and dependent on each other (there is a relationship between the two sets of variables).

Note that the squared canonical correlations represent, for each pair, the percentage of variance in one canonical variate explained by the variation in the other one, but say nothing about the extent to which the canonical variates themselves account for variation in the original variables. Then, around 95.4% of the variation in the first canonical variate for HOR (U_1) is described by the variation in the first canonical variate for IR (V_1), and almost 71% of the variation in U_2 is explained by V_2. This fact suggests that both canonical correlations are important. Figure 10 displays how the values of the canonical variates are spread over the plane. The linear relation in each pair is clearly visible in these scatterplots. Likewise, it is possible to draw conclusions about which ACs behave similarly during the first wave. The results are in accordance with multiple studies about the COVID-19 pandemic in Spain (see, for example, [19]).



(**a**) First canonical variate pair (**b**) Second canonical variate pair

Figure 10. Scatterplot for the first and second canonical variate pairs.

Table 3. Estimates of the canonical correlations next to χ^2 values associated with Bartlett's omnibus statistic, degrees of freedom and p-values for each canonical variate pair.

Canonical Correlation	Squared Canonical Correlation	Stat	df	*p*-Value
0.9765693	0.9536876	55.82721	6	<0.001
0.8398669	0.7053764	15.88673	2	<0.001

Additionally, the estimated canonical coefficients (loadings) for the HOR and RI variables are in Tables 4 and 5, respectively. The magnitudes of these coefficients give the contributions of the individual variables to the corresponding canonical variable. Hence, the canonical variables are determined as follows

$$\begin{aligned} U_1 &= -0.1767528 \times \xi_1^{y_2} + 0.1214623 \times \xi_1^{y_1} \\ U_2 &= -3.2637387 \times \xi_1^{y_2} + 0.2556912 \times \xi_1^{y_1} \\ V_1 &= 0.0336045 \times \xi_1^{x_1} + 0.1877252 \times \xi_1^{x_2} - 0.0044276 \times \xi_1^{x_3} \\ V_2 &= 0.1094736 \times \xi_1^{x_1} - 1.0135127 \times \xi_1^{x_2} + 0.0047968 \times \xi_1^{x_3} \end{aligned}$$

Once the raw canonical coefficients have been estimated, the following step is to interpret each canonical component. For that purpose, the squared correlations between the variables in each group and the canonical components are computed in Table 6 and 7 for the HOR and IR groups, respectively. These parameters indicate the fraction of HOR and IR variance associated with each of their components separately. Let us observe that, for the second canonical variables (U_2, V_2), none of the correlations are large, so this pair provides very little information about the variables. Regarding the first canonical variate pair (U_1, V_1), all the correlations with the variables are uniformly high. This means that U_1 and V_1 are an overall measure of HOR and IR variables, respectively, with U_1 being highly correlated with hospitalizations and V_1 more correlated with positive cases and deceased people.

Table 4. Canonical coefficients for HOR variables.

	U_1	U_2
ICU	−0.1767528	−3.2637387
Hospitalized	0.1214623	0.2556912

Table 5. Canonical coefficients for IR variables.

	V_1	V_2
Cases	0.0336045	0.1094736
Deceased	0.1877252	−1.0135127
Recovered	−0.0044276	0.0047968

These outcomes expose that the level of saturation in the hospitals are particularly determined by the number of hospitalized people; meanwhile, the response to the pandemic is governed by the number of positive cases and deaths. Despite the fact that the number of people in UCI and the number of recovered people play an important role in the canonical variates, their contribution is smaller.

Table 6. Squared correlations between the HOR variables and the canonical variables.

	U_1	U_2	V_1	V_2
ICU	0.8158880	0.184111953	0.7781023	0.129868216
Hospitalized	0.9970756	0.002924353	0.9508986	0.002062769

Table 7. Squared correlations between the IR variables and the canonical variables.

