Revista Colombiana de Estadística - Applied Statistics
January 2025, volume 48, issue 1, pp. 177–193
http://doi.org/10.15446/rce.v48n1.113398

Modeling Experimental Designs Including Longitudinal Data and a Functional Covariate

Modelación de diseños experimentales incluyendo datos longitudinales y una covariable funcional

Gustavo Adolfo Gómez-Escobar[1,a], Mercedes Andrade-Bejarano[1,b], Ramón Giraldo[2,c]

[1]Escuela de Estadística, Facultad de Ingeniería, Universidad del Valle, Santiago de Cali, Colombia

[2]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

The study of longitudinal measures of chlorophyll concentrations is key to reducing the risk of yield-limiting deficiencies or costly over fertilizing. Factors as irrigation and fertilization can influence the chlorophyll content. In this research we analyzed data from a experimental design of chlorophyll concentrations in chili pepper plants under the effect of two factors (fertilizer and irrigation, both with four levels) recorded weekly (for seven weeks). The spectral signature curves obtained for each plant was included in the model as a functional covariate. We propose an alternative for the analysis of data from experimental designs involving longitudinal data (LD) and a functional covariate. Two smoothing approaches using basis functions and functional principal component reduce the problem to the application of a Linear Mixed Model (LMM) to LD in the presence of multiple scalar covariates. In both approaches, the results indicate that the inclusion of the functional covariate (spectral signature) contributes to explain the relationship between the chlorophyll concentration and the factors analyzed.

Key words: Basis functions; Chlorophyll concentration; Functional data analysis; Functional principal components analysis; Random coefficient model; Spectral signature.

[a]MSc. E-mail: gustavoges90@gmail.com
[b]Ph.D. E-mail: mercedes.andrade@correounivalle.edu.co
[c]Ph.D. E-mail: rgiraldoh@unal.edu.co

Resumen

El estudio de mediciones longitudinales de concentraciones de clorofila es clave para reducir el riesgo de deficiencias que limiten el crecimiento o de una fertilización excesiva y costosa. Factores como la irrigación y la fertilización pueden influir en el contenido de clorofila. En esta investigación analizamos datos de un diseño experimental de concentraciones de clorofila en plantas de ají picante, bajo el efecto de dos factores (fertilizante e irrigación, ambos con cuatro niveles), registrados semanalmente (durante siete semanas). Las curvas de firma espectral obtenidas por cada planta se incluyeron en el modelo como una covariable funcional. Proponemos una alternativa para el análisis de datos de diseños experimentales que involucran datos longitudinales (DL) y una covariable funcional. Dos enfoques de suavización que utilizan funciones base y componentes principales funcionales reducen el problema a la aplicación de un modelo lineal mixto (MLM) a DL en presencia de múltiples covariables escalares. En ambas alternativas, los resultados indican que la inclusión de una covariables funcional (firma espectral) contribuye a explicar la relación entre la concentración de clorofila y los factores analizados.

***Palabras clave*:** Análisis de componentes principales funcionales; Análisis de datos funcionales; Base de funciones; Concentración de clorofila; Firma espectral; Modelo de coeficientes aleatorios.

1. Introduction

Statistical analysis of longitudinal data is widely used in agronomy (Fenzi et al., 2017; Bonamy et al., 2020; Lark et al., 2020). In particular, the study of longitudinal measures of chlorophyll concentrations is key to reducing the risk of yield-limiting deficiencies or costly over fertilizing (Ling et al., 2011). Many factors (for example, irrigation and fertilization) can influence the chlorophyll content. Thus, analyzing data generated by experimental designs with a longitudinal response is an essential statistical function in this area. In this paper, we analyze a dataset where the chlorophyll concentration in response to four levels of irrigation and fertilization is assessed (Figure 1). We also considered the spectral curve obtained weekly in each plant as the realization of a functional covariate (Gómez-Escobar, 2017). The spectral signatures were measured with wavelengths between 500 and 950 (nm). The data were collected over seven weeks in an experiment that was carried out at the agronomic experimental center of Universidad del Valle, Cali, Colombia.

Longitudinal data (LD) occur when we repeatedly measure the same variable across time on the study subjects (Weiss, 2005). These records are not necessarily obtained at the same timepoints or the same number of times (unbalanced data). Many approaches are available for the analysis of LD. Repeated measures ANOVA (Davis, 2002), generalized estimating equations (GEE) (Hardin & Hilbe, 2002), multivariate ANOVA (Hedeker & Gibbons, 2006), and linear mixed models (LMM) (Fitzmaurice et al., 2008) are some alternatives. Currently, LMM is probably the most common method for analyzing LD (Fitzmaurice et al., 2008). One advantage of using LMM is that it is possible to perform hypothesis test-

ing on correlation parameters. Classical methods such as the likelihood ratio test (Crainiceanu et al., 2004) and Bayesian information criterion (Jones, 2011) can be used to test and compare model fit. LD is usually recorded under several experimental conditions (combinations of the levels of factors), and generally, some covariates are also considered. There are different types of covariates: continuous, discrete, multinomial, and ordered multinomial (Weiss, 2005).

