Un modelo lineal mixto con covariable funcional aplicado a datos de concentración de clorofila

Gustavo Adolfo Gómez Escobar

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2017

Un modelo lineal mixto con covariable funcional aplicado a datos de concentración de clorofila

Gustavo Adolfo Gómez Escobar

Tesis o trabajo de investigació presentada(o) como requisito parcial para optar al título de:
Magister en Estadística

Director(a):
Ph.D Mercedes Andrade Benjarano
Codirector(a):
Ph.D Ramón Giraldo Henao

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2017

(Dedicatoria)

Este trabajo de investigación está dedicado a Francia Honoría Escobar, porque su memoria es y será mi fuente de inspiración.

Agradecimientos

Primeramente quiero agradecer a Dios por llenar mi camino de bendiciones, a mi madre Fanny Escobar por ser mi apoyo, mi luz y mi horizonte. A mis hermanos; Aura María Gómez y Jairo Andrés Gómez por darles ese toque de felicidad a mi vida.

Aprovecho para brindarle mi gratitud a la directora de este trabajo de investigación, la doctora Mercedes Andrade, que gracias a su entusiasmo y dedicación fue posible este trabajo. También agradezco al doctor Ramón Giraldo por sus valiosos aportes. A mis compañeros de estudios y de lucha Luis Carlos Bravo y Jennyfer Yela.

Resumen

El presente trabajo de investigación tiene como objetivo modelar la concentración de clorofila en plantas de ají de tabasco a través de un modelo lineal mixto con covariable funcional. Las plantas han sido sometidos a dos fuentes de estrés causados por el tipo de fertilizante y el nivel de riego, también se usa la firma espectral como covariable funcional. Se propone dos alternativas para involucra la firma espectral como covariable funcional en el modelo lineal mixto. Por medio de bandas de confianza se encontró que la firma espectral es significativo para explicar la concentración de clorofila.

Palabras claves: Modelo mixto, covariable funcional, clorofila, firma espectral

Abstract

The aim of this research work is to model the chlorophyll concentration in Tabasco pepper plants through a linear mixed model with functional covariable. The plants have been subjected to two sources of stress caused by the type of fertilizer and the level of irrigation, the spectral signature was also used as a functional covariate. Two alternatives are proposed to involve the spectral signature as a functional covariable in the linear mixed model. Through confidence bands it found that the spectral signature is meaningful to explain the chlorophyll concentration

Keywords: Mixed model, functional covariante, chlorophyll, spectral signature

Contenido

Lista de Figuras

Lista de Tablas

Declaración

Me permito afirmar que he realizado el presente Trabajo de Grado de manera autónoma y con la única ayuda de los medios permitidos y no diferentes a los mencionados en el propio trabajo. Todos los pasajes que se han tomado de manera textual o figurativa de textos publicados y no publicados, los he reconocido. Ninguna parte del presente trabajo se ha empleado en ningún otro tipo de Tesis o Trabajo de Grado.

Igualmente declaro que los datos utilizados están protegidos por las correspondientes cláusulas de confidencialidad.

Santiago de Cali, 08.11.2017

(Gustavo Adolfo Gómez Escobar)

1 Introducción

El análisis de las firmas espectrales ha surgido como una herramienta de suma importancia en la modelación en agricultura. Se emplea como variable predictora de, entre otras, las concentraciones de clorofila y nitrógeno o de la absorción de líquidos (De la Cruz-Durán et al. 2011). Por ejemplo, en el trabajo de Murrillo & Carbonell (2012) se estima la concentración de clorofila a partir de índices de vegetación que se calculan con base en la información espectral de dos o más longitudes de onda.

El avance tecnológico permite obtener muchos registros discretos para cada uno de los individuos en una muestra en particular. Estos pueden ser considerados, después de llevar a cabo una etapa preliminar de suavizado o interpolación, como funciones (curvas o superficie). Un caso particular de esta situación es el de las firmas espectrales, en los que muchos datos de reflectancia se obtienen en función de longitud de onda. El análisis de datos funcionales (ADF) (Ferraty & Vieu 2006) ha surgido como una herramienta estadistica apropiada para el tratamiento de este tipo de información. Dentro del ADF las funciones bajo estudio se llaman datos funcionales (Ramsay & Silverman 2005) y su mecanismo generador variable o covariable funcional.

Algunas referencias muy relevantes en este contexto son Ramsay & Silverman (2005), Horváth & Kokoszka (2012) y Ferraty & Vieu (2006). Casi todas las metodologías estadísticas clásicas han sido extendidas al tratamiento de datos funcionales. Análisis exploratorio, técnicas multivariadas (componentes principales, correlación canónica, análisis discriminante y cluster), modelos linealies y series de tiempo, son entre otros, campos de la estadística que tienen su análogo funcional. En el caso particular del modelo funcional las primeras referencias son Ramsay & Dalzell (1991), Frank & Friedman (1993) y Cardot et al. (1999). En los modelos propuestos por estos autores se tiene una respuesta escalar y una variable funcional.

Algunas aplicaciones de la regresión funcional son las dadas entre otros por Frank & Friedman (1993) quienes emplearon este tipo de modelos para el análisis de datos

quimiométricos y por Ramsay & Silverman (1997) quienes usaron en datos climatológicos, modelando la cantidad total de precipitación anual a partir de la información de la curva de temperatura mensual para los datos de las estaciones meteorológicas de Canadá. En cuanto al uso de las firmas espectrales, Ferraty & Vieu (2002) modelan la relación entre la concentración de grasas en piezas de carne a partir del espectro de absorción de la luz, el cual es una curva indexada sobre longitudes de ondas.

Los modelos de respuesta escalar y variable explicatoria funcional se extendieron al caso de distribuciones de la familia exponencial (McCulloch & Nelder 1989), es decir, se proponen el modelo generalizado funcional con respuesta escalar. En particular existen varias aplicaciones en el uso de datos binarios (Marx & Eilers (1999), Cardot & Sarda (2005), Escabias et al. (2006), Preda et al. (2007), Crainiceanu et al. (2009), Mousavi (2015)).

En este proyecto de investigación se retoma el trabajo de grado titulado "Evaluación de la respuesta espectral en plantas de ají tabasco bajo diferentes condiciones de riego y fertilización" (Ocampo 2015), en el cual se buscaba establecer la relación entre la concentración de clorofila (unidades spad) de plantas bajo condiciones de estrés causadas por el tipo de fertilizante y el nivel de riego aplicados. En total se consideraron 16 tratamientos (4 condiciones de riego y 4 niveles de fertilización) con 8 réplicas (8 plantas de ají tabasco). En cada planta se midió la concentración de clorofila y la firma espectral con longitudes de onda entre 500 y 950 en nanómetros (nm). Las mediciones fueron tomadas en forma longitudinal durante 7 semanas consecutivas.

El experimento descrito corresponde a un diseño de medidas repetidas, en el que además de los tratamientos se tiene la firma espectral de cada planta como covariable. La estructura longitudinal de la respuesta limita el uso del modelo de regresión funcional con respuesta escalar y covariable funcional (Cardot et al. 1999) arriba descrito, pues éste supone independencia. Para solucionar este problema surge de manera natural la posibilidad de emplear un modelo mixto (Verbeke & Molenberghs 2000). Teniendo en cuenta lo anterior, en este estudio se busca proponer alternativas para adaptar la teoría de modelos mixtos en el análisis de datos longitudinales cuando hay una covariable funcional. Esto como solución al análisis del experimento anteriormente descrito. En particular se propone dos métodos para incluir la covariable funcional en el modelo mixto. En el primero se emplea una descomposición en términos de base de funciones, de manera similar a lo descrito en Ramsay & Silverman (2005) en el caso del modelo de regresión funcional con respuesta escalar. La segunda alternativa consiste en hacer un análisis de componentes principales en

la covariable funcional y se incluyen los scores como covariable escalares, de manera análoga a a como se propone en Aguilera et al. (2006) en el modelo funcional logístico. Se describe el proceso de estimación de los modelos y en cuanto a la inferencia, se propone realizar las bandas de confianza para el parámetro funcional por medio de método de Bootstrap

Goldsmith et al. (2012) utilizan un modelo de efectos mixtos con covariable funcional aplicado a datos cognitivos. En su aproximación la covariable funcional se representa en términos de funciones propias y el parámetro funcional a través de una base truncada de potencias. En el modelo se asume que el bloque de la matriz de varianzas y covarianzas de los efectos aleatorios es diagonal, lo cual en los modelos de coeficientes aleatorios no se cumple, ya que la pendiente y el intercepto aleatorios son indirectamente correlacionados (Verbeke & Molenberghs 2000).

Las firmas espectrales han sido modeladas de las siguientes formas

- Ferraty & Vieu (2002) propone un modelo no paramétrico funcional y con variable respuesta escalar. El modelo es ilustrado con datos de quimiometría, donde el registro que contiene 215 piezas de carne. El objetivo es modelar el porcentaje de grasa en función de las curvas espectrales de absorbancia.

- Botero et al. (2009), usa la espectometría para la determinación de características foliares de plantas de bananos, donde se correlacionan la concentraciones de nutrientes y los valores de la respuestas espectrales entre las longitudes de onda de 400 a 1050 nm. Mediante una regresión PLS y concluye que este modelo genera buenas estimaciones de las variables de concentración de nutrientes.

- Ordóñez et al. (2010) modela el contenido de agua en 80 hojas de plantas de vid, en función de los valores de reflectancia obtenidos entre el rango de longitudes de onda de 0 a $2000nm$. Proponen un modelo funcional con funciones de base radial.

- Warren et al. (2017) por medio una regresión funcional con respuesta escalar. Tomando las mediciones hiperespectrales entre las longitudes de onda 350 a 2500 nm de la superficie de larvas Lucilia sericata, los autores predicen el día dentro de la etapa posterior a la alimentación. Concluyen que este modelo facilita la predicción. Se destaca también de este trabajo que usaron una base de funciones B-spline con 100 nodos igualmente espaciados para la suavización la firma espectral.

El presente trabajo de investigación se encuentra organizado de la siguiente manera: en el capítulo 2 se muestran los objetivos que se trazaron en el estudio, el capítulo 3 se presenta un marco conceptual sobre el problema de la espectrocopía y un marco teorico estadístico que incluye conceptos de modelo mixto y el análisis de datos funcionales, en el capítulo 4 se presenta la metodología y en los capítulos 5 y 6 se dan los resultados y las conclusiones más relevantes.

2 Objetivos

2.1. Objetivo General

Modelar la concentración de clorofila en unidades SPAD por medio de un modelo mixto en el contexto de datos longitudinales que permita incluir la firma espectral como una covariable funcional.

2.1.1. Objetivos Específicos

- Proponer una metodología para la inclusión de una covariable funcional en un modelo de efectos mixtos usando en términos de bases de funciones y aplicarlo en el estudio de la relación entre la concentración de clorofila (unidades spad) y la firma espectral.
- Proponer una metodología para la inclusión de una covariable funcional en un modelo de efectos mixtos usando análisis de componentes principales funcional y aplicarlo en el estudio de la relación entre la concentración de clorofila (unidades spad) y la firma espectral.
- Comparar la concentración de clorofila de plantas de ají de tabasco para diferentes condiciones de riego y fertilización mediante análisis ANOVA basado en Modelos Mixtos.

3 Marco Teórico

En este capítulo se hace una revisión teórica de los modelos mixtos y análisis de datos funcionales, que son las herramientas posteriormente usadas en la propuesta metodológica del capítulo 4.
Para la modelación de la concentración de clorofila se propone el uso de un modelo de efectos mixtos con cavariable funcional y variable de respuesta escalar, con lo cual se debe de resaltar o tener en cuenta algunos conceptos teóricos, tales como la teoría del modelo mixto y del análisis de datos funcionales. Otros conceptos claves son de contextualización del problema de la espectrocopía.

3.1. Marco Conceptual

A continuación se dan las definiciones de algunos conceptos básicos de la espectroscopia

La teledetección se describe como la técnica que permite adquirir y estudiar los datos de la superficie terrestre desde sensores instalados en plataformas espaciales, en virtud de la interacción electromagnética existente entre la tierra y el sensor, y proviniendo la fuente de radiación ya sea del sol (teledetección pasiva) o del propio sensor (teledetección activa) (Chuvieco 2007).

La espectroscopia es el área que estudia la luz en función de las longitudes de ondas que fueron emitidas, reflejadas o dispersas por un solido, un liquido o un gas (Warner et al. 2009).

La reflectancia es la razón del flujo de radiación reflejado por una superficie a diferentes longitudes de onda (Warner et al. 2009).

El comportamiento espectral de la vegetación, es decir, la cantidad de energía reflectante medida en cada individuo o planta a lo largo del espectro, depende de la naturaleza de esta misma, de sus interacciones con la radiación solar y otros factores climáticos, y de la disposición de nutrientes y agua en su medio ambiente. Por medio de la información espectral

se puede obtener (Jensen 2005):

- La fracción de la cobertura vegetal,
- el contenido de clorofila,
- el índice de área foliar verde,
- Entre otros parámetros biofísicos de las plantas

3.2. Modelo Lineal Mixto

El modelo lineal de efectos mixtos (Verbeke & Molenberghs 2000) puede ser escrito como

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{3-1}$$

donde Y_i es el vector de respuestas n_i dimensional con componentes y_{ij}, $j = 1, 2, ..., n_i$ $i = 1, 2, ..., N$, X_i es la matriz de efectos fijos con covariables conocidas de dimensión $(n_i \times p)$, Z_i es la matriz de efectos aleatorios, con covariables de dimensión $n_i \times q$. β es un vector de dimensión $p \times 1$ de parámetros desconocidos asociados a los efectos fijos del modelo, b_i corresponde a un vector de efectos aleatorios. Se asume que su distribución es $Normal(\mathbf{0}, D)$ y los $b_1, b_2, ..., b_N$ son independientes, por último e_i es un vector de errores residuales $n_i \times 1$ con distribución $Normal(\mathbf{0}, \Sigma_i)$, donde se supone que $e_1, e_2, ..., e_N$ son independientes entre **y** con respecto a los *bi*.
Por ejemplo para un individuo i en el instante j si se encuentra en un conjunto de datos longitudinales, el modelo lineal mixto se puede expresar como (Verbeke & Molenberghs 2000)

$$y_{ij} = x_{ij}^T\beta + z_{ij}b_i + e_{ij},$$

donde x_{ij} y y_{ij} son la j ésima fila respectivas de las matrices X_i y Z_i.

El Modelo Lineal Mixto puede ser expresado por medio de la distribución marginal de Y_i n_i dimensional.

$$y_i \sim \mathbf{Normal}(X_i\beta, V_i)$$

, donde $V_i = Z_i D Z_i^T + \Sigma_i$ Por otro lado la distribución condicional de $Y_i|b_i$ viene dado

$$Y_i|b_i \sim \mathbf{Normal}(X_i\beta + Z_i b_i,\ \Sigma_i)$$

.