	V_1	V_2	U_1	U_2
Cases	0.9660682	0.03366429	0.9213272	0.02374600
Deceased	0.9269440	0.07237154	0.8840149	0.05104917
Recovered	0.7485881	0.04012734	0.7139192	0.02830487

Finally, a canonical redundancy analysis was performed in order to study the percentage of variance of one group of variables which is accounted for by the other (in the usual least squares sense). The results of this analysis can be seen in Table 8, and the correlations between each set of variables and the opposite group of canonical variates in Tables 6 and 7. Table 8 shows that both components of the first canonical pair are a good overall predictor of the opposite set of variables, since the explained proportions of variance for HOR and IR are 0.864 and 0.839, respectively. Nevertheless, despite the significant correlation for the second pair, these variables do not account for a great amount of variability. This statement is corroborated by the squared correlations displayed in Table 6 and 7. These measures indicate that the first canonical variate of the IR group has an outstanding predictive power for the number of hospitalized (95.09%) and a considerable influence for the number of people in ICU (77.81%) as well. Similar interpretations are reached for the first canonical variable of HOR, which is a superb predictor of the number of cases and deaths (92.13% and 88.40%, respectively), and to lesser extent, of the number of recuperated (71.39%). The second canonical variables add virtually nothing given that the fraction of variance in each variable set attributable to the other group through the respective canonical variates barely overcomes 10% of the total variability.

Table 8. Total fraction of HOR (IR) variance accounted by IR (HOR) variables, through each canonical variate in first row (second row).

	U_1	U_2	V_1	V_2
HOR	-	-	0.86450046	0.06596549
RI	0.83975376	0.03436668	-	-

5. Conclusions

The current economic and sanitary crisis provoked by the virus SARS-CoV-2 has taken up most of the planet's attention since the World Health Organization declared a worldwide emergency state in the middle of March 2020. In order to control the propagation of the virus, the scientific community is immersed in the development of statistical models that enable governments to control the behaviour of the pandemic and to mitigate the devastating effects of the COVID-19 illness. Thus, it is essential to build powerful models to guarantee accurate predictions. Taking into account the nature of the variables of interest (for instance, number of positive cases, deceases, recovered, hospitalised and people in intensive care units), a wide variety of models have been tackled by considering Functional Data Analysis methodologies. Nevertheless, the good performance of these models depends on the quality of the data, which is not always as good as one might expect, especially in periods of pandemic, where the data are usually incomplete. On this matter, an extension of function-on-function linear regression is proposed for the imputation of missing values in the response, where the functional coefficients are estimated by means of principal components regression. The motivation for this work is to forecast the curves of hospitalized and intensive care people (functional responses) from the curves of positive cases, deaths and recoveries (functional predictors) for several Spanish Autonomous Communities that changed the means of recording data related to hospital occupancy rate. The imputation of these curves is made once the linear model is estimated, with a training sample composed by the remainder of communities that did not modify their way of registering data. The performance of the model is outstanding for the training sample, since the observed and predicted curves are very similar for both functional responses.

Regarding the prediction sample, the obtained forecasts can be considered as an imputation of what should have been the real behaviour of these curves in the observation period if the mode of data communication did not change. It can be observed that the model captures the trend of the curves up to the change. Additionally, once the missing data were imputed, a canonical correlation analysis was carried out in order to study the possible relationship between the two groups of variables: hospital occupancy rate (number of hospitalized people and ICU admissions) and illness response (number of positive cases, deaths and recovered people). The first principal component score of each variable was selected to form the canonical analysis, since only the first principal component explains almost all the variability in the five functional variables. After an exhaustive analysis, both sets of variables were shown to be highly correlated with each other and, moreover, each of the first canonical variables is a good overall predictor of the opposite group of variables. At this point, the variables with more predictive power are the number of hospitalizations, positives and deceases. In sum, the present document introduces a new mechanism for the imputation of missing at random functional response curves and shows the relationship among interesting functional variables associated with the COVID-19 pandemic.