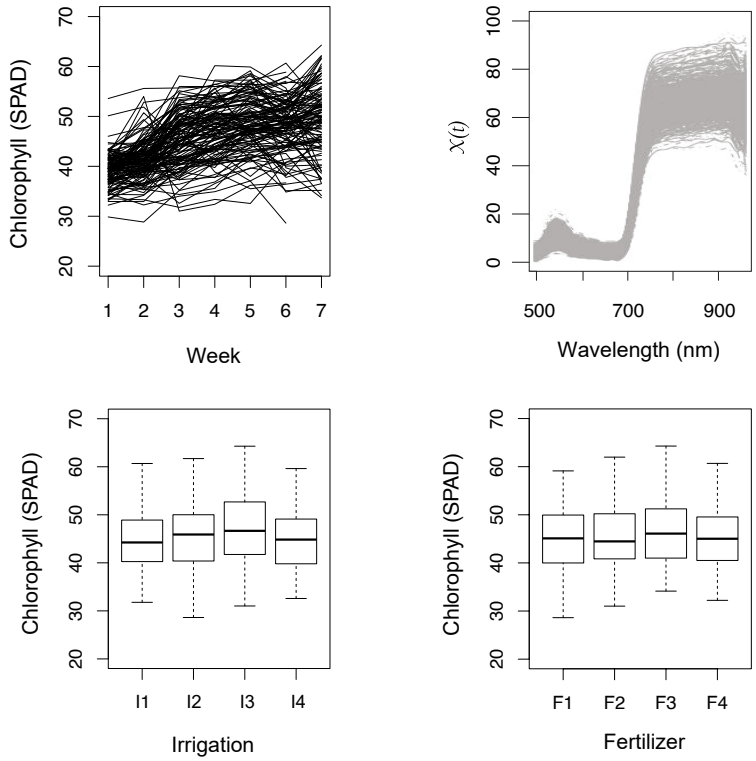

FIGURE 1: Above: Chlorophyll weekly data (left) and spectral signatures curves (right). Bottom: Boxplots of chlorophyll data for each level of irrigation (left) and fertilizer (right).

Since the early 1990s, functional data analysis (FDA) has been a highly developed field in statistics (Ramsay & Silverman, 2005). In FDA, initially, many records of some characteristic of interest for each individual in the sample (usually obtained over time) are fit as a curve (function) using smoothing techniques (nonparametric smoothing, basis functions, polynomial regression, etc.). Subsequently, the curves become the objects of study (Ramsay & Silverman, 2005). FDA encompasses the set of statistical methodologies that allow descriptive and inferential analyses to be carried out with functional data. Regression, ANOVA, mixed models, or multivariate methods (principal components, cluster, and discriminant analysis), among others, have been proposed for this class of variables (Ramsay & Silverman, 2005; Ferraty & Vieu, 2006; Febrero-Bande et al., 2010; Horváth & Kokoszka, 2012). FDA has also been used to model LD when a large set of records

is obtained for each individual in the sample. The traditional LMM has been adapted to this framework generating the so-called functional linear mixed model (FLMM). Many relatively recent works with theoretical and applied perspectives consider this problem (Guo, 2002, 2004; Liu & Guo, 2012; Park & Staicu, 2015; Cederbaum et al., 2016; Liu et al., 2017). Kundu et al. (2016) propose estimation in regression models for longitudinally collected functional covariates; their methodology extends the analysis of functional linear models by relating a scalar outcome to a functional predictor both observed longitudinally; this approach may be viewed as an extension of longitudinal mixed effects models, replacing scalar predictors by functional predictors. Staicu et al. (2020) extend the work by Kundu et al. (2016). They develop a longitudinal dynamic functional regression (LDFR) framework to study time-varying association between responses from the exponential family and functional covariates that are observed in a longitudinal design that enables to recover the full response trajectory.

In this paper, we propose two alternatives (Sections 2.2 and 2.3) for modeling the dataset described in Section 1. A detailed presentation of the data is given in Section 3. The methodologies can be also used in the analysis of other datasets from experimental designs including LD and a functional covariate. We consider data from a longitudinal response and a functional covariate recorded under the levels of various factors. A solution based on functional principal components and the use of confidence bands generated by bootstrap are methodological contributions of this work. Specifically the confidence bands are obtained using a combination of parametric bootstrap in mixed models and bootstrap methods for functional data (Febrero-Bande et al., 2010; Febrero-Bande & Oviedo de la Fuente, 2012). After using basis functions (B-splines, Fourier, etc.) or functional principal components analysis to smooth the data, the problem becomes that of a classical experimental design of LD with multiple scalar covariates. Experimental designs involving LD play an essential role in agricultural sciences. These are routinely used in this field to study changes over time (see for example Fenzi et al. 2017 and Miqueloni, 2019). The proposed methodology is of interest in this context.

The article is organized as follows: Sections 2 and 3 present the methodology proposed and the application to the dataset of interest, respectively. The paper ends with a discussion and suggestions for further research.

2. Statistical Methods

In this section, we begin by defining the extension of the classical linear mixed model to include simultaneously longitudinal data and a functional covariate. Next, we introduce two alternative methods for parameter estimation in this framework.

2.1. Mixed Model Including Longitudinal Data and a Functional Covariate

A classical LMM (Verbeke & Molenberghs, 2000) with a longitudinal variable is defined as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $\mathbf{Y}_i$ is the n_i dimensional response vector for subject i, $i = 1, \ldots, N$, N the number of subjects, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are, respectively $(n_i \times p)$ and $(n_i \times q)$ matrices for the fixed and random effects, $\boldsymbol{\beta}$ is a p-dimensional vector of the fixed effects, $\mathbf{b}_i$ is a q-dimensional vector of the random effects, with $\mathbf{b_i} \sim \mathcal{N}(\mathbf{0},D)$, and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_i)$. It is assumed that $\mathbf{b}_1, \ldots, \mathbf{b}_N$ and $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N$ are independent. The extension of Model (1) to the case where a functional covariate is considered can be defined as follows:

$$\mathbf{Y}_i = \int_T \chi_i(t)\psi(t)dt + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \tag{2}$$

with $\chi_i(t), t \in T$ and $i = 1, \ldots, N$, a functional variable and $\psi(t)$ a functional parameter.