3.2.1. Estimación por Máxima Verosimilitud de los efectos fijos

En el modelo lineal mixto gaussiano (Jiang 2007), el interés se centra en la inferencia de la distribución marginal del vector Y_i (Verbeke & Molenberghs 2000), el cual distribuye $y_i \sim \mathbf{Normal}(X_i\beta, V_i)$, donde $V_i = Z_i D Z_i^T + \Sigma_i$. Sea θ el vector que contiene los parámetros de varianza contenidas en V_i, asumiendo que θ es conocido. El estimador de máxima verosimilitud (MLE) para β es

$$\hat{\beta}(\theta) = \left(\sum_{i=1}^{N} X_i^T V_i^{-1}(\theta) X_i \right)^{-1} \sum_{i=1}^{N} X_i^T V_i^{-1}(\theta) Y_i, \tag{3-2}$$

$$= \left(\sum_{i=1}^{N} X_i^T W_i X_i \right)^{-1} \sum_{i=1}^{N} X_i^T W_i Y_i, \tag{3-3}$$

donde W_i es igual a $V_i^{-1}(\theta)$. En la práctica V_i no es conocido y es estimado por $\hat{V}_i = \hat{W}_i^{-1} = V(\hat{\theta})$, donde las componentes de varianzas θ se pueden estimar por máxima verosimilitud o máxima verosimilitud restringida

3.2.2. Estimación por Máxima Verosimilitud Restringida de los efectos fijos

La estimación de máxima verosimilitud restringida (REML) es una estimación de máxima verosimilitud basada en una transformación de la variable respuesta (Jiang 2007).

$$Y_i^* = A_i Y_i,$$

donde la matriz A_i es tal que Y^* no depende de β o sea $E[y^*] = 0$. Una matriz que cumple con esta condición es la matriz A_i que proyecta a Y_i a los residuos de las estimaciones por Mínimos Cuadrados o de Máxima Verosimilitud, dicha matriz viene dada como

$$A_i = I_{n_i} - H_i = I_{n_i} - X_i (X_i^T X_i)^{-1} X_i^T.$$

Por lo tanto la distribución de Y^* es

$$Y^* \sim \mathbf{Normal}(\mathbf{0}, \mathbf{A_i^T V_i}(\theta)\mathbf{A_i}).$$

Así que la función de log verosimilitud restringida viene dado por

$$l_{REML}(\theta) = c - \frac{1}{2}\log(|A^T V A|) - \frac{1}{2} Y^{*T} (A^T V A)^{-1} Y^*,$$

donde c es una constante; el estimador para θ es el valor $\hat{\theta}$ que satisface que $\frac{\partial l_{REML}}{\partial \theta_j} = 0$, donde

$$\frac{\partial l_{REML}}{\partial \theta_j} = \frac{1}{2}\left\{ Y^T P \frac{\partial V}{\partial \theta_j} P - tr(P\frac{\partial V}{\partial \theta_j}) \right\},$$

con $P = A(A^T V A)^{-1} A^T$.

3.2.3. Inferencia para los Efectos Fijos

Como se mencionó anteriormente (en 3.2.1), la estimación de los parámetros de los efectos fijos del modelo lineal mixto 3-1 β depende del vector de los componentes de varianza θ, que puede ser obtenido por MLE o REML. (Verbeke & Molenberghs 2000). $\hat{\beta}(\theta) \sim$ **Normal**$(\beta, Var[\hat{\beta}(\theta)])$ y la matriz de varianzas y covarianzas viene dado por la siguiente expresión

$$\begin{aligned} Var[\hat{\beta}(\theta)] &= \left(\sum_{i=1}^{N} X_i^T V_i^{-1}(\theta) X_i \right)^{-1} \left(\sum_{i=1}^{N} X_i^T V_i^{-1}(\theta) Var(Y_i) V_i^{-1}(\theta) X_i \right) \\ &\quad \left(\sum_{i=1}^{N} X_i^T V_i^{-1}(\theta) X_i \right)^{-1} \\ &= \left(\sum_{i=1}^{N} X_i^T V_i^{-1}(\theta) X_i \right)^{-1} = \left(\sum_{i=1}^{N} X_i^T W_i X_i \right)^{-1}, \end{aligned}$$

donde W_i es igual a $V_i^{-1}(\theta)$. Ya obtenida la distribución del estimador de $\hat{\beta}(\theta)$ de los efectos fijos se presenta a continuación dos aproximaciones para realizar inferencia en el vector β Aproximación Prueba de Wald Para el caso en el que se requiera realizar inferencia en los parámetros individuales de los efectos fijos β_j, $j = 1, 2, ..., p$. La aproximación de la prueba de Wald se puede realizar por medio del estadístico de prueba

$$z = \frac{\hat{\beta}_j - \beta_j}{\hat{se}(\hat{\beta}_j)},$$

que distribuye asintóticamente a una distribución normal estándar. En general para una matriz L, una prueba de hipótesis de la forma

$$H_0 : L\beta = 0 \;\; H_a : L\beta \neq 0,$$

se contrasta con el estadístico de Wald(W) definido como

$$W = (\hat{\beta} - \beta)^T L^T \left[L \left(\sum_{i=1}^{N} X_i^T V_i^{-1}(\hat{\theta}) X_i \right)^{-1} L^T \right] L(\hat{\beta} - \beta),$$

el que tiene una distribución asintótica Chi cuadrado ($\chi^2_{g.l}$) con $g.l$ = rango de L grados de libertad.
El estadístico de Wald está basado en el error estándar estimado de $\hat{\beta}$, el cual subestima la verdadera variabilidad de dicho estimador (Verbeke & Molenberghs 2000), pues no introduce la variabilidad generada por la estimación de θ (Vector de las componentes de variabilidad en $V_i(\theta)$). Una forma de resolver este inconveniente es usar la aproximación por los estadísticos t y F para las hipótesis sobre β.
La inferencia en los parámetros individuales de los efectos fijos β_j, $j = 1, 2, ..., p$ viene dado por $t_j = \frac{\hat{\beta}_j - \beta_j}{\hat{se}(\hat{\beta}_j)}$ que contrasta con una distribución t student, por otro lado para una prueba $H_0 : L\beta = 0$ $H_a : L\beta \neq 0$ se tiene el estadístico F

$$F = \frac{(\hat{\beta} - \beta)^T L^T \left[L \left(\sum_{i=1}^{N} X_i^T V_i^{-1}(\hat{\theta}) X_i \right)^{-1} L^T \right] L(\hat{\beta} - \beta)}{rango(L)}.$$

Bajo la hipótesis nula F presenta una distribución F con los grados del libertad del numerador es igual a $rango(L)$. Los grados de libertad del denominador deben ser estimados a partir de los datos (Verbeke & Molenberghs 2000).

3.2.4. Estimación e inferencia Sobre $\mathbf{b}_i$

Si bien es importante la estimación de los efectos fijos, también se puede considerar fundamental la estimación de los efectos aleatorios $\mathbf{b}_i$ que es de gran ayuda para realizar predicciones sobre los perfiles de los individuos (Verbeke & Molenberghs 2000). Partiendo del hecho que $\mathbf{b}_i$ es considerada una variable aleatoria y considerando que:

- $Y_i|\mathbf{b}_i \sim Normal(X_i\beta + Z_i b_i, \Sigma_i)$, en otras palabras $f(\mathbf{y}_i|\mathbf{b}_i)$,
- $b_i \sim Normal(\mathbf{0}, D)$, en otras palabras $f(\mathbf{b}_i)$,

es razonable considerar que la distribución de $\mathbf{b}_i$ es una distribución a priori, pues no depende del conjunto de datos Y_i, es así que se puede usar el enfoque Bayesiano para hacer estimaciones de $\mathbf{b}_i$, esto es, con base en una distribución a posteriori $f(\mathbf{b}_i|Y_i = \mathbf{y}_i)$

$$f(\mathbf{b}_i|Y_i = y_i) = \frac{f(\mathbf{y}_i|\mathbf{b}_i)f(\mathbf{b}_i)}{\int f(\mathbf{y}_i|\mathbf{b}_i)f(\mathbf{b}_i)d\mathbf{b}_i}.$$

La distribución a posteriori es una normal multivariada con vector de medias y matriz de varianzas y covarianzas dada por

$$\hat{\mathbf{b}}_i(\theta) = E[\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i] = DZ_i^T W_i(\theta)(\mathbf{y}_i - X_i\beta),$$

$$Var[\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i] = DZ_i^T \left\{ W_i - W_i X_i \left(\sum_{i=1}^{N} X_i^T W_i X_i \right)^T X_i^T W_i \right\} Z_i D,$$

$\hat{\mathbf{b}}_i(\theta)$ es estimada a partir de la media aposteriori $E[\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i]$ con una variabilidad de $Var[\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i]$.

3.3. Diagnóstico de Residuales

En cuanto al analisis de los supuestos de los modelos planteados, éstos son los mismos que los que se plantean en un modelo mixto tradicional, estos es; la normalidad multivariada tanto en los efectos aleatorios como a los errores, la homogeneidad de varianzas.

En este trabajo se explora el cumplimiento de estos supuestos por medio de método gráficos, a partir de la descomposición de Cholesky de la matriz de varianzas y covarianzas estimada de los errores $\hat{\Sigma} = LL^T$, donde L es una matriz triangular inferior. Para ello se define los siguientes residuales

$$r_i^* = L^{-1}(\mathbf{y}_i - \tilde{X}_i\hat{\beta}) \tag{3-4}$$

Con esto se realiza un gráfico de cuantil cuantil para evaluar de manera visual el supuesto de normalidad en los errores, en cuanto a la homogeneidad de varianzas ésta se hace por medio de gráfico de los residuales versus valores ajustados.

3.4. Análisis de Datos Funcionales

3.4.1. Espacio L^2

Si perder generalidad, el espacio L^2 o $L^2[0,1]$ es el conjunto de las funciones reales f medibles en el conjunto cerrado $[0,1]$ que cumplen la condición de ser de cuadrado integrables (Horváth & Kokoszka 2012), es decir.

$$\int_T f^2(t)dt < \infty.$$

El espacio L^2 es un espacio de Hilbert separable con producto interno $<,>$ definido como

$$< f, g >= \int_T f(t)g(t)dt,$$

donde f y $g \in L^2$. Otras propiedades del espacio L^2 de las funciones cuadrado integrable se enuncian a continuación
Si f, g y h pertenecen a $L^2[T]$ y $c \in R$, entonces se tiene que:

- $< f, g >=< g, f >$ propiedad de simetría,
- $< f, f >= 0$ si y solo si $f = 0$,
- $< af + bg, h >= a < f, h > +b < g, h >$,

la norma del espacio $L^2(T)$ se define como:

$$||f(t)|| = \sqrt{\int_T f^2(t)dt}$$

y por consiguiente

$$||f(t)||^2 = \int_T f^2(t)dt.$$

Algunas propiedades de la norma $L^2[T]$ son:

1. $||f|| > 0$ y $||f|| = 0$ si y solo si $f = 0$
2. $||af|| = |a|||f||$ siendo $a \in R$
3. $||f + g|| \leq ||f|| + ||g||$

Una vez definidos los espacio L^2 y el espacio de Hilbert, se pueden dar las siguientes definiciones respecto a los datos funcionales.

Definición 1: Una variable aleatoria $\boldsymbol{\chi}$ se llama variable funcional $(v.f)$ si toma valores en un espacio de dimensión infinita o espacio funcional y una observación χ de $\boldsymbol{\chi}$ es llamado un dato funcional. (Ferraty & Vieu 2006).

La $v.f$ $\chi = \{\chi(t) : t \in T\}$ y donde la indexación T, $T \in \mathbb{R}$, habitualmente es tomada como el tiempo, aunque en el caso de este trabajo T indexará la longitud de onda de la firma espectral. Por otro lado se puede resaltar que T no necesariamente debe de ser unidimensional, por ejemplo, $T \in \mathbb{R}^2$, en este caso se tendría una superficie aleatoria.

3.4.2. De datos discretizados a datos funcionales

En la práctica $\chi(t)$ se registra de manera discreta en un conjunto de puntos $t_1 < t_2 < \cdots < t_m \in T$.
Uno de los mecanismos más usados para la obtención de datos funcionales a partir de datos discretizados es el uso de una base de funciones (Ramsay & Silverman 2005). Esto es, se asume que cada trayectoria $\chi(t)$ puede ser reconstruida a partir de una base de funciones conformada por $\{\phi_1(t), \phi_2(t), ..., \phi_k(t)\}$, es decir

$$\chi_i(t) = \sum_{j=1}^{k} c_{ij}\phi_j(t) = \mathbf{c}_i^T\phi, \quad i = 1, 2, ..., n \tag{3-5}$$

donde $\mathbf{c}_i$ es el vector de coeficientes $\phi(t)$, es un vector de funciones cuyos componentes corresponden a las funciones bases ϕ_j, $j = 1, 2, ..., k$. La elección del número de elementos k es fundamental, pues incide en la calidad de la representación de la variable funcional. Existen varios métodos para la selección de k. Uno de ellos es el criterio de validación cruzada generalizada (GCV)(Ramsay & Silverman 2005) .

3.4.3. Estimación de Coeficientes

La estimación por Mínimos Cuadrados de los coeficientes $\mathbf{c}_i$ en 3-5 se halla minimizando respecto a $\mathbf{c}_i$

$$\boldsymbol{\Phi}_i = (\phi_j(t_{ik}))_{mp}$$

Se obtiene por mínimos cuadrado (Ramsay & Silverman 2005)

$$SCR(\mathbf{c}_i) = (\chi_i(t) - \mathbf{\Phi}_i\mathbf{c}_i)^T (\chi_i(t) - \mathbf{\Phi}_i\mathbf{c}_i). \tag{3-6}$$

Después de derivar se obtiene

$$\hat{\mathbf{c}}_i = \left(\mathbf{\Phi}_i^T\mathbf{\Phi}_i\right)^{-1}\mathbf{\Phi}_i^T\chi_i(t).$$

3.4.4. Tipos de Bases de Funciones

Base de B-Splines

En De Boor (1977) se presenta la definición de las funciones *B-Splines* de forma recursiva. Si se denota $\kappa_1 < \kappa_2 < \cdots < \kappa_k$ particiones del intervalo T (nodos) y k corresponde al número de nodos, entonces el conjunto de funciones bases de orden 1 para los k nodos puede ser formulada como:

$$B_{j,1} = \begin{cases} 1 & Si \quad \kappa_j \leq t \leq \kappa_{j+1} \\ 0 & \qquad Otro\ Caso \end{cases} \tag{3-7}$$

En general las bases de funcionales de orden $p+1$ (de grado p) se define recursivamente como

$$B_{j,p+1}(t) = \frac{t-\kappa_j}{\kappa_{j+p}-\kappa_j}B_{j,p}(t) + \frac{\kappa_{j+p+1}-t}{\kappa_{j+p+1}-\kappa_{j+1}}B_{j+1,p}(t) \tag{3-8}$$

Cuando el orden es 4, es decir, de grado 3 se le conoce como base de funciones **B-Splines** cúbicos.

Base Fourier

Es quizás la expansión básica más conocida y se construye como sumas de funciones trigonometricas, de senos y cosenos. La i ésima trayectoria se obtiene como (Ramsay & Silverman 2005)

$$\chi_i(t) = c_0 + c_1\sin(wt) + c_2\cos(wt) + c_3\sin(2wt) + c_4\cos(2wt) + \cdots$$

En este caso las funciones son

$$\phi_1(t) = 1,\ \phi_{2j-1}(t) = \sin(jwt),\ \phi_{2j}(t) = \cos(jwt),$$

con $j = 1, 2, ...$ y $w = \frac{2\pi}{T}$

Base Wevelet

Si se considera una función Wavelet madre de Haar $\psi()$ (Haar 1910) se puede obtener las funciones básicas (Ramsay & Silverman 2005)

$$\psi_{kj} = 2^{k/2}\psi(s^k t - t),$$

con k y j enteros. Usualmente la función Wavelet de Haar asegura que la base sea ortogonal, es decir, que el producto interno de las funciones de la base sea igual a 1.

Base Funciones Constantes

Suponga una partición del intervalo T, que define los nodos $a_0 < a_1 < \cdots < a_k$. Entonces se tiene los intervalos de la forma $(a_{j-1}, a_j]$ para $j = 1, 2, ...,$. Las funciones básicas están dado por (Ramsay & Silverman 2005)

$$\phi_i(t) = (a_j - a_{j-1})^{-1/2}\mathbb{I}(t)_{(t_{j-1},t_j)}.$$

Base Polinomiales

Las funciones de la forma

$$\phi_j(t) = (t - w)^j \quad con \ \ j = 1, 2, ..., p,$$

conforma la base polinomial, donde w es un parámetro a elegir. Usualmente se escoge w como el centro del intervalo T (Ramsay & Silverman 2005).