Author Contributions: All authors contributed equally in developing the methodology, implementing the proposed estimation algorithms with the software R, performing the simulation study and application, as well as in writing this manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by project MTM2017-88708-P of the Spanish Ministry of Science, Innovation and Universities (also supported by the FEDER program), project FQM-307 of the Government of Andalusia (Spain) and the Ph.D. grant (FPU18/01779) awarded to Christian Acal.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available at https://drive.ugr.es/index.php/s/7VWt12CYdiAIIYt.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AC	Autonomous Community
CCA	Canonical Correlation Analysis
FPCA	Functional Principal Component Analysis
FDA	Functional Data Analysis
HOR	Hospital occupancy rate
IR	Illness Response
MFFR	Multiple Function-on-Function Regression
PC	Principal Component
PCA	Principal Component Analysis
PLS	Partial Least Squares
PC-MFFR	Principal Components Multiple Function-on-Function Regression

References

1. Dong, E.; Du, H.; Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **2020**, *20*, 533–534. [CrossRef]
2. Berihuete, A.; Sanchez-Sanchez, M.; Suarez-Llorens, A. A Bayesian Model of COVID-19 Cases Based on the Gompertz Curve. *Mathematics* **2021**, *9*, 228. [CrossRef]
3. Mora, J.C.; Pérez, S.; Dvorzhak, A. Application of a Semi-Empirical Dynamic Model to Forecast the Propagation of the COVID-19 Epidemics in Spain. *Forecasting* **2020**, *2*, 452–469. [CrossRef]
4. Agarwal, P.; Jhajharia, K. Data analysis and modeling of COVID-19. *J. Stat. Manag. Syst.* **2021**, *24*, 1–16. [CrossRef]