2.2. Estimation Based on Basis Functions

In practice some basis functions (B-splines, Wavelets, Fourier) can be used to define curves based on a large set of data $(\chi_i(t_1), \ldots, \chi_i(t_m))$. Using a basis function approach we have for each subject $i = 1, \ldots, N$ in Model (2)

$$\begin{aligned}\chi_i(t) &= \sum_{j=1}^{k} c_{ij}\phi_j(t) \\ &= c_{i1}\phi_1(t) + \ldots + c_{ik}\phi_k(t) \\ &= \mathbf{c}_i^T\boldsymbol{\phi}, \text{ with } \mathbf{c}_i^T = (c_{i1}, \ldots, c_{ik}) \text{ and } \boldsymbol{\phi} = (\phi_1(t), \ldots, \phi_k(t))^T\end{aligned} \tag{3}$$

The optimal number of basis functions k is generally derived by generalized cross validation (GCV) Ramsay & Silverman (2005). Similarly the functional parameter in Model (2) can be defined as

$$\begin{aligned}\psi(t) &= \sum_{j=1}^{k} d_j\theta_j(t) \\ &= d_1\theta_1(t) + \cdots + d_k\theta_k(t) \\ &= \boldsymbol{\theta}^T\mathbf{d}, \text{ with } \boldsymbol{d} = (d_1, \ldots, d_k)^T \text{ and } \boldsymbol{\theta}^T = (\theta_1(t), \ldots, \theta_k(t))\end{aligned} \tag{4}$$

For simplicity, the same k is usually considered in both representations. Based on Equations (3) and (4) the model in (2) can be written as follows

$$\begin{aligned}\mathbf{Y}_i &= \int_T \mathbf{c}_i^T\boldsymbol{\phi}\boldsymbol{\theta}^T\mathbf{d}dt + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \\ &= \mathbf{c}_i^T\mathbf{J}\mathbf{d} + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \; i = 1, \ldots, N\end{aligned} \tag{5}$$

with

$$\mathbf{J} = \begin{bmatrix} \int_T \phi_1(t)\theta_1(t)dt & \cdots & \int_T \phi_1(t)\theta_k(t)dt \\ \vdots & \ddots & \vdots \\ \int_T \phi_k(t)\theta_1(t)dt & \cdots & \int_T \phi_k(t)\theta_k(t)dt \end{bmatrix}$$

The model in Equation (5) for each subject is given by

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{bmatrix} = \begin{bmatrix} \mathbf{c}_{i1}^T\mathbf{J} & x_{i11} & \cdots & x_{i1p} \\ \mathbf{c}_{i2}^T\mathbf{J} & x_{i21} & \cdots & x_{i2p} \\ \vdots & \vdots & & \\ \mathbf{c}_{in_i}^T\mathbf{J} & x_{in_i1} & \cdots & x_{in_ip} \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_k \\ \boldsymbol{\beta} \end{bmatrix} + \begin{bmatrix} z_{i11} & \cdots & z_{i1q} \\ z_{i21} & \cdots & z_{i2q} \\ \vdots & \ddots & \vdots \\ z_{in_i1} & \cdots & z_{in_iq} \end{bmatrix} \begin{bmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{iq} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{in_i} \end{bmatrix}$$

If the basis functions used to define functional data and the functional parameter are orthonormal we have

$$\int_T \phi_l(t)\theta_m(t)dt = \begin{cases} 1 & \text{if } l = m \\ 0 & \text{if } l \neq m \end{cases}, \; l, m = 1, \ldots, k$$

i.e, in (5) $\mathbf{J} = \mathbf{I}_{k\times k}$. Then, the Model (2) becomes

$$\mathbf{Y_i} = \widetilde{\mathbf{X}}_i\widetilde{\boldsymbol{\beta}} + \mathbf{Z}_i\mathbf{b_i} + \boldsymbol{\varepsilon_i}, \tag{6}$$

where $\widetilde{\mathbf{X}}_i = \left[\mathbf{c}_i^T, \mathbf{X}_i\right]$. After calculating $\widetilde{\mathbf{X}}_i$, the estimations of the model can be obtained through maximum likelihood or restricted maximum likelihood (REML) (Patterson & Thompson, 1971). Let $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \boldsymbol{\Sigma}_i$. This matrix can be estimated maximizing the REML likelihood function (Verbeke & Molenberghs, 2000)

$$l_{REML}(\boldsymbol{\gamma}) = \left|\sum_{i=1}^N \widetilde{\mathbf{X}}_i^T\mathbf{V}_i^{-1}\widetilde{\mathbf{X}}_i\right|^{-1/2} \prod_{i=1}^N (2\pi)^{-n_i/2}\left|\mathbf{V}_i(\boldsymbol{\alpha}')\right|^{-1/2} \exp\left(\frac{-1}{2}(\mathbf{Y_i} - \widetilde{\mathbf{X}}_i\boldsymbol{\beta})^T\mathbf{V}_i^{-1}(\alpha)(\mathbf{Y_i} - \widetilde{\mathbf{X}}_i\boldsymbol{\beta})\right)$$

with respect to all parameters simultaneously $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$; where $\boldsymbol{\alpha}$ is an unknown vector of variance components. The fixed effects are estimated by Verbeke & Molenberghs (2000)

$$\widehat{\widetilde{\boldsymbol{\beta}}}(\boldsymbol{\alpha}) = \left(\sum_{i=1}^N \widetilde{\mathbf{X}}_i^T\mathbf{W}_i\widetilde{\mathbf{X}}_i)\right)^{-1} \sum_{i=1}^N \widetilde{\mathbf{X}}_i^T\mathbf{W}_i\mathbf{y_i},$$

where $\mathbf{W}_i = \mathbf{V}_i^{-1}$. When an estimate $\widehat{\boldsymbol{\alpha}}$ is available $\widehat{\mathbf{V}_i} = \widehat{\mathbf{V}_i}(\widehat{\alpha}) = \widehat{\mathbf{W}_i^{-1}}$. $\mathbf{b_i}$ in Equation (6) can be estimated by Verbeke & Molenberghs (2000)

$$\mathbf{b_i}(\boldsymbol{\gamma}) = \mathbf{D}\mathbf{Z}_i^T\mathbf{W}_i(\mathbf{y_i} - \widetilde{\mathbf{X}}_i\boldsymbol{\beta}).$$