3.5. Análisis descriptivo funcional

El análisis exploratorio o análisis descriptivo funcional definido en Ramsay & Silverman (2005) son

- Media $\bar{\chi}(t) = \frac{1}{n}\sum_{i=1}^{n}\chi_i(t)$
- Varianza $Var(\chi(t)) = \frac{1}{n-1}\sum_{i=1}^{n}(\chi_i(t) - \bar{\chi}(t))^2$
- Covarianza $Cov(\chi(t), \chi(s)) = \frac{1}{n-1}\sum_{i=1}^{n}(\chi_i(t) - \bar{\chi}(t))(\chi_i(s) - \bar{\chi}(s))$

3.6. Modelo de Regresión Funcional con Respuesta Escalar

Definición 2. Sea $y_1, y_2, ..., y_n$ una muestra aleatoria y $\chi_1(t), \chi_2(t), ..., \chi_n(t)$ un conjunto de funciones que constituye realizaciones de un proceso estocástico $\chi = \{\chi(t) : t \in T\}$. Se define entonces el modelo de regresión lineal funcional con respuesta escalar y covariable funcional como (Ramsay & Silverman 2005)

$$y_i = \alpha + \int_T \chi_i(t)\beta(t)dt + \epsilon_i, \tag{3-9}$$

con $\beta(t) \in L_2(T)$, $\epsilon_i \sim N(0, \sigma^2)$ y $Cov(\epsilon_i, \epsilon_j) = 0$ para $i \neq j$.
Para el desarrollo del modelo funcional la variable y el parámetro funcional se representan, usualmente, en términos de una base de funciones, es decir

$$\beta(t) = \sum_{j=1}^{k_\beta} b_j\theta_j(t), \quad \beta(t) = \theta^T\mathbf{b},$$

y

$$\chi_i(t) = \sum_{j=1}^{k_\chi} c_{ij}\phi_j(t), \quad \chi_i(t) = \mathbf{c}_i\phi,$$

donde $\mathbf{b}$ y θ son vectores de k_β componentes, $\mathbf{c}_i$ y ϕ son vectores de dimensión k_χ, para $i = 1, 2, ..., n$. Con esta representación el modelo funcional se escribe como

$$\begin{aligned} y_i &= \int_T \chi_i(t)\beta(t)dt + \epsilon_i \\ &= \int_T \mathbf{c}_i\phi\theta^T\mathbf{b}dt + \epsilon_i = \mathbf{c}_i\mathbf{J}_{\phi\theta}\mathbf{b} + \epsilon_i, \end{aligned}$$

con

$$\mathbf{J}_{\phi\theta} = \int_T \phi(t)\theta^T(t)dt = \begin{bmatrix} \int_T \phi_1(t)\theta_1(t)dt & \cdots & \int_T \phi_1(t)\theta_{k_\beta}(t)dt \\ \vdots & \ddots & \vdots \\ \int_T \phi_{k_x}(t)\theta_1(t)dt & \cdots & \int_T \phi_{k_x}(t)\theta_{k_\beta}(t)dt \end{bmatrix}.$$

El modelo completo para toda la muestra $y_1, y_2, ..., y_n$ se puede expresarse matricialmente como

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{c}_1\mathbf{J}_{\phi\theta} \\ 1 & \mathbf{c}_2\mathbf{J}_{\phi\theta} \\ \vdots & \vdots \\ 1 & \mathbf{c}_n\mathbf{J}_{\phi\theta} \end{bmatrix} \begin{bmatrix} \alpha \\ b_1 \\ \vdots \\ b_{k_\beta} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}. \tag{3-10}$$

La ecuación tiene la forma de un modelo lineal simple

$$\mathbf{Y} = \mathbf{Z}\xi + \epsilon,$$

El vector de parámetros ξ se puede estimar por mínimos cuadrados ordinarios como

$$\hat{\xi} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y},$$

con varianza

$$Var[\hat{\xi}] = Var[(\hat{\alpha}, \hat{\mathbf{b}})] = \sigma_\epsilon^2(\mathbf{Z}^T\mathbf{Z})^{-1}. \tag{3-11}$$

La varianza σ_ϵ^2 en la ecuación 3-11 se estima por

$$\begin{aligned} \hat{\sigma}_\epsilon^2 &= \frac{SSE}{n - Tr(S)} \\ &= \frac{(\mathbf{Y} - \hat{\mathbf{Y}})^T(\mathbf{Y} - \hat{\mathbf{Y}})}{n - Tr(S)}. \end{aligned}$$

La matriz S es la matriz que proyecta a $\mathbf{Y}$ en $\hat{\mathbf{Y}}$, es decir,

$$\hat{\mathbf{Y}} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y} = S\mathbf{Y}.$$

Bajo el supuesto de normalidad un intervalo de confianza del $100(1-\alpha)\,\%$ para $\beta(t)$ es

$$= \theta^T(t)\hat{\mathbf{b}} \pm 2\sqrt{\theta^T(t)Var[\hat{\mathbf{b}}]\theta(t)}$$

3.6.1. Análisis de Componentes Principales Funcionales

La matriz de diseño generada anteriormente es posible que presente problemas de multicolinealidad debido a las columnas generadas por el producto de $C^T\mathbf{J}_{\phi\theta}$. Una forma de amortiguar este inconveniente y reducir la dimensión de la matriz de diseño $\tilde{X}_i$, es por medio del uso de Análisis de Componentes Funcionales (ACPF) (Aguilera et al. 2006), el cual se describe a continuación.

Para iniciar el ACPF se debe de centrar los datos funcionales, esto es: Se define la variable $\chi^c(t) = \chi(t) - \mu(t)$, donde $E[\chi(t)] = \mu(t)$ corresponde a la esperanza funcional y dada la covarianza funcional $E[\chi^c(t), \chi^c(s)] = \Sigma(t,s) = \Sigma$, la cual se puede descomponer según el teorema de Karhunen y Loève (Loeve 1997) en

$$\Sigma(t,s) = \sum_{K=1}^{\infty} \lambda_k \xi_k(t) \xi_k(s), \tag{3-12}$$

el cual se puede aproximar o truncar por medio de la suma de los primeros K términos.

$$\Sigma(t,s) = \sum_{K=1}^{K} \lambda_k \xi_k(t) \xi_k(s), \tag{3-13}$$

donde los valores λ_k son los valores propios ($\lambda_1 \geq \lambda_2 \geq \cdots$) asociados a las funciones propias ortogonales $(\xi_1(t), \xi_2(t), ...)$. $\chi^c(t)$ se puede expandir de la forma

$$\chi(t) - \mu(t) = \chi^c(t) = \sum_{k=1}^{\infty} \alpha_k \xi_k(t), \tag{3-14}$$

y para las primeras K funciones propias

$$\chi(t) - \mu(t) = \chi^c(t) = \sum_{k=1}^{K} \alpha_k \xi_k(t), \tag{3-15}$$

donde $\alpha_k = \int_T \chi^C(t) \xi_k(t) dt$ corresponde al k ésimo componente principal funcional o también conocido como Score. Los α_k son variables no correlacionados por pares y con valor esperado $E[\alpha_k] = 0$ y varianza $Var[\alpha_k] = \lambda_k$.

Las ecuaciones propias funcionales corresponden a

$$\int_T \Sigma(t,s) \xi_k(t) dt = \lambda_k \xi_k(s), \tag{3-16}$$

donde $\Sigma(t,s)$ es la autocovarianza funcional de $\chi^c(t)$. $\xi_k(t)$ y λ_k corresponden a las k ésimos funciones y valores propios. Si los datos son expandidos por k_χ funciones de una base de funciones, se tiene que

$$\chi^c(t) = \sum_{k=1}^{k_\chi} c_k \phi_k(t), \quad \chi(t)^c = \mathbf{c}\phi. \tag{3-17}$$

De igual forma se expande la k èsima función propia como la suma de las mismas funciones básicas usadas en la descripción de los datos funcionales

$$\xi_k(t) = \sum_{k=1}^{K_\chi} \delta_k \phi_k(t), \quad \xi_k(t) = \delta_k \phi. \tag{3-18}$$

Entonces la función de autocovarianza funcional está dada por

$$\Sigma(t,s) = n^{-1}\phi(s)^T \mathbf{c}^T \mathbf{c}\phi(t). \tag{3-19}$$

Remplazando en las ecuaciones propias

$$\begin{aligned}
\int_T \Sigma(t,s)\xi_k(t)dt &= \lambda_k \xi_k(s), \\
n^{-1}\phi(s)^T \mathbf{c}^T \mathbf{c} \int_T \phi(t)\phi(t)dt\delta_k &= \lambda_k \phi(s)^T \delta_k, \\
n^{-1}\phi(s)^T \mathbf{c}^T \mathbf{c}\mathbf{J}\delta_k &= \lambda_k \phi(s)^T \delta_k, \\
n^{-1}\mathbf{c}^T \mathbf{c}\mathbf{J}\delta_k &= \lambda_k \delta_k,
\end{aligned}$$

sujeto $\|\xi_k(s)\| = 1$ o también se puede expresar $\delta_k^T \mathbf{J}\delta_k = 1$. Si se define $u_k = \mathbf{J}^T\delta_k$, entonces las ecuaciones propias se puede escribir como

$$n^{-1}\mathbf{J}^{1/2}\mathbf{c}^T\mathbf{c}\mathbf{J}^{1/2}u_k = \lambda_k u_k, \tag{3-20}$$

sujeto a $u_k^T u_k = 1$. Al resolver la ecuación propia se puede obtener los coeficientes de los funciones propias.

$$\delta_k = \mathbf{J}^{1/2}u_k. \tag{3-21}$$

3.7. Diagnóstico de Residuales

En cuanto al análisis de los supuestos de los modelos planteados, éstos son los mismos que los que se plantean en un modelo mixto tradicional, estos es; la normalidad multivariada

tanto en los efectos aleatorios como a los errores, la homogeneidad de varianzas.

En este trabajo se explora el cumplimiento de estos supuestos por medio de método gráficos, a partir de la descomposición de Cholesky de la matriz de varianzas y covarianzas estimada de los errores $\hat{\Sigma} = LL^T$, donde L es una matriz triangular inferior. Para ello se define los siguientes residuales

$$r_i^* = L^{-1}(\mathbf{y}_i - \tilde{X}_i\hat{\beta}) \tag{3-22}$$

Con esto se realiza un gráfico de cuantil cuantil para evaluar de manera visual el supuesto de normalidad en los errores, en cuanto a la homogeneidad de varianzas ésta se hace por medio de gráfico de los residuales versus valores ajustados.

4 Metodología

En esta sección se describe los pasos que se consideran pertinentes destacar para alcanzar los objetivos trazados en el presente trabajo de investigación. En primera instancia se describen los datos analizados y posteriormente se plantean dos alternativas de modelación de los mismos.

4.1. Datos Analizados

La información estudiada corresponde a mediciones longitudinales de clorofila en unidades SPAD-502, registradas semanalmente en 5 hojas de cada una de las 128 plantas de Ají de Tabasco. Las mediciones se realizan bajo las combinaciones de 4 niveles de fertilizante y 4 de riego (ver tabla **4-1**), lo cual corresponde a 16 tratamientos y se contó también con 8 replicas por tratamiento.

Tabla 4-1: Niveles de los factores fertilización y de riego de las unidades experimentales

FACTORES	**NIVELES**	
Fertilización	F1	Solución sin Boro
	F2	Solución sin Hierro
	F3	Solución sin Manganesio
	F4	Solución Completa
Riego	R1	225 (ml)
	R2	15 (ml)
	R3	75 (ml)
	R4	150(ml)

En cada planta se midió también la reflectancia utilizando un espectroradiometro EPP2000, el cual posee una resolución de $0{,}5nm$ en un rango de $500-1100nm$, para obtener en total 900 mediciones de reflectancia. Una base de funciones $B-Splines$ fue usada para posteriormente convertir cada conjunto datos discreto en una curva (Ramsay & Silverman (2005); Ferraty & Vieu (2006)).

Los datos de clorofila se registraron durante 7 semanas consecutivas, con el fin de monitoriar la evolución y el efecto de los factores (fertilizante y riego) y de la firma espectral en la concentración de clorofila en medidas SPAD.

4.2. Análisis Descriptivo Funcional y Definición del Modelo Mixto Funcional

Una fase preliminar del análisis de datos funcionales como en el caso clásico, es realizada por medio de gráficos y de estadísticas descriptivas. Las medidas de localización y variabilidad (Ramsay & Silverman 2005). especificadas en 3.5.

En cuanto a la modelación de la clorofila debido a que la muestra no se puede asumir como independiente, ya a que las medidas son tomados en la misma planta en el tiempo, lo cual es posible que produzca una estructura de correlación entre las mediciones (Verbeke & Molenberghs 2000) y ésta debe ser modelada, por tal motivo se acudió al uso de los Modelos Mixtos

$$Y_i = X_i\beta + Z_i b_i + e_i \tag{4-1}$$

Para incluir la covariable funcional en el modelo 4-1 se propone en el siguiente trabajo dos aproximaciones, los cuales se basan en la intervención o configuración de la matriz X_i. La primera propuesta, es a partir del usos de expansión de los datos funcionales como sumas de funciones de una base y la segunda propuesta basada en la primera propuesta; es realizar un Análisis de Componentes Principales Funcional (ACPF) para evitar incluir un mal acondicionamiento en la matriz X_i producido por multicolinealidad.

4.2.1. Alternativa 1: Modelo Mixto con Covariable Funcional

Para la modelación de la clorofila por medio del modelo mixto empleando una covariable funcional y usando sumas de funciones de bases de funciones se siguió la metodología sugerida en Ramsay & Silverman (2005).

Se tomó la base de funciones B-*Splines*, pues esta cuenta con gran versatilidad para describir funciones y en comparación de la base de Fourier, ésta requiere menos funciones básicas para obtener un buen ajuste (Ramsay & Silverman 2005). Por otro lado se eligió el número de funciones K_χ y el criterio que se tuvo en cuenta para la selección de este parámetro fue

el Criterio de Validación Cruzada Generalizada (GCV) (Ramsay & Silverman 2005).

El modelo lineal mixto que se propone, con base en sumas de funciones básicas parte de la generalización del modelo funcional con respuesta escalar descrita en Ramsay & Silverman (2005), donde la covariable funcional es incluida por el operador

$$\Psi(\chi(t)) = \int_T \psi(t)\chi(t)dt, \tag{4-2}$$

donde $\psi(t), \chi(t) \in H$, siendo H el espacio de Hilbert separable o espacio de las funciones cuadrado integrables. Para un individuo i con n_i mediciones repetidas en el tiempo, se tiene el vector de respuestas $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})^T$, el cual se puede modelar (Goldsmith et al. 2012) como

$$\mathbf{y}_i = \int_T \boldsymbol{\chi}_i(t)\psi(t)dt + X_i\beta + Z_i\mathbf{b}_i + \mathbf{e}_i \tag{4-3}$$

donde

- $\psi(t)$ corresponde al parámetro funcional (Firma espectral) y se obtiene.

$$\psi(t) = \sum_{j=1}^{k_\beta} d_j\theta_j(t), \quad \psi(t) = \theta^T\mathbf{d}, \tag{4-4}$$

- $\boldsymbol{\chi}_i(t)$ es un vector de n_i datos funcionales y está conformada por $\boldsymbol{\chi}_i(t) = (\chi_{i1}(t), \chi_{i2}(t), ..., \chi_{in_i}(t))^T$. la representación de estas observaciones funcionales por medio de funciones básicas para $l = 1, 2, ..., n_i$

$$\chi_{il}(t) = \sum_{k=1}^{k_\chi} c_{ijk}\phi_j(t), \quad \chi_i(t) = \mathbf{c}_{il}\phi, \tag{4-5}$$

  donde $\mathbf{d} = (d_1, d_2, ..., d_{K_\psi})$ y $\mathbf{c}_{il} = (c_{il1}, c_{il2}, ..., c_{ilK_\chi})$ son vectores de coeficientes de tamaño K_ψ y K_χ respectivamente. $\phi = (\phi_1(t), \phi_2(t), ..., \phi_{k_\chi}(t))$ y $\theta = (\theta_1(t), \theta_2(t), ..., \theta_{K_\psi}(t))$ son las primeras funciones de una base de funciones, por ejemplo de la B-*Spline*.

- X_i y Z_i corresponden a las matrices de efectos fijos y aleatorios respectivamente.

- β Corresponde al vector de parámetros de efectos fijos y $\mathbf{b}_i$ es el vector de efectos aleatorios con distribución $Normal(\mathbf{0}, \mathbf{D})$ tal como un Modelo Mixto tradicional

- $\mathbf{e}_i$ es un vector de errores aleatorios con distribución $Normal(\mathbf{0}, \mathbf{\Sigma})$ y además se supone que $\mathbf{b}_i$ y e_i son independientes.