5. Tobias, A. Evaluation of the lockdowns for the SARS-CoV-2 epidemic in Italy and Spain after one month follow up. *Sci. Total Environ.* **2020**, *725*, 138539. [CrossRef] [PubMed]
6. Maleki, M.; Mahmoudi, M.R.; Heydari, M.H.; Pho, K.H. Modeling and forecasting the spread and death rate of coronavirus (COVID-19) in the world using time series models. *Chaos Solitons Fractals* **2020**, *140*, 110151. [CrossRef]
7. Zeroual, A.; Harrou, F.; Dairi, A.; Sun, Y. Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. *Chaos Solitons Fractals* **2020**, *140*, 110121. [CrossRef] [PubMed]
8. Qi, H.; Xiao, S.; Shi, R.; Ward, M.P.; Chen, Y.; Tu, W.; Su, Q.; Wang, W.; Wang, X.; Zhang, Z. COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis. *Sci. Total Environ.* **2020**, *728*, 138778. [CrossRef]
9. Briz-Redon, A. The impact of modelling choices on modelling outcomes: A spatio-temporal study of the association between COVID-19 spread and environmental conditions in Catalonia (Spain). *Stoch. Environ. Res. Risk Assess.* **2021**. [CrossRef]
10. Zanin, M.; Papo, D. Assessing functional propagation patterns in COVID-19. *Chaos Solitons Fractals* **2020**, *138*, 109993. [CrossRef]
11. Pak, D.; Langohr, K.; Ning, J.; Cortés-Martínez, J.; Gómez-Melis, G.; Shen, Y. Modeling the Coronavirus Disease 2019 Incubation Period: Impact on Quarantine Policy. *Mathematics* **2020**, *8*, 1631. [CrossRef]
12. Mansour, M.; Farsi, M.; Mohamed, S.; Elrazik, M. Modeling the COVID-19 Pandemic Dynamics in Egypt and Saudi Arabia. *Mathematics* **2021**, *9*, 827. [CrossRef]
13. Ramsay, J.O.; Silverman, B.W. *Functional Data Analysis*, 2nd ed.; Springer: New York, NY, USA, 2005.
14. Ramsay, J.O.; Silverman, B.W. *Applied Functional Data Analysis: Methods and Case Studies*; Springer: New York, NY, USA, 2002.
15. Ramsay, J.O.; Hooker, G.; Graves, S. *Functional Data Analysis with R and MATLAB*; Springer: New York, NY, USA, 2009.
16. Ferraty, F.; Vieu, P. *Nonparametric Functional Data Analysis. Theory and Practice*; Springer: New York, NY, USA, 2006.
17. Horvath, L.; Kokoszka, P. *Inference for Functional Data with Applications*; Springer: New York, NY, USA, 2012.
18. Tang, C.; Wang, T.; Zhang, P. Functional data analysis: An application to COVID-19 data in the United States. *arXiv* **2020**, arXiv:2009.08363.
19. Acal, C.; Aguilera, A.M.; Escabias, M. New Modeling Approaches Based on Varimax Rotation of Functional Principal Components. *Mathematics* **2020**, *8*, 2085. [CrossRef]
20. Carroll, C.; Bhattacharjee, S.; Chen, Y.; Dubey, P.; Fan, J.; Gajardo, A.; Zhou, X.; Müller, H.G.; Wang, J.L. Time dynamics of COVID-19. *Sci. Rep.* **2020**, *10*, 21040. [CrossRef] [PubMed]
21. Torres-Signes, A.; Frías, M.P.; Ruiz-Medina, M.D. COVID-19 mortality analysis from soft-data multivariate curve regression and machine learning. *arXiv* **2021**, arXiv:2008.06344.
22. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2019.
23. Graham, J.W. *Missing Data: Analysis and Design*; Springer Science & Business Media: New York, NY, USA, 2012.
24. He, Y.; Yucel, R.; Raghunathan, T.E. A functional multiple imputation approach to incomplete longitudinal data. *Stat. Med.* **2011**, *30*, 1137–1156. [CrossRef] [PubMed]
25. Ferraty, F.; Sued, M.; Vieu, P. Mean estimation with data missing at random for functional covariables. *Statistics* **2013**, *47*, 688–706. [CrossRef]
26. Ling, N.; Liang, L.; Vieu, P. Nonparametric regression estimation for functional stationary ergodic data with missing at random. *J. Stat. Plan. Inference* **2015**, *162*, 75–87. [CrossRef]
27. Ling, N.; Liu, Y.; Vieu, P. Conditional mode estimation for functional stationary ergodic data with responses missing at random. *Statistics* **2016**, *50*, 991–1013. [CrossRef]
28. Crambes, C.; Henchiri, Y. Regression imputation in the functional linear model with missing values in the response. *J. Stat. Plan. Inference* **2019**, *201*, 103–119. [CrossRef]
29. Febrero-Bande, M.; Galeano, P.; González-Manteiga, W. Estimation, imputation and prediction for the functional linear model with scalar response with responses missing at random. *Comput. Stat. Data Anal.* **2019**, *131*, 91–103. [CrossRef]
30. Ciarleglio, A.; Petkova, E.; Harel, O. Multiple imputation in functional regression with applications to EEG data in a depression study. *arXiv* **2020**, arXiv:2001.08175.
31. Rao, A.R.; Reimherr, M. Modern multiple imputation with functional data. *Stat* **2021**, *10*, e331. [CrossRef]
32. Aguilera, A.M.; Escabias, M.; Ocaña, F.A.; Valderrama, M.J. Functional Wavelet-Based Modelling of Dependence Between Lupus and Stress. *Methodol. Comput. Appl. Probab.* **2015**, *17*, 1015–1028. [CrossRef]
33. Valderrama, M.; Ocaña, F.; Aguilera, A.; Ocaña-Peinado, F. Forecasting pollen concentration by a two-step functional model. *Biometrics* **2010**, *66*, 578–585. [CrossRef]
34. Qi, X.; Luo, R. Function-on-function regression with thousands of predictive curves. *J. Multivar. Anal.* **2018**, *163*, 51–66. [CrossRef]
35. Lima, I.; Cao, G.; Billor, N. Robust simultaneous inference for the mean function of functional data. *Test* **2019**, *28*, 785–803. [CrossRef]
36. Chiou, J.M.; Müller, H.G.; Wang, J.L. Functional response models. *Stat. Sin.* **2004**, *14*, 659–677.
37. Escabias, M.; Aguilera, A.M.; Valderrama, M.J. Principal component estimation of functional logistic regression: Discussion of two different approaches. *J. Nonparametr. Stat.* **2004**, *16*, 365–384. [CrossRef]
38. Müller, H.G.; Stadtmüller, U. Generalized functional linear models. *Ann. Stat.* **2005**, *33*, 774–805. [CrossRef]
39. Aguilera-Morillo, M.C.; Aguilera, A.M.; Escabias, M.; Valderrama, M.J. Penalized spline approaches for functional logit regression. *Test* **2013**, *22*, 251–277. [CrossRef]