The number of basis functions k that defines the functional parameter is determined through the marginal Akaike information criterion (mAIC) (Greven & Kneib, 2010) as

$$\text{mAIC} = -2l\left(\mathbf{Y} \mid \widehat{\widetilde{\boldsymbol{\beta}}}, \widehat{\mathbf{V}}\right) + 2(p+q),$$

where $\widehat{\mathbf{V}}$ is the covariance matrix given by $\widehat{\mathbf{V}} = \mathbf{Z}\widehat{\mathbf{D}}\mathbf{Z}^T + \widehat{\boldsymbol{\Sigma}}$, $l\left(\mathbf{Y} \mid \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{V}}\right)$ is the maximized log likelihood and p, q are then number of parameters in the model.

2.3. Estimation Based on Functional Principal Components Analysis (FPCA)

Consider the model in Equation (2). Suppose that based on the set of curves $\chi_1(t), \ldots, \chi_k(t)$ a functional principal components analysis is carried out obtaining the eigenfunctions $\xi_1(t), \ldots, \xi_k(t)$. The functional data and functional parameters in Model (2) are defined in terms of the eigenfunctions as

$$\chi_i(t) = \sum_{j=1}^{k} \alpha_{ij}\xi_j(t), \tag{7}$$

and

$$\psi(t) = \sum_{j=1}^{k} \tau_j\xi_j(t), \tag{8}$$

Using these representations (7) and (8) the model in (2) can be written as

$$\begin{aligned}
\mathbf{Y}_i &= \int_T \chi_i(t)\psi(t)dt + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \qquad (9)\\
\mathbf{Y}_i &= \int_T \mathbf{A}_i^T\boldsymbol{\xi}(t)\boldsymbol{\xi}^t(t)\boldsymbol{\tau}dt + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i\\
\mathbf{Y}_i &= \mathbf{A}_i^T\left(\int_T \boldsymbol{\xi}(t)\boldsymbol{\xi}^t(t)dt\right)\boldsymbol{\tau} + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i,
\end{aligned}$$

where $\mathbf{A}_i^T = (\alpha_{ij})$ is the j-th score of the i-th functional datum. Considering that the eigenfunctions belong to an orthonormal base, we have $\int_T \boldsymbol{\xi}(t)\boldsymbol{\xi}^t(t)dt = I_{k\times k}$, and the model can be written as

$$\mathbf{Y}_i = \mathbf{A}_i^T\boldsymbol{\tau} + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \tag{10}$$

The Model (10) can be defined as

$$\mathbf{Y}_i = \widetilde{\mathbf{X}}_i\widetilde{\boldsymbol{\beta}} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \tag{11}$$

with $\widetilde{\mathbf{X}}_i = (\mathbf{A}_i^T, \mathbf{X}_i)$ and $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{\tau}, \boldsymbol{\beta})$. The estimation of the parameters in Model (11) is carried out as in Section 2.2.

2.4. Verifying Model Assumptions

The validation of the assumptions is carried out graphically. The normality is evaluated using a QQ plot of the standardized residuals and the homogeneity of variance trough a predicted versus standardized residuals plot. The residuals are defined as

$$\mathbf{r}_i^* = \mathbf{L}^{-1}(\mathbf{y}_i - \widetilde{\mathbf{X}}_i\widehat{\boldsymbol{\beta}})$$

where $\widehat{\boldsymbol{\Sigma}} = \mathbf{L}\mathbf{L}^T$ is the Cholesky decomposition of the estimated errors covariance matrix.

2.5. Confidence Bands Approach via Bootstrap

To obtain the confidence bands for the functional parameter $\psi(t)$ in Equation (2), we perform a combination of fully parametric bootstrap methods in mixed and functional models (Lahiri, 2003; Febrero-Bande et al., 2010; Febrero-Bande & Oviedo de la Fuente, 2012). For the model in (9), the confidence area (CA) is defined as

$$CA(\psi(t)) = \left\{\psi(t) \in L^2 : \left\|\psi(t) - \widehat{\psi}(t) \leqslant D_\alpha\right\|\right\},$$

where the statistic $D\alpha$, meets the condition of

$$P\left(\psi(t)\right) \in CA(\psi(t)) = 1 - \alpha$$
$$P\left(\left\|\psi(t) - \widehat{\psi}(t)\right\|\right) = 1 - \alpha$$

The procedure to obtain the confidence bands is as follows:

- Step 1: Fit Model (9) with the original data to obtain the respective estimates.
- Step 2: Generate $\{\mathbf{b}_i\}_{i=1}^N$ of size N of a q-dimensional normal distribution with means vector $\mathbf{0}$ and covariance matrix $\widehat{\mathbf{D}}$ (estimated in step 1).
- Step 3: Generate N samples for $\{\boldsymbol{\varepsilon}_i\}_i^N$ of size n_i of a multivariate normal distribution with means vector $\mathbf{0}$ and covariances matrix $\widehat{\boldsymbol{\Sigma}}$ estimated in step 1.
- Step 4: Obtain the bootstrap observation of the response variable as

$$\mathbf{y}_i^* = \int_T \boldsymbol{\chi}(t)\psi(t)dt + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i^* + \boldsymbol{\varepsilon}_i^*$$

- Step 5: Fit Model (9) taking the bootstrap observations of the response variable $\mathbf{y}_i^*$, and obtain the estimates of the functional parameter denoted as $(\psi(t))^j$.
- Step 6: Replicate steps 1 to 5 B times.