Ahora bien, el modelo expresado para un individuo i viene dado por

$$\begin{aligned}
\mathbf{y}_i &= \int_T \boldsymbol{\chi}_i(t)\boldsymbol{\psi}(t)dt + X_i\beta + Z_i\mathbf{b}_i + \mathbf{e}_i \\
\mathbf{y}_i &= \int_T \mathbf{C}_i^T\boldsymbol{\phi}\boldsymbol{\theta}^T\mathbf{d}dt + X_i\beta + Z_i\mathbf{b}_i + \mathbf{e}_i \\
\mathbf{y}_i &= \mathbf{C}_i^T\mathbf{J}_{\phi\theta}\mathbf{d} + X_i\beta + Z_i\mathbf{b}_i + \mathbf{e}_i.
\end{aligned}$$

$\mathbf{J}_{\phi\theta}$ es una matriz de k_χ por k_β y se define

$$\mathbf{J}_{\phi\theta} = \int_T \phi\theta^T dt = \begin{bmatrix} \int_T \phi_1(t)\theta_1(t)dt & \cdots & \int_T \phi_1(t)\theta_{k_\beta}(t)dt \\ \vdots & \ddots & \vdots \\ \int_T \phi_{k_x}(t)\theta_1(t)dt & \cdots & \int_T \phi_{k_x}(t)\theta_{k_\beta}(t)dt \end{bmatrix}.$$

De forma matricial se obtiene la siguiente expresión

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix} = \begin{bmatrix} \mathbf{c}_{i1}^T\mathbf{J}_{\phi\theta} & x_{i11} & \cdots & x_{i1p} \\ \mathbf{c}_{i2}^T\mathbf{J}_{\phi\theta} & x_{i21} & \cdots & x_{i2p} \\ \vdots & \vdots & & \\ \mathbf{c}_{in_i}^T\mathbf{J}_{\phi\theta} & x_{in_i1} & \cdots & x_{in_ip} \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_{k_\psi} \\ \beta \end{bmatrix} + \begin{bmatrix} z_{i11} & \cdots & z_{i1q} & \\ z_{i21} & \cdots & z_{i2q} & \\ \vdots & \vdots & \ddots & \vdots \\ z_{in_i1} & \cdots & z_{in_iq} & \end{bmatrix} \begin{bmatrix} b_{1i} \\ b_{2i} \\ \vdots \\ b_{iq} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

En el caso en el cual las bases de funciones elegidas para describir los datos funcionales y el parámetro funcional sean las mismas y ortonormales, es decir, que el producto interno entre dos funciones $\int_T \phi_i(t)\theta_j(t)dt$ es igual a 0 si $i \neq i$ y es igual a 1 si $i = j$. En tal caso la matriz de productos internos corresponde a una matriz identidad de orden $K = K_\chi = K_\beta$, entonces el modelo viene dado por:

$$Y_i = \tilde{X}_i\tilde{\beta} + Z_ib_i + e_i, \tag{4-6}$$

donde $\tilde{X}_i = [C^T\mathbf{J}_{\phi\theta},\ X_i]$ o en el caso de las bases ortonormales la matrix de efectos fijos es $\tilde{X}_i = [C^T,\ X_i]$. Después de obtener estas matrices se pueden realizar las estimaciones del modelo por métodos tradicionales del modelo mixto, como máxima verosimilitud y máxima verosimilitud restringida (Patterson & Thompson 1971). Si se asume que $V(\theta) = Z_iDZ_i^t + \Sigma_i$,

siendo θ el vector de parámetros desconocidos de la de la componente de varianza, entonces se puede estimar $V(\theta)$ maximizando

$$l_{REML}(\theta) = c - \frac{1}{2}\log(|A^T V A|) - \frac{1}{2}Y^{*T}(A^T V A)^{-1}Y^*, \tag{4-7}$$

donde $A_i = I_{n_i} - H_i = I_{n_i} - \tilde{X}_i(\tilde{X}_i^t\tilde{X}_i)^{-1}\tilde{X}_i^t$ corresponde a la matriz que proyecta a $\mathbf{y}$ a los residuales. Los efectos fijos son estimados por medio de Mínimos Cuadrados Generalizados (Verbeke & Molenberghs 2000) asumiendo $V(\theta) = \hat{V}(\theta)$

$$\hat{\beta}(\theta) = \left(\sum_{i=1}^{N}\tilde{X}_i^T\hat{V}^{-1}(\theta)\tilde{X}_i\right)^{-1}\sum_{i=1}^{N}\tilde{X}_i^T\hat{V}_i^{-1}(\theta)Y_i. \tag{4-8}$$

$$\hat{b_i} = DZ_i^T\hat{V}(\theta)_i^{-1}(\mathbf{y}_i - X_i\beta). \tag{4-9}$$

La determinación del número de bases *B Splines* necesarias para describir el parámetro funcional K_β es obtenido por medio del Criterio de Información de Akaike Marginal (mAIC) (Greven & Kneib 2010). El mAIC de un modelo marginal de la forma $\mathbf{Y} \sim \mathbf{N}(X\beta, V)$, donde V corresponde a la matriz de componentes de varianzas marginales definida $V = \Sigma + ZDZ^T$ con θ que define el vector de d parámetros que requeridos para conformar V, entonces el mAIC se define (Greven & Kneib 2010)

$$mAIC = -2l(\mathbf{Y}|\hat{\beta}, \hat{V}) + 2(p + d), \tag{4-10}$$

donde $l(\mathbf{Y}|\hat{\beta}, \hat{V})$ es la log verosimilitud maximizada y $p = dim(\beta)$, donde $\beta = (\mathbf{b}, \beta)$

4.2.2. Alternativa 2: Modelo Mixto Mediante ACPF

Dado el modelo

$$\mathbf{y}_i = \int_T \psi(t)\boldsymbol{\chi}_i(t)dt + X_i\beta + Z_i\mathbf{b}_i + \mathbf{e}_i, \tag{4-11}$$

suponiendo que $\boldsymbol{\chi}_i^c(t) = \sum_{k=1}^{K}\alpha_{ik}\xi_k(t)$ y $\psi(t) = \sum_{j=1}^{K}\tau_k\xi_k(t)$, que tanto los datos funcionales como el parámetro funcional están siendo descritos con las mismas funciones propias y a su vez estas funciones propias se describen por medio de K_χ funciones de una base, por ejemplo B-*Spline*

$$\mathbf{y}_i = \int_T \boldsymbol{\chi}_i^c(t)\psi(t)dt + X_i\beta + Z_i\mathbf{b}_i + e_i,$$
$$\mathbf{y}_i = \int_T \mathbf{A}_i^T\boldsymbol{\xi}(t)\boldsymbol{\xi}^T(t)\boldsymbol{\tau}dt + X_i\beta + Z_i\mathbf{b}_i + e_i,$$
$$\mathbf{y}_i = \mathbf{A}_i^T\int_T \boldsymbol{\xi}(t)\boldsymbol{\xi}^T(t)dt\boldsymbol{\tau} + X_i\beta + Z_i\mathbf{b}_i + e_i.$$

Las columnas de la matriz $\mathbf{A}_i$ corresponden a los K scores del individuo i, debido a que los funciones propias pertenecen a una base ortonormal, entonces se puede obtener que $\int_T \boldsymbol{\xi}(t)\boldsymbol{\xi}^T(t)dt = \mathbf{I}_K$. La matriz de la integral de productos internos es igual a la matriz identidad de orden K

$$\mathbf{y}_i = \mathbf{A}_i^T\boldsymbol{\tau} + X_i\beta + Z_i\mathbf{b}_i + e_i, \tag{4-12}$$

donde $\mathbf{A}_i^T = (\alpha_{il})$ corresponde al l ésimo score del i ésimo dato funcional, El modelo final viene dado por

$$Y_i = \tilde{X}_i\tilde{\beta} + Z_ib_i + e_i, \tag{4-13}$$

así $\tilde{X}_i = (\mathbf{A}_i^T, X_i)$ y $\tilde{\beta} = (\boldsymbol{\tau}, \beta)$

4.2.3. Aproximación de banda de confianza vía Bootstrap

En la generación de la banda de confianza para el parámetro funcional $\psi(t)$, se realiza una combinación de métodos de Bootstrap completamente paramétrico en modelos mixtos (Lahiri 2003) y de métodos de Bootstrap en modelos funcionales. Se toman como referencia los procedimientos descritos por Febrero-Bande et al. (2010) y Febrero-Bande & Oviedo de la Fuente (2012), donde describen el método de remuestreo para los modelos de regresiones funcional con respuesta escalar. Suponga el modelo:

$$\mathbf{y}_i = \int_T \boldsymbol{\chi}_i(t)\psi(t)dt + X_i\beta + Z_i\mathbf{b}_i + \mathbf{e}_i. \tag{4-14}$$

Se define el área de confianza AC como:

$$AC(\psi(t)) = \left\{\psi(t) \in L^2 : \left\|\psi(t) - \widehat{\psi}(t)\right\| \leq D_\alpha\right\}, \tag{4-15}$$

donde el estadístico D_α cumple la condición de

$$P\left(\psi(t) \in AC(\psi(t))\right) = 1 - \alpha$$

$$P\left(\left\|\psi(t) - \widehat{\psi}(t)\right\| \leq D_\alpha\right) = 1 - \alpha.$$

El procedimiento es el siguiente.

1. Se ajusta el modelo 4-14 con los datos originales obteniendo sus respectivas estimaciones.

2. Se generan $\{\mathbf{b}_i\}_{i=1}^N$ de tamaño N de una distribución normal multivariada q-dimensional con vector de medias $\mathbf{0}$ y matriz de varianzas y covarianzas $\hat{D}$ estimado en el paso (1).

3. Se generan N muestras para $\{\mathbf{e}_i\}_{i=1}^N$ de tamaños n_i de una distribución normal multivariada con vector de medias $\mathbf{0}$ y la matriz de varianzas y covarianzas $\hat{\Sigma}$ estimada en el paso (1).

4. Se genera una observación bootstrap de la variable respuesta

$$\mathbf{y}_i^* = \int_T \psi(t)\boldsymbol{\chi}_i(t)dt + X_i\beta + Z_i\mathbf{b}_i^* + \mathbf{e}_i^*.$$

5. Se ajusta el modelo 4-14 tomando observaciones bootstrap de la variable respuesta $\mathbf{y}_i^*$, se obtienen las estimaciones del parámetro funcional $\psi(t)^j$.

6. Se replican los pasos anteriores un número de B veces.

7. El valor D_α es estimado $\hat{D}_\alpha$ por medio del percentil $(1-\alpha)$ de

$$d^j = \left\|\widehat{\psi}(t) - \widehat{\psi}(t)^j\right\|$$

$$d^j = \left\|\widehat{\psi}(t) - \widehat{\psi}(t)^j\right\|$$

$$d^j = \int_T \left(\widehat{\psi}(t) - \widehat{\psi}(t)^j\right)^2 dt$$

$P\left(d \leq \hat{D}_\alpha\right) \approx 1 - \alpha.$

8. Finalmente se gráfican las estimaciones bootstrap tales que: $\left\|\widehat{\psi}(t) - \widehat{\psi}(t)^j\right\| \leq \hat{D}_\alpha$.

4.2.4. Modelo para los Datos de Clorofila

Para la modelación de la concentración de clorofila en medidas SPAD usando covariable la firma espectral de las plantas se tiene el siguiente modelo, donde la indexación del parámetro y de los datos funcionales se considera como $\lambda \in [500 - 1100nm]$ con el fin de diferenciar del tiempo. El primer nivel del modelo considerado viene dado por

$$\textbf{Clorofila}_i = \beta_{0i} + \beta_{1.} + \beta_{2.} + \beta_{3i}t + \beta_{4.}t + \beta_{5.}t + \beta_{6.}t + \int_T \psi(\lambda)\boldsymbol{\chi}_i(\lambda)dt + \mathbf{e}_i \quad (4\text{-}16)$$

En el segundo nivel se estudia el cambio de intercepto de los sujetos específicos y la variación de sus respectivas pendientes en términos del tiempo. Se plantea las ecuaciones:

$$\begin{aligned}
\beta_{0i} &= \beta_0 + b_{0i} \\
\beta_{1.} &= \beta_1 F_1 + \beta_2 F_2 + \beta_3 F_3 + \beta_4 R_1 + \beta_5 R_2 + \beta_6 R_3 \\
\beta_{2.} &= \beta_7 F_1 R_1 + \beta_8 F_1 R_2 + \beta_9 F_1 R_3 + \beta_{10} F_2 R_1 + \beta_{11} F_2 R_2 + \beta_{12} F_2 R_3 \\
&\quad + \beta_{13} F_3 R_1 + \beta_{14} F_3 R_2 + \beta_{15} F_3 R_3 \\
\beta_{3i} &= \beta_{16} + b_{1i} \\
\beta_{4.} &= \beta_{17} F_1 + \beta_{18} F_2 + \beta_{19} F_3 \\
\beta_{5.} &= \beta_{20} R_1 + \beta_{21} R_2 + \beta_{22} R_3 \\
\beta_{6.} &= \beta_{23} F_1 R_1 + \beta_{24} F_1 R_2 + \beta_{25} F_1 R_3 + \beta_{26} F_2 R_1 + \beta_{27} F_2 R_2 + \beta_{28} F_2 R_3 \\
&\quad + \beta_{29} F_3 R_1 + \beta_{30} F_3 R_2 + \beta_{31} F_3 R_3,
\end{aligned}$$

donde se tiene las variables indicadoras F_1, F_2, F_3 y R_1, R_2, R_3 las cuales expresan los niveles de los factores Fertilizante y Riego.

$$\begin{cases}
Fertilizante\ 1 & Si \quad F_1 = F_2 = F_3 = 0 \\
Fertilizante\ 2 & Si \quad F_1 = 1\ y\ F_2 = F_3 = 0 \\
Fertilizante\ 3 & Si \quad F_2 = 1\ y\ F_1 = F_3 = 0 \\
Fertilizante\ 4 & Si \quad F_3 = 1\ y\ F_1 = F_2 = 0
\end{cases} \quad (4\text{-}17)$$

$$\begin{cases}
Riego\ 1 & Si \quad R_1 = R_2 = R_3 = 0 \\
Riego\ 2 & Si \quad R_1 = 1\ y\ R_2 = R_3 = 0 \\
Riego\ 3 & Si \quad R_2 = 1\ y\ R_1 = R_3 = 0 \\
Riego\ 4 & Si \quad R_3 = 1\ y\ R_1 = R_2 = 0
\end{cases} \quad (4\text{-}18)$$

- $\mathbf{y}_i$ es el vector que contiene 7 medidas de clorofila para el individuo i, $i = 1, 2, ..., 128$

- β_0 representa la media general de concentración de clorofila.
- β_1, β_2 y β_3 corresponden a los parámetros de regresión de los efectos asociados al considerar los fertilizantes 2, 3, 4.
- β_4, β_5 y β_6 corresponden a los parámetros de regresión de los efectos asociados al considerar los riegos 2, 3, 4.
- $\beta_7, \beta_8, ..., \beta_{31}$ corresponden a los parámetros de regresión de los efectos asociados a las interaciones de las variables F_1, F_2, F_2, R_1, R_2 y R_3
- $\beta_{17}, \beta_{18}, \beta_{19}$ corresponden a los parámetros de regresión de los efectos asociados a las interaciones de las variables F_1, F_2, F_3, con el tiempo
- $\beta_{19}, \beta_{20}, \beta_{21}$ corresponden a los parámetros de regresión de los efectos asociados de las interaciones de las variables R_1, R_2, R_3, con el tiempo
- $\beta_{23}, \beta_{24}, ..., \beta_{19}$ corresponden a los parámetros de regresión de los efectos asociados a las interaciones de las variables $F_1, F_2, F_3, R_1, R_2, R_3$, con el tiempo
- β_{16} corresponde al parámetro de regresión del efecto fijo asociados a la pendiente del tiempo
- b_{0i} corresponde al intercepto aleatorio para el i ésimo individuo.
- b_{1i} corresponde al efecto aleatorio de la pendiente del tiempo.
- b_{0i} y b_{1i} se suponen que distribuyen $N(\mathbf{0}, D)$. Se tomó a D como una estructura de matriz de varianzas y covarianzas no estructurada (Verbeke & Molenberghs 2000) de la forma

$$D = \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix}, \tag{4-19}$$

  donde d_{12} e, por la correlación negativa entre el intercepto y la pendiente aleatoria.
- ϵ_i es vector de los errores aleatorios del individuo i, el cual se supone que distribuye $N(\mathbf{0}, \Sigma_i)$.