40. Escabias, M.; Aguilera, A.; Aguilera-Morillo, M. Functional PCA and Base-Line Logit Models. *J. Classif.* **2014**, *31*, 296–324. [CrossRef]
41. Aguilera, A.M.; Aguilera-Morillo, M.C.; Preda, C. Penalized versions of functional PLS regression. *Chemom. Intell. Lab. Syst.* **2016**, *154*, 80–92. [CrossRef]
42. Preda, C.; Saporta, G. PLS regression on a stochastic process. *Comput. Stat. Data Anal.* **2005**, *48*, 149–158. [CrossRef]
43. Escabias, M.; Aguilera, A.M.; Valderrama, M.J. Functional PLS logit regression model. *Comput. Stat. Data Anal.* **2007**, *51*, 4891–4902. [CrossRef]
44. Aguilera, A.M.; Escabias, M.; Preda, C.; Saporta, G. Using basis expansion for estimating functional PLS regression. Applications with chemometric data. *Chemom. Intell. Lab. Syst.* **2010**, *104*, 289–305. [CrossRef]
45. Delaigle, A.; Hall, P. Methodology and theory for partial least squares applied to functional data. *Ann. Stat.* **2012**, *40*, 322–352. [CrossRef]
46. Febrero-Bande, M.; Galeano, P.; González-Manteiga, W. Functional principal component regression and functional partial least squares regression: An overview and a comparative study. *Int. Stat. Rev.* **2017**, *85*, 61–83. [CrossRef]
47. Aguilera, A.M.; Acal, C.; Aguilera-Morillo, M.C.; Jiménez-Molinos, F.; Roldán, J.B. Homogeneity problem for basis expansion of functional data with applications to resistive memories. *Math. Comput. Simul.* **2021**, *186*, 41–51. [CrossRef]
48. Aguilera, A.M.; Ocaña, F.A.; Valderrama, M.J. An approximated principal component prediction model for continuous-time stochastic processes. *Appl. Stoch. Model. Data Anal.* **1997**, *13*, 61–72. [CrossRef]
49. Aguilera, A.M.; Ocaña, F.A.; Valderrama, M.J. Forecasting with unequally spaced data by a functional principal component approach. *Test* **1999**, *8*, 233–254. [CrossRef]
50. Deville, J.C. Méthodes statistiques et numériques de l'analyse harmonique. *Ann. De L'INSEE* **1974**, *15*, 3–101. [CrossRef]
51. Dauxois, J.; Pousse, A.; Romain, Y. Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *J. Multivar. Anal.* **1982**, *12*, 136–156. [CrossRef]
52. Ocaña, F.A.; Aguilera, A.M.; Valderrama, M.J. Functional Principal Components Analysis by Choice of Norm. *J. Multivar. Anal.* **1999**, *71*, 262–276. [CrossRef]
53. Hall, P.; Hosseini-Nasab, M. On properties of functional principal components analysis. *J. R. Stat. Soc. B* **2006**, *68*, 109–126. [CrossRef]
54. Ruiz-Castro, J.E.; Acal, C.; Aguilera, A.M.; Aguilera-Morillo, M.C.; Roldán, J.B. Linear-Phase-Type probability modelling of functional PCA with applications to resistive memories. *Math. Comput. Simul.* **2021**, *186*, 71–79. [CrossRef]
55. Ocaña, F.A.; Aguilera, A.M.; Escabias, M. Computational considerations in functional principal component analysis. *Comput. Stat.* **2007**, *22*, 449–465. [CrossRef]
56. Nie, Y.; Wang, L.; Liu, B.; Cao, J. Supervised functional principal component analysis. *Stat. Comput.* **2018**, *28*, 713–723. [CrossRef]