- Step 7: The estimated value $\widehat{D_\alpha}$ is obtained through the percentile $(1-\alpha)$ of

$$\text{diff}^j = \left\| \widehat{\psi}(t) - \left(\widehat{\psi}(t)\right)^j \right\| = \int_T \left(\widehat{\psi}(t) - \left(\widehat{\psi}(t)\right)^j \right) dt$$

- Step 8: Plot the estimated values $\text{diff}^j \leq \widehat{D_\alpha}$.

3. Application to Chlorophyll Data

Table 1 contains information on the levels of fertilizer and irrigation used in the experiment described in Section 1.

TABLE 1: Levels of the factors fertilizer and irrigation used in the experiment.

Factors	Levels	
Fertilizer	F1	Boron Solution
	F2	Solution without Iron
	F3	Solution without Magnesium
	F4	Complete Solution
Irrigation	I1	225(ml)
	I2	15(ml)
	I3	75(ml)
	I4	150(ml)

3.1. Model

For modeling the chlorophyll concentration in SPAD measurements, using the spectral signature of the plants as a covariate, we specified the first level of the model as

$$\text{Chlorophyll}_i = \beta_{0i} + \beta_{1.} + \beta_{2.} + \beta_{3i}t + \beta_{4.}t + \beta_{5.}t + \beta_{6.}t + \int_T \psi(\lambda)\boldsymbol{\chi}(\lambda)dt + e_i \quad (12)$$

where the indexing of the parameter and the functional variable is denoted as λ, $\lambda \in [500 - 1100nm]$, to differentiate it from time (t). In the second level of the model, we study the change ln the intercept of the subject-specific and the variation of their respective time slopes. The follow equations are specified:

$$\begin{aligned}
\beta_{0i} &= \beta_0 + b_{0i} \\
\beta_{1.} &= \beta_1 F_1 + \beta_2 F_2 + \beta_3 F_3 + \beta_4 I_1 + \beta_5 I_2 + \beta_6 I_3 \\
\beta_{2.} &= \beta_7 F_1 I_1 + \beta_8 F_1 I_2 + \beta_9 F_1 I_3 + \beta_{10} F_2 I_1 + \beta_{11} F_2 I_2 + \beta_{12} F_2 I_3 \\
&\quad + \beta_{13} F_3 I_1 + \beta_{14} F_3 I_2 + \beta_{15} F_3 I_3 \\
\beta_{3i} &= \beta_{16} + b_{1i} \\
\beta_{4.} &= \beta_{17} F_1 + \beta_{18} F_2 + \beta_{19} F_3 \\
\beta_{5.} &= \beta_{20} I_1 + \beta_{21} I_2 + \beta_{22} I_3 \\
\beta_{6.} &= \beta_{23} F_1 I_1 + \beta_{24} F_1 I_2 + \beta_{25} F_1 I_3 + \beta_{26} F_2 I_1 + \beta_{27} F_2 I_2 + \beta_{28} F_2 I_3 \\
&\quad + \beta_{29} F_3 I_1 + \beta_{30} F_3 I_2 + \beta_{31} F_3 I_3
\end{aligned}$$

where F_1, F_2, and F_3 and I_1, I_2, and I_3 are indicator variables, that express the levels of the fertilizer and irrigation factors

$$\begin{cases} Fertilizer\ 1 & if \quad F_1 = F_2 = F_3 = 0 \\ Fertilizer\ 2 & if \quad F_1 = 1 \text{ and } F_2 = F_3 = 0 \\ Fertilizer\ 3 & if \quad F_2 = 1 \text{ and } F_1 = F_3 = 0 \\ Fertilizer\ 4 & if \quad F_3 = 1 \text{ and } F_1 = F_2 = 0 \end{cases}$$

$$\begin{cases} Irrigation\ 1 & if \quad I_1 = I_2 = I_3 = 0 \\ Irrigation\ 2 & if \quad I_1 = 1 \text{ and } I_2 = I_3 = 0 \\ Irrigation\ 3 & if \quad I_2 = 1 \text{ and } I_1 = I_3 = 0 \\ Irrigation\ 4 & if \quad I_3 = 1 \text{ and } I_1 = I_2 = 0 \end{cases}$$

- $\boldsymbol{y}_i$ is the vector that contains the seven (7) measurements of clorophyll for individual i, $i = 1, 2, \ldots, 128$.
- β_0 represents the fixed intercept parameter.
- β_1, β_2 and β_3 correspond to the regression parameters for the variables associated with fertilizer levels 2, 3 and 4.
- β_4, β_5 and β_6 are the regression parameters for the variables associated with irrigation levels 2, 3, and 4.
- β_7, $\beta_8, \ldots, \beta_{15}$ are the regression parameters for the interactions doubles between the variables F_1, F_2, F_3, I_1, I_2 y I_3.
- β_{17}, β_{18}, and β_{19} correspond to the regression parameters for the effects associated with the interactions between the variables F_1, F_2, and, F_3, and time.
- β_{20}, β_{21}, β_{22} are the regression parameters for the interactions of the variables I_1, I_2, I_3, and time.
- β_{23}, $\beta_{24}, \ldots, \beta_{31}$ correspond to the regression parameters for the triple interactions among the variables F_1, F_2, F_3, I_1, I_2, I_3, and time
- β_{16} is the regression parameter associated with the time slope.
- b_{0i} is the random effect corresponding to the intercept for the i-th individual
- b_{1i} corresponds to the random effect of the time slope.
- b_{0i} and b_{1i} are assumed to follow the distribution $\mathcal{N}(\mathbf{0}, \mathbf{D})$. We assume the variance-covariance matrix $\mathbf{D}$ is unstructured (Verbeke & Molenberghs, 2000).
- $\boldsymbol{\varepsilon}_i$ is the random error vector for individual i, which is assumed to follow the distribution $\mathcal{N}(0, \boldsymbol{\Sigma}_i)$.