El caso en que la matriz de covarianza residual sea definida como $\sigma = \sigma^2 I$, supondría independencia condicional de $\mathbf{b}_i$. Este supuesto es posible que no sea cierto, es decir, que los efectos aleatorios no logren captar la correlación entre las mediciones longitudinales. Por tal caso en este trabajo se evalúan algunas estructuras de varianzas residuales, tales como: estructura de covarianza de simetría compuesta, no estructurada, autorregresivas, entre otras (Verbeke & Molenberghs 2000). Donde la estructura autorregresiva de orden uno fue la más idónea que viene de la forma:
Estructura autorregresiva de orden uno

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 \end{bmatrix} \tag{4-20}$$

Estructura de simetría compuesta

$$\Sigma = \begin{bmatrix} \sigma_1^2 + \sigma^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma^2 \end{bmatrix} \tag{4-21}$$

No estructurada

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix} \tag{4-22}$$

4.3. Diagnóstico de Residuales

Finalmente se analiza el cumplimiento de los supuestos del modelo como se menciona en la sección 3.7.

5 Resultados

En éste capítulo se presenta la aplicación de la metodología propuesta para el ajuste de un modelo mixto con covariable funcional. Primero se hace un análisis descriptivo de la concentración de clorofila y un análisis descriptivo funcional de las firmas espectrales. Posteriormente se ajusten los modelos mixtos con covariable funcional sin usar y usando el ACPF.

5.1. Análisis Exploratorio

La concentración de clorofila en medidas SPAD varia entre 28.6 y 64.3 (SPAD) y su promedio es 45.5 (SPAD), el coeficiente de variación es 14 %. lo que sugiere homogeneidad. Se puede destacar de la tabla **5-1** que, existe una tendencia creciente en el promedio de la concentración de clorofila en el transcurso de las semanas, esto es, en la primera semana se tiene un promedio de 39,18, en la segunda semana un promedio de 41,35 y así sucesivamente en las semanas del 3 al 7 de 45,1, 47,13, 48,67, 47,96 y 49,44 respectivamente. En la tabla **5-1** y en las figuras **5-1** y **5-2** se muestran algunos indicadores descriptivos y la distribución de la concentración de clorofila, desagregado por tipo de fertilizante, riego y semanas de medición. Se destaca que hay tendencia creciente en la evolución temporal de la concentración de clorofila con respecto a las semana de medición. Las plantas que se le aplica un riego de 75 ml(R3) son los que presentan en promedio más concentración de clorofila con 47 frente a los otros niveles de riego.

Tabla 5-1: Medidas descriptivas de la concentración de clorofila (SPAD) según los factores fertilizante, riego y el tiempo (semana 1 a semana 7), D.E = Desviación Estándar y C.V = Coeficiente de Variación

		Media	Mínimo	Máximo	D.E	C.V(%)
Fertilizante	F1	45,2	28,6	59,1	6,5	14
	F2	45,4	31,0	62,0	6,5	14
	F3	46,2	34,1	64,3	6,5	14
	F4	45,0	32,2	60,7	5,9	13
Riego	R1	44,7	31,8	60,7	6,5	15
	R2	45,4	28,6	61,7	6,6	15
	R3	47,0	31,0	64,3	6,7	14
	R4	44,8	32,6	59,6	5,5	12
Semanas	Semana 1	39,2	29,9	53,6	3,3	9
	Semana 2	41,3	28,8	55,6	4,3	10
	Semana 3	45,1	31,0	58,1	5,5	12
	Semana 4	47,1	32,4	60,2	5,6	12
	Semana 5	48,7	32,6	59,9	5,5	11
	Semana 6	48,0	28,6	60,7	5,5	11
	Semana 7	49,4	33,7	64,3	6,8	14
General		45,5	28,6	64,3	6,4	14

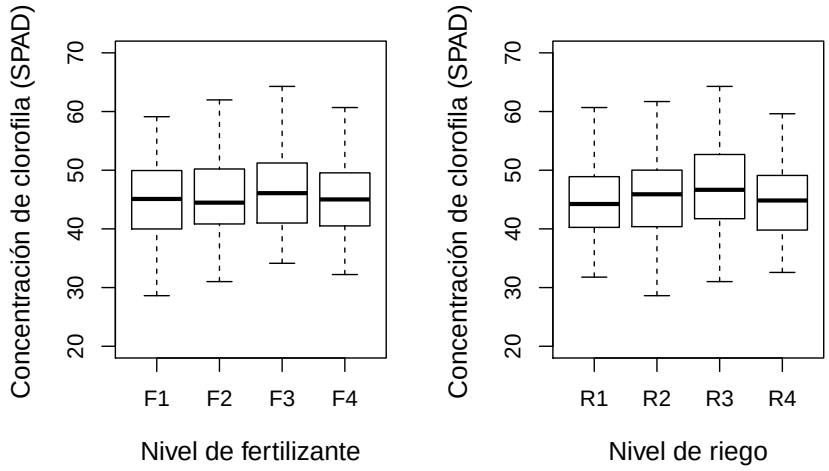

Figura 5-1: Diagrama de cajas de la concentración de clorofila por nivel de fertilizante y riego

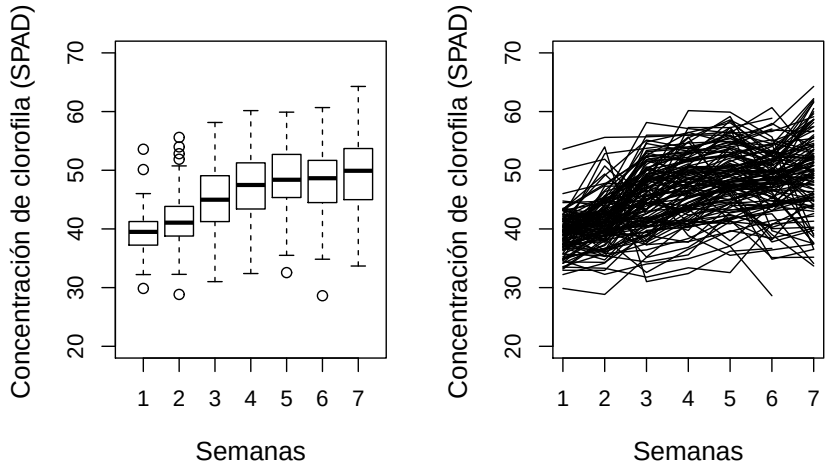

Figura 5-2: Boxplot de la concentración de clorofila por semana de medición (izquierda: diagrama de cajas y derecha: Línea de tendencia de cada planta)

En la figura **5-3** se muestra el diagrama de interacción entre los dos factores considerados. Se puede observar que las trazas de los valores medios de la variable respuesta se cortan en

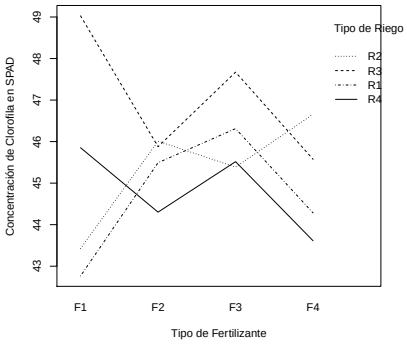

Figura 5-3: Gráfico de interacción entre los factores fertilizante y riego

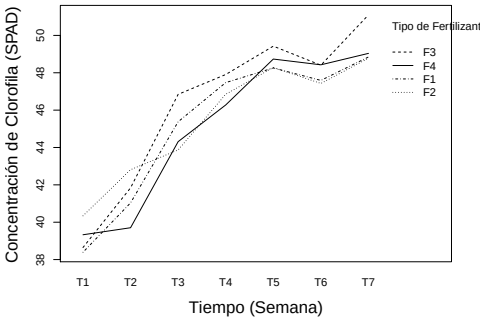

Figura 5-4: Gráfico de interacción entre los factores fertilizante y tiempo

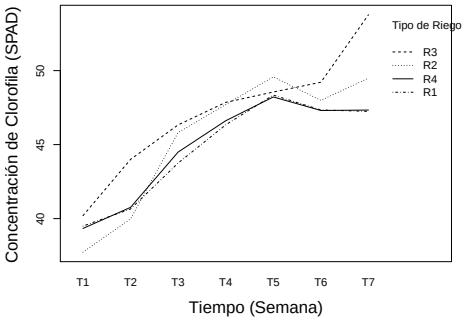

Figura 5-5: Gráfico de interacción entre los factores riego y tiempo

algunos puntos, lo cual sugiere la presencia del efecto de la interacción. Las figuras **5-4** y **5-5** muestran la interacción entre cada factor contra el tiempo. En los dos casos se presenta de nuevo el corte entre lineas, lo cual nuevamente sugiere la presencia de una interacción significativa entre los factores y el tiempo.

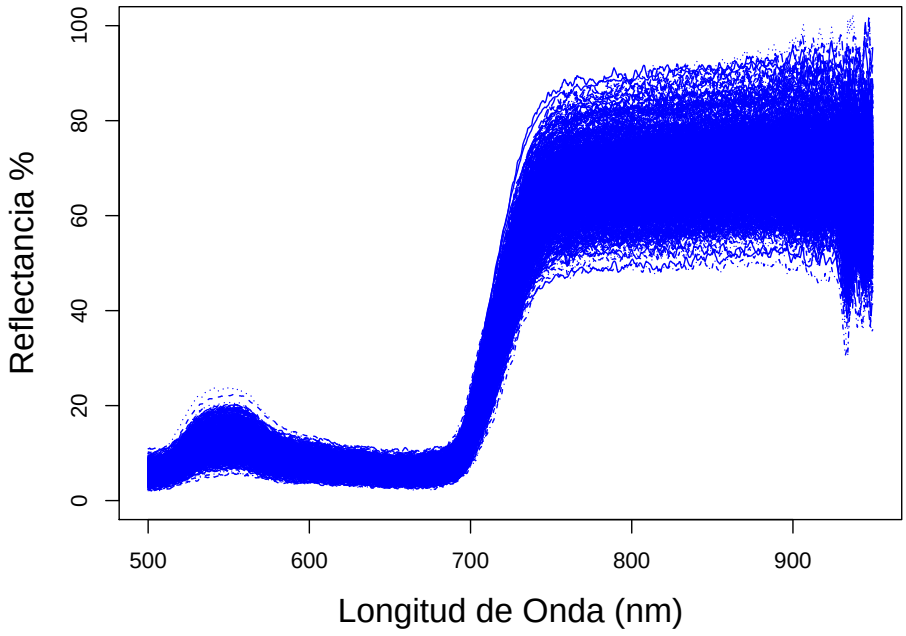

Figura 5-6: Firmas espectrales de las plantas de Ají de Tabasco

En la figura **5-6** se muestra las curvas de las firmas espectrales obtenidas en el experimento. En total son 896 curvas. En figura **5-7** se muestra el análisis descriptivo funcional de la variable funcional firma espectral, en el panel izquierdo se muestra la media y la desviación estándar funcional, donde se destaca que la variabilidad de la curva o firma espectral está en función directa de la media funcional, es decir, a mayor reflectancia media mayor variabilidad. En el panel derecho se muestra la correlación entre cada longitud de onda de la firma espectral con respecto a la concentración de clorofila $Corr(Y, \chi(\lambda))$, en el cual se puede observar que para cualquier longitud de onda la reflectancia se correlaciona de forma negativa con la variable respuesta. En la longitud de onda de 708(nm) se tiene una correlación de -0.59. Lo que indica que en dicho punto la correlación entre la concentración de clorofila y la reflectancia es inversa.

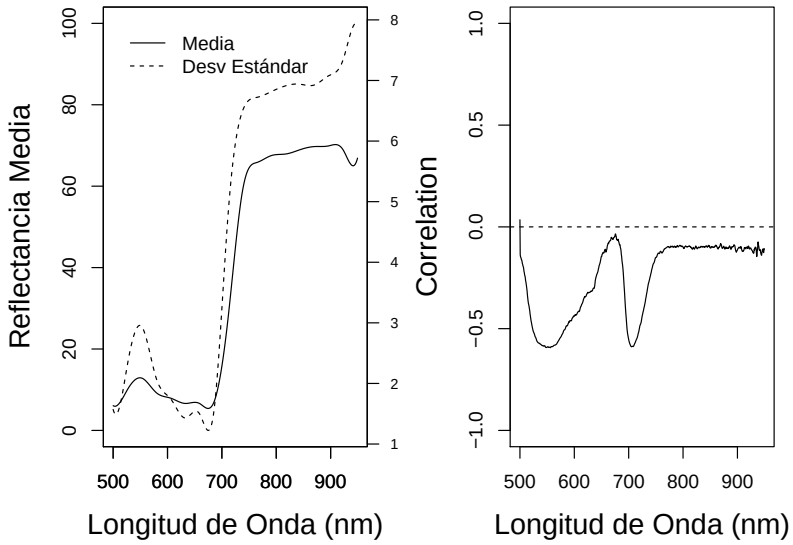

Figura 5-7: Media y desviación estándar funcional de las firmas espectrales (izquierda). Correlación funcional de la firma espectral y la concentración de clorofila (derecha).

5.2. De Datos a Curvas

Para la selección del número de funciones de la base (K_χ) se hizo la Validación Cruzada Generalizada (GVC) variando $K_\chi = 2, 3, ..., 50$. En la figura **5-8** se observa que los valores de GVC decrecen rápidamente, pero no es posible encontrar un mínimo. Entonces se determina un punto de corte en 20 funciones de la bases, pues a partir de aquí se considera despreciable la disminución de la GVC. Tomando $K_\chi = 20$, se suavizan los datos de las firmas espectrales como se muestra en el panel derecho de la figura **5-9**
Las firmas espectrales originales y suavizadas por medio de 20 funciones *B-Spline* se muestran en la figura **5-9**

Inicialmente se supone que la estructura de varianza para los residuales es $= Var(\epsilon_i) = \Sigma_i = \sigma^2 I$ y se a partir del Criterio de Información de Akaike marginal, se determina el número de base de funciones adecuado para describir el parámetro funcional ($\psi(\lambda)$).

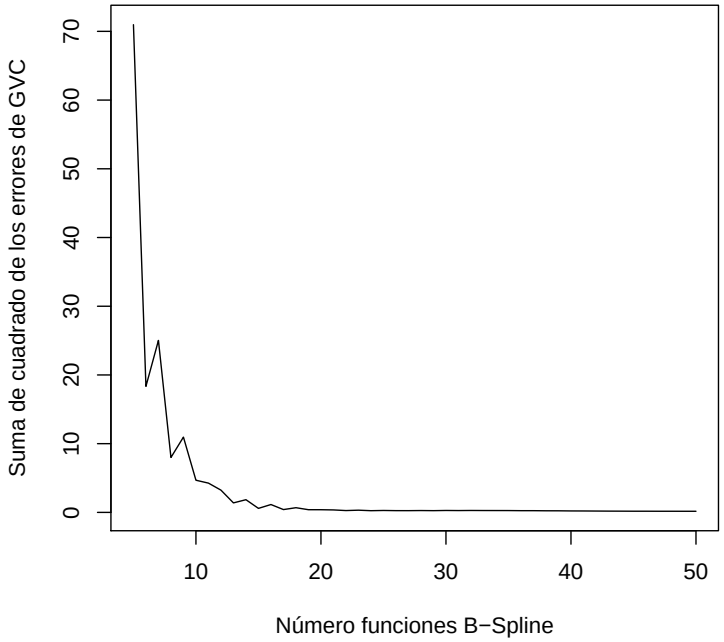

Figura 5-8: Suma de cuadrado del error de Validación Cruzada Generalizada (GVC) según el número de funciones de la base *B-Splines*

Tabla 5-2: Criterio de Información de Akaike marginal contra número de funciones para describir el parámetro funcional

	nb	mAIC
1	6	4914.36
2	8	4786.69
3	10	4786.19
4	12	4788.70
5	14	4796.99
6	16	4807.50
7	18	4814.01
8	20	4822.13

nb = Número de funciones *B-Spline*

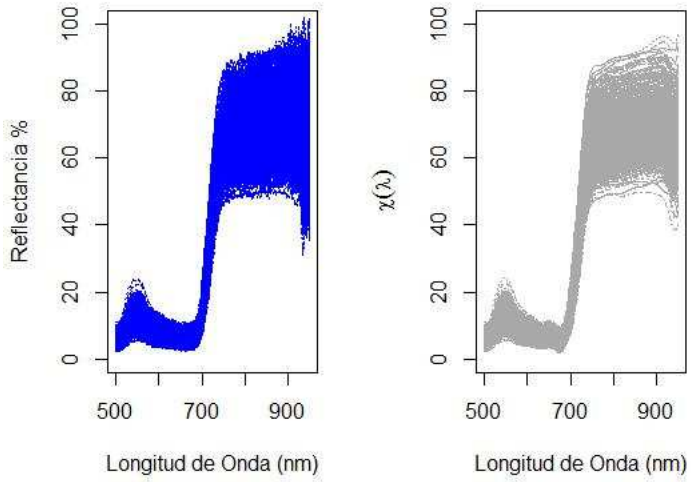

Figura 5-9: Firma espectral original(izquierda) y firma espectral suavizada(derecha) por medio de 20 funciones *B-Spline*

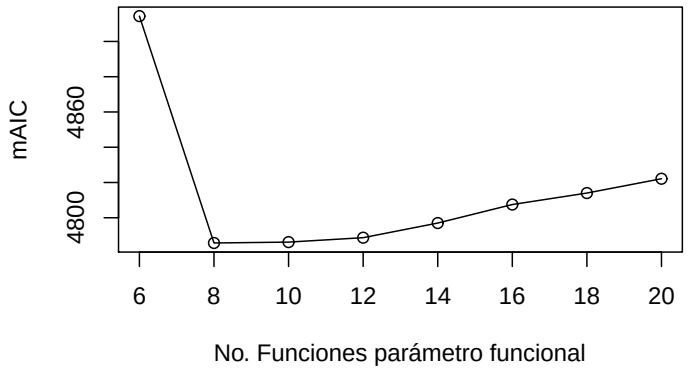

Figura 5-10: Criterio de Información de Akaike contra número de funciones para describir el parámetro funcional

De acuerdo al criterio de mAIC (figura **5-10** y tabla **5-2**) se determina que el número de funciones básicas *B-Spline* necesario para el parámetro funcional es de $\cdot\ K_\psi = 10$.