3.2. Results

The chlorophyll concentration in SPAD measurements varied between 28.6 and 64.2 (top left, Figure 1). We observed an increasing trend over time (top left Figure 1) and that there were no marked differences among the levels of the factors (irrigation and fertilizer) with respect to the chlorophyll concentration (bottom left and right panels in Figure 1). With respect to the spectral signature curves we noted high variability after 750 nm (top right Figure 1). This may be due to a combination of factors related to the absorption and reflection of electromagnetic radiation in the environment, the sensitivity of the detectors, the optical properties of the materials, and the behavior of the radiation. It is important to consider these factors when interpreting spectral measurements in the near-infrared region. Next, we present the results obtained with the modeling approaches given in Sections 2.2 to 2.5.

3.2.1. Linear Mixed Model Based on Basis Functions

Consider the model defined in Section 3.1. Initially, we assumed that the covariance structure of the errors is $var(\epsilon_i) = \sigma^2$. Through the marginal Akaike information criterion (mAIC) (Greven & Kneib, 2010), we determined that $K_{\psi(\lambda)} = 10$ is the optimal number of B-spline basis functions to describe the functional parameter $\psi(\lambda)$ (Gómez-Escobar, 2017). Using AIC and Bayesian information criterion (BIC), we established that AR(1) was an appropriate covariance structure for the errors, and based on these indicators, we determined that a random coefficients model (including intercept and slope) was adequate for modeling the chlorophyll data. Similarly, the model including the functional covariate yielded low values of AIC and BIC (Table 2).

Table 2: Comparison of the model including and excluding the random slope and the model with the intercept and random slope with and without the functional covariate.

Model	df	AIC	BIC
Model with b_{0i}	46	4780.5	4993.3
Model with b_{0i},b_{1i}	47	4768.1	4990.4
Model with χ(t)	47	4768.1	4990.4
Model without χ(t)	37	4948.0	5124.0

Table 3: Test of the fixed effects including the functional covariate.

Source of variation	F	p-value
Intercept	44218.2	0.00
Fertilizer	1.2	0.30
Irrigation	4.8	0.00
Time	543.5	0.00
Fertilizer:Irrigation	0.8	0.56
Fertilizer:Time	2.7	0.04
Irrigation:Time	6.6	0.00
Fertilizer:Irrigation:Time	0.6	0.77

Table 3 shows the variance analysis for fixed effects of the design model. The results reveals that the interactions of irrigation, time and fertilization-time were statistically significant at $\alpha = 0.05$. The parameters estimations for the model (12) can be found on Gómez-Escobar (2017) (pages 43 and 44, https://hdl.handle.net/10893/13510). We used bootstrapping to obtain confidence bands for the functional parameter. Figure 2 shows the bootstrap functional parameters that meet $\left\|\widehat{\psi}(t) - \left(\widehat{\psi}(t)\right)^j\right\| \leq \widehat{D_\alpha}$, where $\leq \widehat{D_\alpha}$ is the percentile $(1-\alpha)$ of the statistics $\text{diff}^j = \left\|\widehat{\psi}(t) - \left(\widehat{\psi}(t)\right)^j\right\|$. The confidence bands indicate that the functional parameter was significantly different from zero.

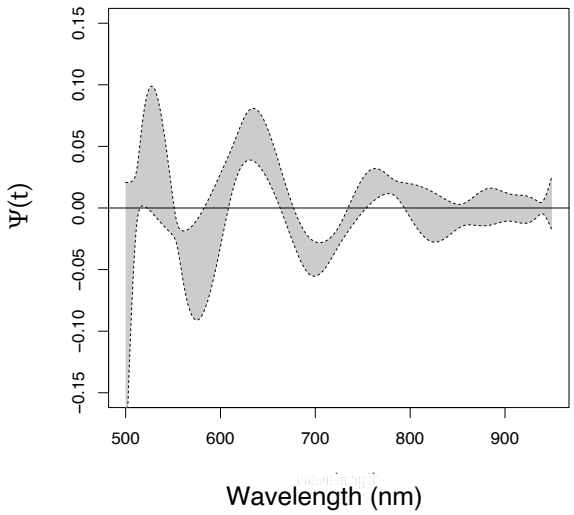

FIGURE 2: Confidence bands for the spectral signature (functional covariate). Case of the linear mixed model based on basis functions.

3.2.2. Mixed Model Through Functional Principal Components Analysis (FPCA)

The first four eigenfunctions $\xi_1(t), \ldots, \xi_4(t)$ (Figure 3) obtained from the centered functional data accounted for 99.3% of the variability. The first eigenfunction explained 91.4% of this variability and followed the mean behavior pattern of the spectral signatures (Figure 1). The remaining three (corresponding to 7.9% of the variability) were associated with curves with a behavior pattern different from the global one, particularly for wavelengths greater than 750 nm.

Three models were fitted (Table 4). The first one did not include the functional covariate $\chi(t)$. In the second case a functional covariate was considered and the corresponding curves were obtained by using a size 20 B-spline basis functions. In the third model, we also included a functional covariate, but in this case, the modeling was carried out through an FPCA, avoiding, on the one hand,

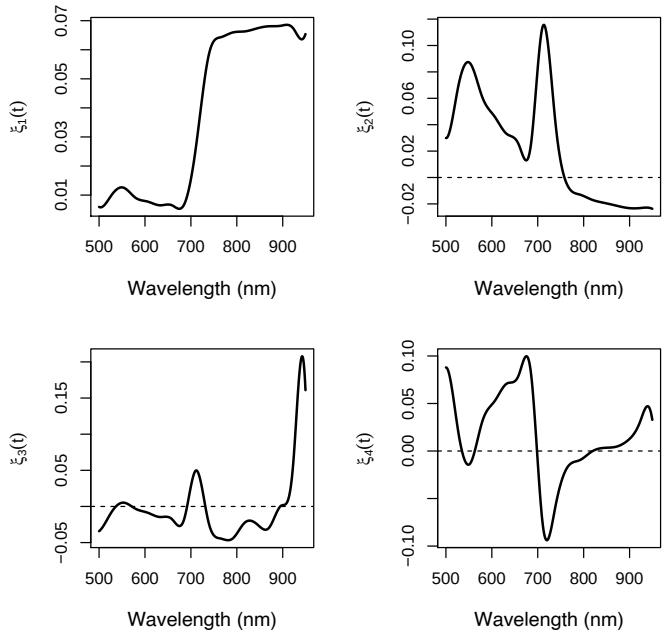

FIGURE 3: Four eigenfunctions obtained from a functional principal component analysis based on centered spectral signature curves. $\xi_1(t)$ explains 91.4% of the variability. The remaining ($\xi_2(t)$ to $\xi_4(t)$) accumulate 7.9%.

a possible collinearity problem, and reducing, on the other hand, the dimensionality of the model. The three models were compared using the AIC and BIC criteria (Table 4). Based on the results, we concluded that it was appropriate to include a functional covariate (spectral signature) to explain the relationship between the chlorophyll concentration and the factors fertilizer and irrigation. It was also clear that using an FPCA contributed to reducing the dimensionality of the problem and improving the goodness of fit. Additionally, the bootstrap confidence bands for the model based on FPCA (Gómez-Escobar, 2017, page 52, https://hdl.handle.net/10893/13510) indicated that the functional parameter was significantly different from zero (or from a constant value), meaning that the functional covariate must be considered in the model. From Table 5, we concluded that the interaction between irrigation and time was significant.

TABLE 4: Comparison among three models of longitudinal chlorophyll data including and excluding the spectral signature $\chi(t)$ as a functional covariate. DF: Degree freedom. AIC: Akaike information criterion. BIC: Bayesian information criterion.

Model	DF	AIC	BIC
1. Without χ(t)	37	4948.2	5123.6
2. With χ(t)	47	4768.1	4990.4
3. With χ(t) by FPCA	41	4751.1	4945.3

Table 5: Test of the fixed effects including the functional covariate.

Source of variation	F value	p value
Intercept	40590.0	0.00
Fertilizer	1.1	0.35
Irrigation	4.4	0.01
Time	506.7	0.00
Fertilizer:Irrigation	0.9	0.55
Fertilizer:Time	2.1	0.09
Irrigation:Time	6.1	0.00
Fertilizer:Irrigation:Time	0.7	0.72

4. Discussion

We propose an alternative for modeling data from experimental designs including LD and a functional covariate. In this context, we include data from the response variable and a functional covariate recorded under the levels of various factors. Basis functions and functional principal component analysis were considered as approaches to include this covariate in the model. In the paper by Kundu et al. (2016) penalized least squares are used while Staicu et al. (2020) propose employing penalized maximum likelihood. In our approach, after using basis functions and functional principal components, the estimation problem becomes a classical problem of a linear mixed model with longitudinal data where the estimation is carried out by restricted maximum likelihood (REML) and generalized least squares. This approach is computationally more straightforward and allows statistical inference in a classical sense.

We also model the covariance matrix of the errors (it was not assumed to be an identity matrix). To assess the statistical significance of the functional parameter, confidence bands were obtained using a combination of parametric bootstrap in mixed models and bootstrap methods for functional data (Febrero-Bande et al., 2010; Febrero-Bande & Oviedo de la Fuente, 2012). The AIC and BIC values highlighted the importance of including the random slope in the model, indicating differences among plants regarding linear growth rates of chlorophyll concentration. In both modeling approaches, smaller values of these statistics were obtained when the functional covariate was included. We also found that the covariance structure AR(1) was the most suitable for modeling the covariance structure of the errors. We propose the FPCA as a method for summarizing the spectral signature information and facilitate its inclusion in a mixed model. This approach avoids dealing with high dimensionality and collinearity (Aguilera et al., 2006). In summary, we give an alternative for modeling experimental designs including longitudinal data and a functional covariate. We used basis functions and functional principal components to smooth the functional data. Other alternatives, such as functional partial least squares, could also be considered.

Remote sensing as a tool for crop characterization is based on the construction of indices (vegetation indices) derived from the observation of spectral signatures. Among the most well-known indices are CTR1 and CTR2 by Carter (1994), the modified Red Edge Ratio simple (mSR705) proposed by Sims & Gamon (2002),

the indices VOG1, VOG2, and VOG3 proposed by Vogelmann et al. (1993), and the NDVI705 and mNDVI705 indices proposed by Gitelson & Merzlyak (1994), among others. Unlike these indices, the proposed model allows for a complete analysis of the spectral signature. Additionally, an alternative to the method given by Goldsmith et al. (2012) for modeling the residual variance-covariance matrix is proposed, which can model dependencies in longitudinal data (Verbeke & Molenberghs, 2000). In this case study, it was found that the first-order autoregressive dependence structure is the most suitable for the residual variance-covariance matrix.

[Received: March 2024 — Accepted: September 2024]

References

Aguilera, A., Escabias, M. & Valderrama, M. (2006), 'Using principal components for estimating logistic regression with high-dimensional multicollinear data', *Computational Statistics & Data Analysis* **50**(8), 1905–1924.