En la tabla **5-3** se muestra la comparación del ajuste del modelo propuesto usando criterios de información como los de AIC y BIC, además del usó de la prueba de razón de verosimilitud. Con base a la covariable funcional descrita con $K_\psi = 10$ se evalua la elección la estructuras de covarianzas de $\Sigma_i = Var(\epsilon_i) = Cov(\mathbf{y}_i|\mathbf{b}_i)$ como se muestra en la tabla **5-3**.

Tabla 5-3: Elección de la estructura de covarianza del los errores

Modelo	**df**	**AIC**	**BIC**	**Razón de Verosimilitud**
Identidad	46	4790.106	5007.623	-2349.053
Simetría Compuesta	47	4792.106	5014.352	-2349.053
AR(1)	47	4768.117	4990.362	-2337.058
MA(1)	47	4772.330	4994.576	-2339.165

Donde se puede observar que la matriz autorregresiva de orden 1, es la que mejor modela la estructura de covarianza residual. Esto teniendo en cuenta que es la que presenta el menor valor en los criterios de información.

Tabla 5-4: Comparación de modelos incluyendo y excluyendo la pendiente aleatoria

Modelo	**df**	**AIC**	**BIC**	**Razón de Verosimilitud**
Modelo con b_{0i}	45	4780.533	4993.322	
Modelo con b_{0i}, b_{1i}	47	4768.117	4990.362	Valor p $<$ 0,0001

En la tabla **5-4** se compararon los modelos de intercepto aleatorio y de coeficientes aleatorios, ambos asumiendo que la matriz de covarianza de los errores tiene una estructura AR(1). Teniendo en cuenta los criterios de información y la prueba de razón de verosimilitud se llega a la conclusión que el modelo de coeficientes aleatorios, es decir, con intercepto y pendiente aleatoria es el más adecuado.

Tabla 5-5: Comparación de modelos incluyendo y excluyendo la covariable funcional

	Model	df	AIC	BIC
Modelo Sin $\chi(t)$	1	37	4948.224	5123.623
Modelo Con $\chi(t)$	2	47	4768.117	4990.362

En la tabla **5-5** se compara los modelos ajustados con matriz autorregresiva de orden (1) incluyendo el parámetro funcional y sin él; se observa que el número de grados de libertad

adquiridos cuando se incluye la covariable funcional al modelo es de 10, pese a esto, los criterios de información AIC y BIC concuerdan en que el mejor modelo es cuando se considera la covariable funcional.

Tabla 5-6: ANOVA del modelo que incluye covariable funcional

	glN	glD	valor F	Valor p
(Intercept)	1	724	44218.23	0.00
Fertilizante	3	112	1.22	0.30
Riego	3	112	4.83	0.00
Tiempo	1	724	543.51	0.00
Fertilizante:Riego	9	112	0.87	0.56
Fertilizante:Tiempo	3	724	2.72	0.04
Riego:Tiempo	3	724	6.62	0.00
Fertilizante:Riego:Tiempo	9	724.00	0.63	0.77

Acorde al análisis de varianza o ANOVA presentada en la tabla **5-6**, se concluye que la interacción entre el riego y el tiempo es estadisticamente significativas a un $\alpha = 0{,}05$. En cuanto a las interacciones Fertiliza:Riego y Fertiliza:Riego:Tiempo no resultaron ser significativas para explicar la concentración de clorofila, por otro lado las interacciones Fertiliza:Tiempo y Riego:Tiempo si aportan al modelo con un nivel de significancia de 0,05

$$\hat{D} = \begin{bmatrix} & Intercepto & Tiempo \\ Intercepto & 0{,}40898 & -0{,}12305 \\ Tiempo & -0{,}12305 & 0{,}15561 \end{bmatrix} \tag{5-1}$$

En la matriz 5-1 se muestra la matriz de varianzas y covarianzas estimada de los efectos aleatorios $\hat{D}$, la cual fue estimada suponiendo que la matriz de varianzas y covarianzas presenta una forma no estructurada como se menciono en la metodología, la desviación estándar estimada de los errores es de 3,27 y el parámetro estimado correspondiente a la estructura autorregresiva de orden 1 es $\rho = 0{,}276$. A partir del modelo estimado, se tiene en la figura **5-11** la estimación del parámetro funcional.

Las bandas de confianza para el parámetro funcional fueron obtenidas por medio del método Bootstrap que se referencia en la metodología, en el cual se decide usar $B = 1{,}000$ corridas o remuetras bootstrap. En la figura **5-12** se presenta dos tipos de bandas de confianzas obtenidas a partir de las mismas remuestras bootstrap; la primera son todos los parámetros

Tabla 5-7: Ajuste de modelo mixto con covariable funcional

	Valores	Error estándar	gl	valor t	valor p
(Intercept)	39,59	1,85	724	21,44	0,00
Fertilizante22	3,36	1,65	112	2,03	0,04
Fertilizante23	-0,44	1,66	112	-0,26	0,79
Fertilizante24	1,42	1,65	112	0,86	0,39
Riego22	-1,87	1,66	112	-1,12	0,26
Riego23	0,87	1,68	112	0,52	0,60
Riego24	1,03	1,65	112	0,63	0,53
Tiempo	0,46	0,29	724	1,58	0,12
Fertilizante22:Riego22	-1,21	2,34	112	-0,52	0,61
Fertilizante23:Riego22	2,55	2,36	112	1,08	0,28
Fertilizante24:Riego22	1,40	2,35	112	0,60	0,55
Fertilizante22:Riego23	-1,53	2,36	112	-0,65	0,52
Fertilizante23:Riego23	0,65	2,37	112	0,28	0,78
Fertilizante24:Riego23	0,60	2,36	112	0,25	0,80
Fertilizante22:Riego24	-2,66	2,34	112	-1,14	0,26
Fertilizante23:Riego24	-0,70	2,34	112	-0,30	0,77
Fertilizante24:Riego24	-0,29	2,34	112	-0,13	0,90
Fertilizante22:Tiempo	-0,24	0,41	724	-0,58	0,56
Fertilizante23:Tiempo	0,48	0,41	724	1,17	0,24
Fertilizante24:Tiempo	0,03	0,41	724	0,07	0,95
Riego22:Tiempo	0,83	0,41	724	2,01	0,05
Riego23:Tiempo	1,24	0,41	724	3,02	0,00
Riego24:Tiempo	0,37	0,41	724	0,90	0,37

Tabla 5-8: Ajuste de modelo mixto con covariable funcional

	Valores	Error estándar	gl	valor t	valor p
Fertilizante22:Riego22:Tiempo	-0,02	0,58	724	-0,03	0,97
Fertilizante23:Riego22:Tiempo	-0,75	0,59	724	-1,29	0,20
Fertilizante24:Riego22:Tiempo	-0,16	0,58	724	-0,28	0,78
Fertilizante22:Riego23:Tiempo	-0,50	0,58	724	-0,86	0,39
Fertilizante23:Riego23:Tiempo	-0,55	0,58	724	-0,94	0,35
Fertilizante24:Riego23:Tiempo	-0,84	0,58	724	-1,44	0,15
Fertilizante22:Riego24:Tiempo	-0,09	0,58	724	-0,16	0,87
Fertilizante23:Riego24:Tiempo	-0,35	0,58	724	-0,61	0,54
Fertilizante24:Riego24:Tiempo	-0,61	0,58	724	-1,06	0,29
*B-Spline*1	-0,05	0,06	724	-0,89	0,37
*B-Spline*2	0,12	0,08	724	1,45	0,15
*B-Spline*3	-0,16	0,06	724	-2,44	0,02
*B-Spline*4	0,15	0,03	724	4,78	0,00
*B-Spline*5	-0,12	0,02	724	-5,58	0,00
*B-Spline*6	0,07	0,02	724	3,30	0,00
*B-Spline*7	-0,04	0,03	724	-1,62	0,11
*B-Spline*8	0,03	0,02	724	1,13	0,26
*B-Spline*9	-0,02	0,02	724	-0,99	0,32
*B-Spline*10	0,01	0,01	724	0,80	0,43

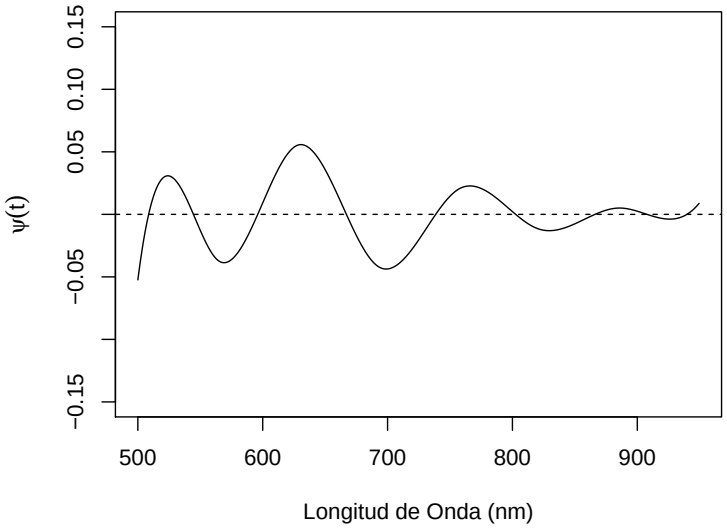

Figura 5-11: Parámetro funcional estimado

funcionales bootstrap que cumplen $\left\|\widehat{\psi}(t) - \widehat{\psi}(t)^j\right\| \leq \hat{D}_\alpha$, donde $\hat{D}_\alpha$ corresponde al percentil $1-\alpha$ del estadístico $d^j = \left\|\widehat{\psi}(t) - \widehat{\psi}(t)^j\right\|$, por otro lado en el panel derecho se muestra la banda de confianza bootstrap obtenida considerando los percentiles $\alpha/2$ y $1-\alpha/2$ de los 1,000 parámetros funcionales tomándolos punto a punto, es decir, para cada valor de de longitud de onda se calcula los percentiles para los 1,000 parámetros funcionales, en ambos casos se toma un nivel de confianza del 95 %.
Se destaca que las bandas de confianza en ambos casos no contienen siempre al valor cero, lo cual indica que el parámetro funcional es significativamente diferente de cero, tampoco ronda alrededor de un valor constante, en el caso en el cual las bandas de confianzas estén alrededor de un valor constante $\psi(t) = c$, entonces $\int_T \psi(t)\chi(t)dt = \int_T c\chi(t)dt = c\int_T \chi(t)dt$, lo cual quiere decir, que que debe de proyectar a la covariable funcional con su integral y trabajarlo como un dato escalar.

5.3. Modelo Basado en ACPF

Si se analiza la correlación entre las columnas de la matriz de diseño de los efectos fijos de la propuesta 1, se tiene que existe correlaciones casi practicamente iguales a 1 como la

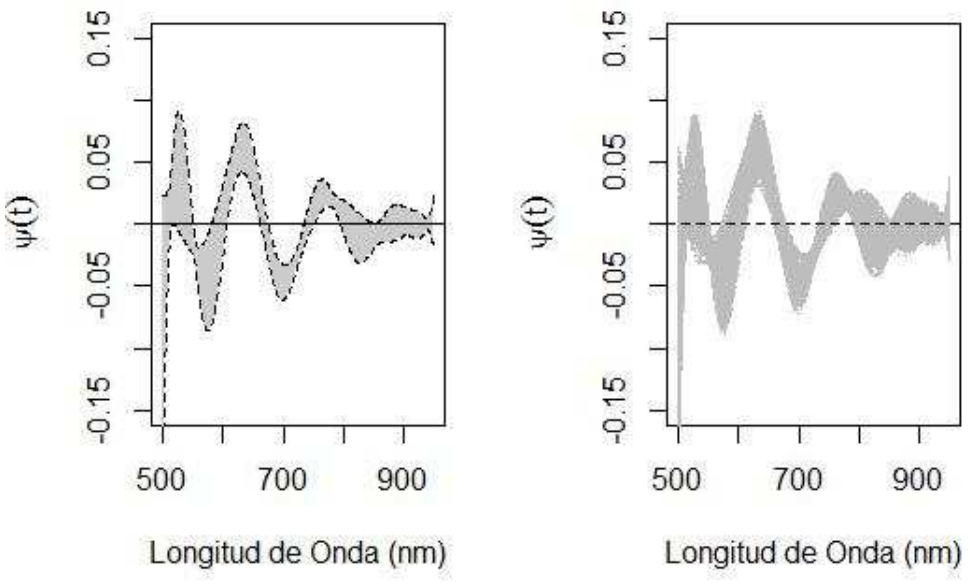

Figura 5-12: Banda de confianza basadas en método Bootstrap. Usando área de confianza AC (izquierda) y usando percentiles punto a punto (derecha).

muestra la figura **5-13**.

En las figuras **5-16** y **5-17** se muestra las primeras 4 funciones propias $\xi_1(t), \xi_2(t), \xi_3(t), \xi_4(t)$ obtenidos a partir del conjunto de datos funcionales centrados **5-14**. Se puede destacar que la primera función propia del conjunto de datos funcionales que acumula el 91,4 % de la variabilidad presenta una forma funcional similar a la que se ilustra en el gráfico descriptivo de la desviación estándar funcional **5-7**. Esto es razonable, pues las componentes van en dirección a la mayor variablidad del conjunto de datos.

Se considera los 4 primeras componentes principales, pues el porcentaje de variabilidad explicada por estos es 99,3 %. El primer componente explica aproximadamente el 91,4 %, mientras que los componentes: 2, 3 y 4 explican conjuntamente 7,8 % de la variabilidad.