Bonamy, M., de Iraola, J., Prando, A., Baldo, A., Giovambattista, G. & Rogberg-Muñoz, A. (2020), 'Application of longitudinal data analysis allows to detect differences in pre-breeding growing curves of 24-month calving angus heifers under two pasture-based system with differential puberty onset', *Journal of the Science of Food and Agriculture* **100**(2), 714–720.

Cederbaum, J., Pouplier, M., Hoole, P. & Greven, S. (2016), 'Functional linear mixed models for irregularly or sparsely sampled data', *Statistical Modelling* **16**(1), 67–88.

Crainiceanu, C., Ciprian, M. & Ruppert, D. (2004), 'Likelihood ratio tests in linear mixed models with one variance component', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **66**(1), 165–185.

Davis, C. (2002), *Statistical methods for the analysis of repeated measurements*, Springer, New York.

Febrero-Bande, M., Galeano, P. & González-Manteiga, W. (2010), 'Measures of influence for the functional linear model with scalar response', *Journal of Multivariate Analysis* **101**(2), 327–339.

Febrero-Bande, M. & Oviedo de la Fuente, M. (2012), 'Statistical computing in functional data analysis: the R package *fda.usc*', *Journal of Statistical Software* **51**(4), 1–28.

Fenzi, M., Jarvis, D., Reyes, L., Moreno, L. & Tuxill, J. (2017), 'Longitudinal analysis of maize diversity in Yucatan, Mexico: influence of agro-ecological factors on landraces conservation and modern variety introduction', *Plant Genetic Resources* **15**(1), 51–63.

Ferraty, F. & Vieu, P. (2006), *Nonparametric functional data analysis: theory and practice*, Springer Science & Business Media.

Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. (2008), *Longitudinal data analysis*, CRC Press, New York.

Gitelson, A. & Merzlyak, M. (1994), 'Spectral reflectance changes associated with autumn senescence of *Aesculus hippocastanum* L. and *Acer platanoides* L. leaves. spectral features and relation to chlorophyll estimation', *Journal of Plant Physiology* **143**(3), 286–292.

Goldsmith, J., Greven, S. & Crainiceanu, C. (2012), 'Corrected confidence bands for functional data using principal components', *Biometrics* **69**(1), 41–51.

Greven, S. & Kneib, T. (2010), 'On the behaviour of marginal and conditional AIC in linear mixed models', *Biometrika* **97**(4), 773–789.

Guo, W. (2002), 'Functional mixed effects models', *Biometrics* **58**(1), 121–128.

Guo, W. (2004), 'Functional data analysis in longitudinal settings using smoothing splines', *Statistical Methods in Medical Research* **13**(1), 49–62.

Gómez-Escobar, G. (2017), Un modelo lineal mixto con covariable funcional aplicado a datos de concentración de clorofila, Tesis de maestría, Universidad del Valle, Facultad de Ingeniería, Escuela de Estadística, Santiago de Cali, Colombia.

Hardin, J. & Hilbe, J. (2002), *Generalized estimating equations*, Chapman and Hall/CRC, New York.

Hedeker, D. & Gibbons, R. (2006), *Longitudinal data analysis*, John Wiley & Sons, New York.

Horváth, L. & Kokoszka, P. (2012), *Inference for functional data with applications*, Springer, Science & Business Media.

Jones, R. (2011), 'Bayesian information criterion for longitudinal and clustered data', *Statistics in Medicine* **30**(25), 3050–3056.

Kundu, M., Harezlak, J. & Randolph, T. (2016), 'Longitudinal functional models with structured penalties', *Statistical Modelling* **16**(2), 114–139.

Lahiri, S. N. (2003), *Resampling methods for dependent data*, Springer Series in Statistics, Springer-Verlag, New York.

Lark, R., Ligowe, I., Thierfelder, C., Magwero, N., Namaona, W., Njira, K., Sandram, I., Chimungu, J. & Nalivata, P. (2020), 'Longitudinal analysis of a long-term conservation agriculture experiment in Malawi and lessons for future experimental design', *Experimental Agriculture* **56**(4), 506–527.

Ling, Q., Huang, W. & Jarvis, P. (2011), 'Use of a SPAD-502 meter to measure leaf chlorophyll concentration in *Arabidopsis thaliana*', *Photosynthesis Research* **107**(2), 209–214.

Liu, B., Wang, L. & Cao, J. (2017), 'Estimating functional linear mixed-effects regression models', *Computational Statistics & Data Analysis* **106**, 153–164.

Liu, Z. & Guo, W. (2012), 'Functional mixed effects models', *Computational Statistics* **4**(6), 527–534.

Miqueloni, D. (2019), 'About the use of longitudinal data analysis in forage legumes breeding: A review', *International Journal of Environment, Agriculture and Biotechnology* **4**(4), 1249–1262.

Park, S. & Staicu, A. (2015), 'Longitudinal functional data analysis', *Stat* **4**(1), 212–226.

Patterson, D. & Thompson, R. (1971), 'Recovery of inter-block information when block sizes are unequal', *Biometrika* **58**(3), 545–554.

Ramsay, J. & Silverman, B. (2005), *Functional data analysis*, Springer-Verlag, New York.

Sims, D. & Gamon, J. (2002), 'Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and development stages', *Remote Sensing of Environment* **81**(2-3), 337–354.

Staicu, A., Islam, M., Dumitru, R. & van Heugten, E. (2020), 'Longitudinal dynamic functional regression', *Journal of the Royal Statistical Society. Series C, Applied Statistics* **69**(1), 25–46.

Verbeke, G. & Molenberghs, G. (2000), *Linear mixed models for longitudinal data*, Springer-Verlag, New York.

Vogelmann, J., Rock, B. & Moss, D. (1993), 'Red edge spectral measurements from sugar maple leaves', *International Journal of Remote Sensing* **14**(2), 1563–1575.

Weiss, R. (2005), *Modeling longitudinal data*, Springer, New York.