Como se muestra en la tabla **5-9**, donde se comparan los tres modelos planteados, el primero no contempla el parámetro funcional, en el segundo se tiene en cuenta la covariable

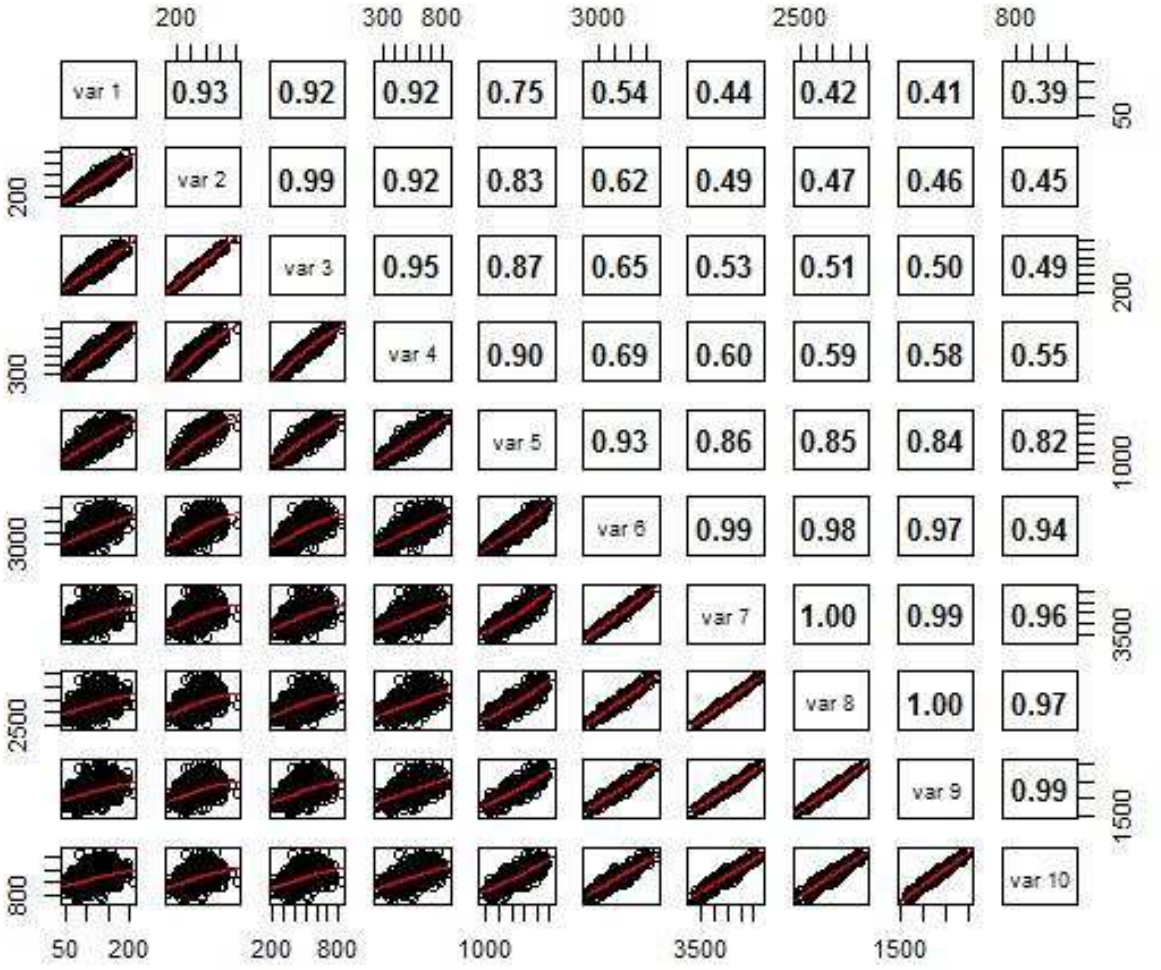

Figura 5-13: Matriz de correlacción entre columnas de la matriz de diseño

funcional, donde el parámetro funcional es descrito por medio de una base de *B-Spline* de 20 funciones y un tercer modelo que también contempla la covariable funcional, pero en esta aproximación se realiza por medio de ACPF, el cual soluciona tanto el problema de multicolinealidad presente en la matriz de diseño y reduce la dimensionalidad del modelo. La comparación de estos tres modelos se realiza por medio de los criterios de información AIC y BIC. Se destaca que los dos modelos donde se incluye la covariable funcional presentan valores AIC Y BIC inferiores al primer modelo, lo cual quiere decir que la inclusión de

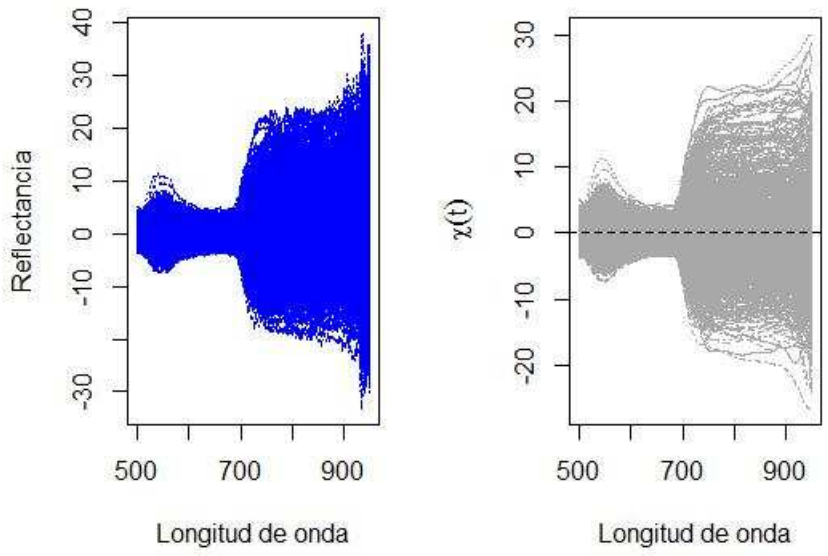

Figura 5-14: Firmas espectrales centradas (izquierda) y firmas espectrales centradas y suavizadas usando una base de funciones (tamaño 20) de *B-splines* (derecha)

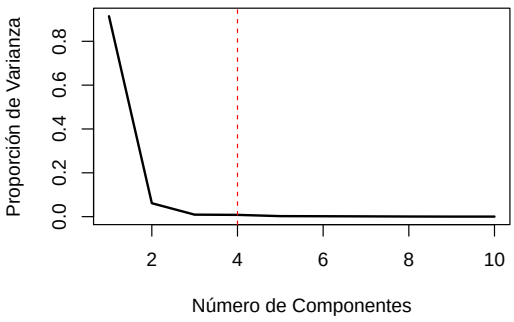

Figura 5-15: Cambio de los valores propios respecto al número de componentes considerados

esta variable si es relevante para explicar la concentración de clorofila en el experimento planteado. Por otro lado, dichos criterios permiten concluir también que entre el balance del poder explicativo y parsimonia entre los modelos 2 y 3; el modelo con ACPF es mejor

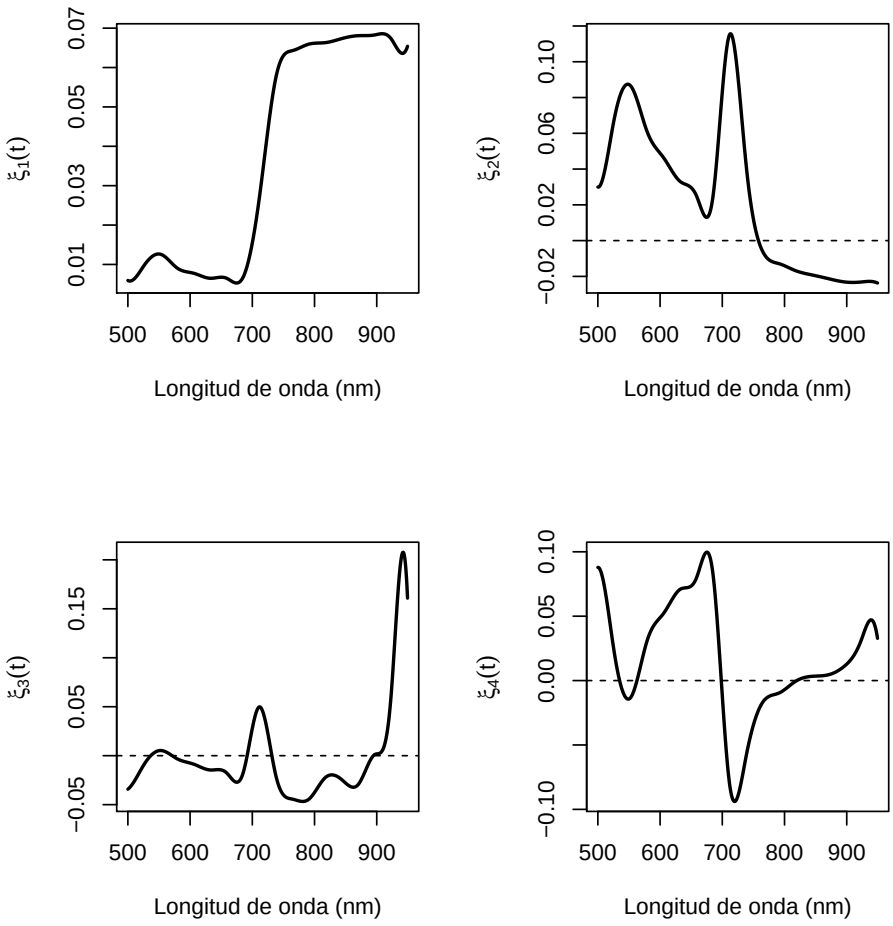

Figura 5-16: Primeros 4 funciones propias de las firmas espectrales centradas y suavizadas usando una base de funciones (tamaño 20) de *B-splines*

Tabla 5-9: Comparación de los criterios AIC y BIC de los modelos incluyendo y excluyendo la covariable funcional.

	Modelo	gl	AIC	BIC
Modelo Sin $\chi(t)$	1	37	4948.224	5123.623
Modelo Con $\chi(t)$	2	47	4768.117	4990.362
Modelo Con $\chi(t)$ ACPF	3	41	4751.115	4945.282

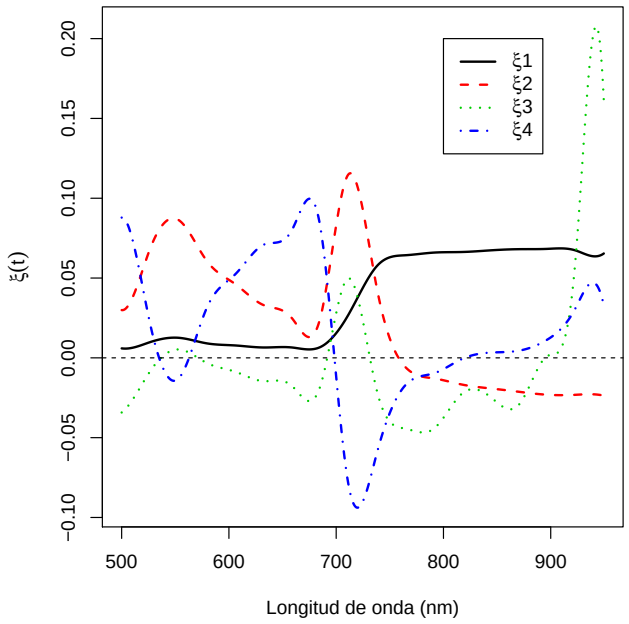

Figura 5-17: Primeros 4 funciones propias de las firmas espectrales centradas y suavizadas usando una base de funciones (tamaño 20) de *B-splines*

es el que mejor describe la concentración de clorofila.

Tabla 5-10: ANOVA del modelo con ACPF

	glN	glD	valor F	Valor p
(Intercept)	1	730.00	40590.04	0.00
Fertilizante	3	112.00	1.11	0.35
Riego	3	112.00	4.42	0.01
Tiempo	1	730.00	506.69	0.00
Fertilizante:Riego	9	112.00	0.88	0.55
Fertilizante:Tiempo	3	730.00	2.13	0.09
Riego:Tiempo	3	730.00	6.16	0.00
Fertilizante:Riego:Tiempo	9	730.00	0.69	0.72

De acuerdo con el ANOVA (tabla **5-10**), se concluye que la interacción ente riego y tiempo

es estadísticamente significativa.

En la siguiente figura **5-18** se presenta el parámetro funcional para el modelo que considera la descomposición de los datos funcionales en 4 componentes principales.

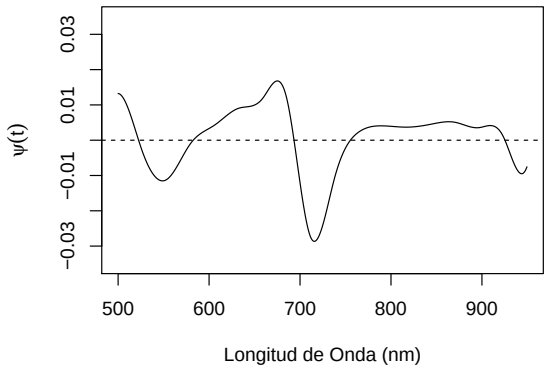

Figura 5-18: Estimación del parámetro funcional del modelo basado en el ACPF

En la figura **5-19** nuevamente se presentan el resultado de las bandas de confianza bootstrap para el modelo ACPF, en los cuales se puede observar que el parametro funcional es diferente a un valor constante de cero o a cualquier valor constante e indica que la inclusión de la covariable funcional es significativa a un $\alpha = 0{,}05$ para la modelación de la concentración de clorofila.

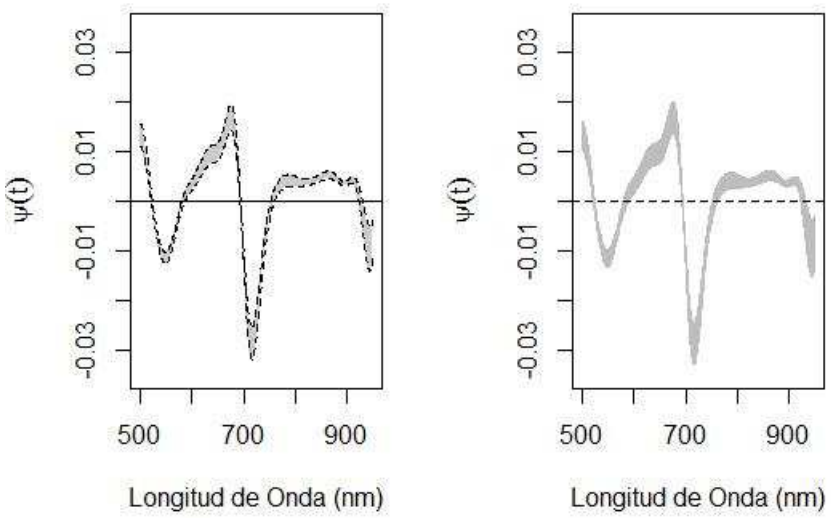

Figura 5-19: Banda de confianza basadas en método Bootstrap del modelo basado en el ACPF. Usando área de confianza AC (izquierda) y usando percentiles punto a punto (derecha).

En los ANEXOS se presentan los gráficos de validación de los supuestos de los modelos propuestos. Donde los gráficos de cuantil cuantil para el modelo 4-3 y el modelo que incluye como covariable los scores del ACPF (figuras 7-2 y 7-6 respectivamente) se muestran colas pesadas en la distribución de los residuales de los modelos, lo cual sugiere la modelación por medio de distribuciones de probabilidad que consideren colas más pesadas que la distribución normal. Las figuras 7-3 y 7-7 se muestra el gráfico de los residuales estandarizados contra valores ajustados de los modelos 4-3 y 4-11, el objetivo de estas gráfica son las de evaluar de manera visual la homocedásticidad en los modelos, se encontró que tanto en 4-3 como en 4-11 no se invalida este supuesto. Por último las figuras 7-4 y 7-8 por medio de la función de autocorrelación no se presenta visualmente correlación residual.

6 Conclusiones y trabajos futuros

6.1. Conclusiones

El objetivo de este trabajo fue modelar la concentración de clorofila en unidades spad en plantas de Ají de Tabasco bajo factores de estrés causados por el tipo de fertilizante y el nivel de riego aplicado. Adicional a estos factores se tuvo en cuenta la firma espectral, como una covariable funcional. La evaluación de la significancia de estos factores se realizó por medio del ajuste de un modelo lineal mixto con covariable funcional bajo dos enfoques: El primero usando sumas de funciones de una base en este caso B-*Splines* y el segundo enfoque por medio de análisis de componentes principales funcionales. A diferencia del model propuesto por Goldsmith et al. (2012), las alternativas de modelación que se presentaron en este trabajo de investigación contemplan la modelación de la matriz de varianzas y covarianzas de los residuales y no asumiendo que esta tiene una estructura identidad. Acorde al análisis de varianza o ANOVA presentada en la tabla **5-6**, se concluye que la interacción entre el riego y el tiempo es estadisticamente significativas a un $\alpha = 0{,}05$. En cuanto a las interacciones Fertiliza:Riego y Fertiliza:Riego:Tiempo no resultaron ser significativas para explicar la concentración de clorofila, por otro lado las interacciones Fertiliza:Tiempo y Riego:Tiempo si aportaron al modelo planteado. Por medio de la prueba de Razón de Verosimilitud se determinó la importancia de incluir la pendiente aleatoria en el modelo, lo cual índica que la diferencias entre la tasa de crecimiento lineal de la concentración de clorofila en las diferentes plantas. En término general, los modelos planteados fueron idóneos para la modelación de la variable de la concentración relacionándola con la firma espectral en el ámbito de datos longitudinales.

La teledetección como herramienta para la caracterización de cultivos, se fundamenta en la construcción de índices (índices de vegetación) basados en la observación de las firmas espectrales, entre los índices más conocidos se tienen CTR1 y CTR2 de Carter (1994), el índice modificado Red Edge Ratio simple (mSR705), planteado por Gamon & Sims (2002), los índices VOG1, VOG2, y VOG3, propuestos por Vogelmann et al. (1993), el índice NDVI705, y mNDVI705, planteado por Gitelson & Merzlyak (1994) y el índice CARI de

Kim et al. (1994) entre otros. A diferencia de estos índices por medio del modelo planteado se pudo analizar la firma espectral de forma completa. Así mismo se propone una alternativa a la dada por Goldsmith et al. (2012) para la modelación de la matriz de varianzas y covarianza residual, la cual puede modelar la dependencia en datos longitudinales (Verbeke & Molenberghs 2000). En este caso de estudio se encontró que la estructura de dependencia autorregresiva de orden uno es la más adecuada para la matriz de varianzas y covarianzas residual.

En esta investigación se se propone la metodología para usar un ACPF con el fin de resumir la información de las curvas espectrales y facilitar así mismo su inclusión en el modelo mixto. Esta metodología evita tratar con el problema de alta dimensionalidad y colinealidad (Aguilera et al. 2006).

Para evaluar la significancia del parámetro funcional se desarrollan bandas de confianza, estimadas combinando los métodos de Bootstrap completamente paramétrico en Modelos Mixtos (Lahiri 2003) enfocado a la covariable funcional, donde se toma como referencia los procedimientos descritos por Febrero-Bande et al. (2010) y Febrero-Bande & Oviedo de la Fuente (2012). Estas bandas de confianza ayudaron de forma visual a comprobar la significancia del parámetro funcional.

Los modelos planteados en este trabajo de investigación se considera como un gran aporte, pues por un lado generaliza a los modelos mixtos (Verbeke & Molenberghs 2000) al poder incluir una o más covariables funcionales y por otro lado generaliza también a los modelos de regresión funcional con respuesta escalar (Ramsay & Dalzell 1991, Frank & Friedman 1993, Cardot et al. 1999), pues estos trabajaban bajo el supuesto de independencia entre observaciones, lo cual no permite el análisis de datos longitudinales.

6.2. Trabajos futuros

Este trabajo de investigación abren la posibilidad de nuevos estudios e incluso posiblemente algunas nuevas lineas de investigación, entre estas se resaltan:

- La estimación de las funciones propias a partir de mínimos cuadrados parciales (PLS) funcionales, pues éste método garantiza que la correlación entre las funciones propias sea máxima con respecto a la variable respuesta.
- Los gráficos cuartil cuartil sugiere la posibilidad de considerar otros modelos distintos

al normal. Goldsmith et al. (2012) ha propuesto en este contexto los modelos lineales generalizados mixtos funcionales.

7 ANEXO

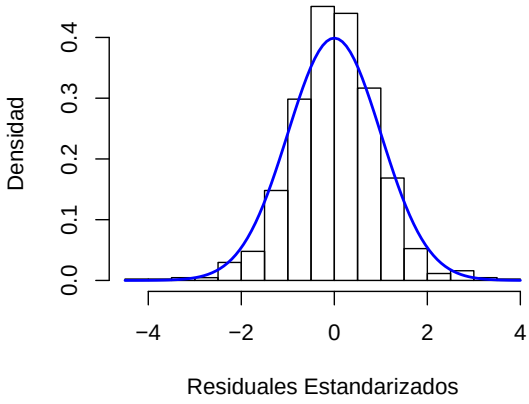

Figura 7-1: Distribución de los residuales estandarizados del modelo ajustado 4-3.

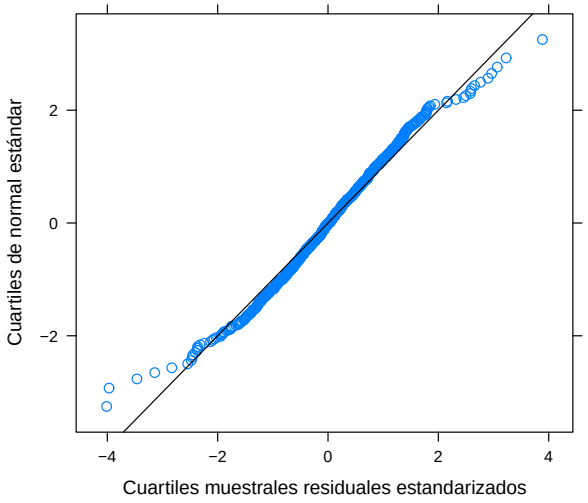

Figura 7-2: Gráfico cuantil cuantil de los residuales estandarizados del modelo ajustado 4-3.

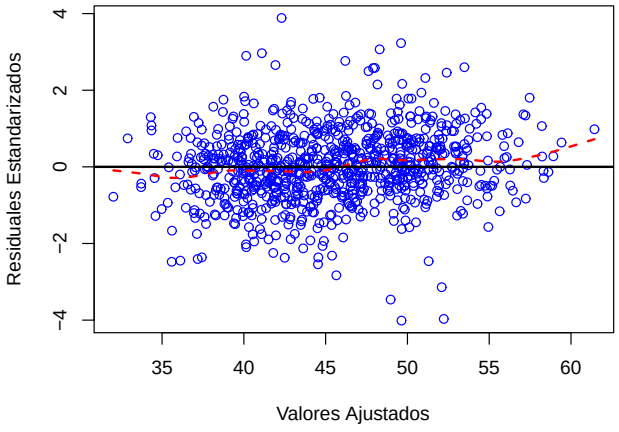

Figura 7-3: Residuales estandarizados contra valores ajustados en el modelo 4-3.

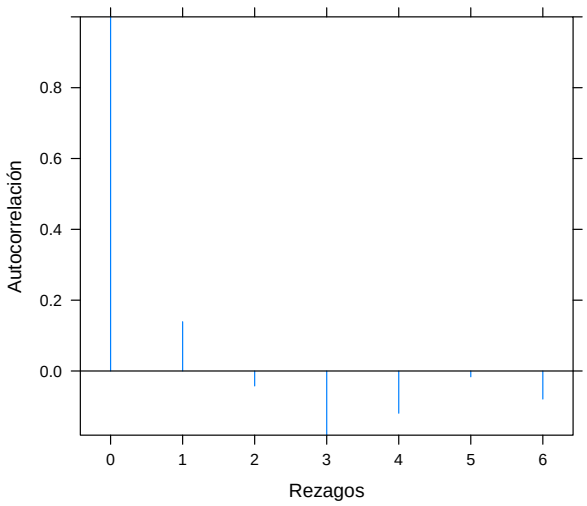

Figura 7-4: Función de autocorrelación simple (FAS) de los residuales del modelo 4-3.

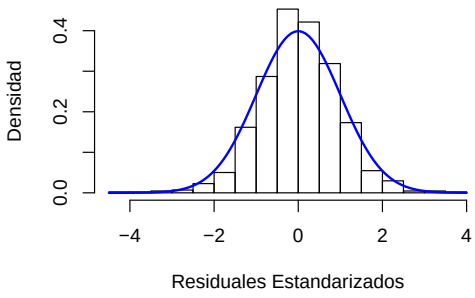

Figura 7-5: Distribución de los residuales estandarizados del modelo que incluye como covariable los scores del ACPf.

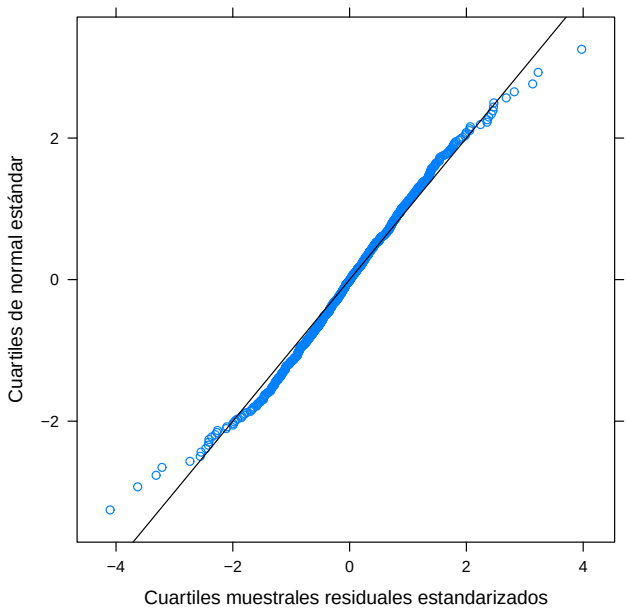

Figura 7-6: Gráfico cuantil cuantil de los residuales estandarizados del modelo que incluye como covariable los scores del ACPf.

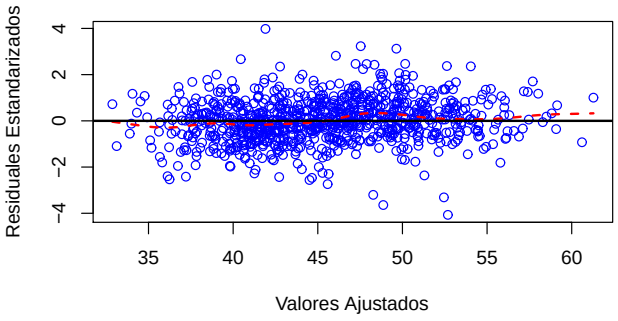

Figura 7-7: Residuales estandarizados contra valores ajustados en el modelo que incluye como covariable los scores del ACPf.

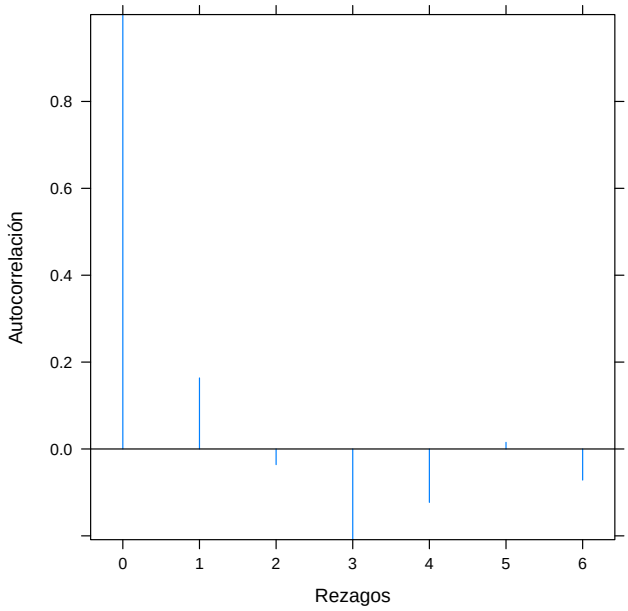

Figura 7-8: Función de autocorrelación simple (FAS) de los residuales del modelo que incluye como covariable los scores del ACPf.

Bibliografía

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.

Botero, J., Parra, L. & Cabrera, K. (2009), 'Determinación del nivel de nutrición foliar en banano por espectrometría de reflectancia', *Revista Facultad Nacional de Agronomía, Medellín* **62**(2), 5089–5098.

Cardot, H., Ferraty, F. & Sarda, P. (1999), 'Functional linear model', *Statistics & Probability Letters* **45**(1), 11–22.

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

Carter (1994), 'Ratios of leaf reflectance in narrow wavebands as indicators of plants stress', *International Journal of Remote Sensing* **15**(3), 697–703.

Chuvieco, E. (2007), *Teledetección Audiovisual: La Observación de la Tierra desde el Espacio*, Barcelona: Ariel.

Crainiceanu, C. M., Staicu, A. M. & Di, C. Z. (2009), 'Generalized multilevel functional regression', *Journal of the American Statistical Association* **104**(488), 1550–1561.

De Boor, C. (1977), 'Package for calculating with b-splines', *SIAM Journal on Numerical Analysis* **14**(3), 441–472.

De la Cruz-Durán, J., Sánchez-García, P., Galvis-Spínola, A. & Carrillo-Salazar, J. (2011), 'Índices espectrales en pimiento para el diagnóstico nutrimental de nitrógeno', *Terra Latinoamericana* **29**(3), 259–265.

Escabias, M., Aguilera, A. M. & Valderrama, M. J. (2006), 'Functional pls logit regression model', *Computational Statistics & Data Analysis* **51**(10), 4891–4902.

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.

Ferraty, F. & Vieu, P. (2002), 'The functional nonparametric model and application to spectrometric data', *Computational Statistics* **17**(4), 545–564.

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

Frank, l. & Friedman, J. H. (1993), 'A statistical view of some chemometrics regression tools', *Technometrics* **35**(2), 109–135.

Gamon, J. & Sims, D. (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), 337–354.

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 Plant Physiology* **143**, 286–292.

Goldsmith, J., Crainiceanu, C. M., Caffo, B. & Reich, D. (2012), 'Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **61**(3), 453–469.

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

Haar, A. (1910), 'Zur theorie der orthogonalen funktionensysteme', *Mathematische Annalen* **69**(3), 331–371.

Horváth, L. & Kokoszka, P. (2012), *Inference for Functional Data with Applications*, Springer Science & Business Media.

Jensen, J. (2005), *It Introductory Digital Image Processing: A Remote Sensing Perspective*, Pearson Prentice Hall.

Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and their Applications*, Springer Science & Business Media.

Kim, M., Daughtry, C., Chappelle, E., Mcmurtrey, J. & Walthall, C. (1994), 'The use of high spectral resolution bands for estimating absorbed photosynthetically active radiation (apar)', *In: Proceedings of the Sixth Symposium on Physical Measurements and Signatures in Remote Sensing, Val D?Isere, France January 17?21* p. 299?306.

Lahiri, S. N. (2003), *Resampling Methods for Dependent Data*, Springer-Verlag; New York.

Loeve, M. (1997), *Probability Theory I*, Springer-Verlag; New York.

Marx, B. D. & Eilers, P. H. (1999), 'Generalized linear regression on sampled signals and curves: a p-spline approach', *Technometrics* **41**(1), 1–13.

McCulloch, P. & Nelder, J. (1989), *Generalized linear Models*, 2 edn, London: Chapmanan and Hall.

Mousavi, S. N. (2015), Analysis of Functional Data with Focus on Multinomial Regression and Multilevel Data, PhD thesis, Department of Mathematical Sciences, Faculty of Science, University of Copenhagen.

Murrillo, P. & Carbonell, J. (2012), Principios y aplicaciones de la percepción remota en el cultivo de la caña de azúcar en colombia, Technical report, CENICAÑA.

Ocampo, J. (2015), Evaluación de la respuesta espectral en plantas de ají tabasco bajo diferentes condiciones de riego y fertilización, Master's thesis, Universidad del Valle.

Ordóñez, C., Martínez, J., Matías, J. M., Reyes, A. & Rodríguez-Pérez, J. R. (2010), 'Functional statistical techniques applied to vine leaf water content determination', *Mathematical and Computer Modelling* **52**(7), 1116–1122.

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

Preda, C., Saporta, G. & Lévéder, C. (2007), 'Pls classification of functional data', *Computational Statistics* **22**(2), 223–235.

Ramsay, J. O. & Dalzell, C. (1991), 'Some tools for functional data analysis', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 539–572.

Ramsay, J. & Silverman, B. (1997), *Functional Data Analysis*, Springer-Verlag.

Ramsay, J. & Silverman, B. W. (2005), *Functional Data Analysis*, Springer-Verlag; New York.

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.

Warner, T. A., Foody, G. M. & Nellis, M. D. (2009), *The SAGE handbook of remote sensing*, 2nd edn, Sage Publications.

Warren, J., Ratnasekera, T., Campbell, D. & Anderson, G. (2017), 'Spectral signatures of immature lucilia sericata (meigen) (diptera: Calliphoridae)', *Insects* **8**(2), 34–54.