ENFOQUE BAYESIANO DE UN MODELO SEMIPARAMÉTRICO MIXTO CON DISTRIBUCIÓN BETA

Por

Liz del Rosario Teran Herrera

Tesis sometida en cumplimiento parcial de los requisitos para el grado de:

MAESTRÍA EN CIENCIAS

en

MATEMÁTICAS (ESTADÍSTICA)

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ

Mayo, 2017

Aprobada por:

Pedro A. Torres Saavedra, Ph.D Presidente, Comité Graduado	Fecha
Raúl E. Macchiavelli, Ph.D Miembro, Comité Graduado	Fecha
Dámaris Santana Morant, Ph.D Miembro, Comité Graduado	Fecha
José R. Ferrer López, Ed.D Representante de Estudios Graduados	Fecha
Olgamary Rivera Marrero, Ph.D Directora del Departamento	Fecha

Resumen de Disertación Presentado Escuela Graduada
de la Universidad de Puerto Rico como requisito parcial de los
Requerimientos para el grado de Maestría en Ciencias

ENFOQUE BAYESIANO DE UN MODELO SEMIPARAMÉTRICO MIXTO CON DISTRIBUCIÓN BETA

Por

Liz del Rosario Teran Herrera

Mayo 2017

Consejero: Pedro A. Torres-Saavedra
Departamento: Ciencias Matemáticas

En este trabajo se propone un modelo semiparamétrico mixto con distribución beta con un enfoque bayesiano usando algoritmos de Monte Carlo vía Cadenas de Markov (MCMC, por sus siglas en inglés). En el modelo propuesto se incorporan técnicas de suavizamiento para el modelamiento de curvas de forma más flexible. Enfoques existentes para modelar variables respuestas que están restringida en el intervalo $(0, 1)$ usan modelos con respuesta normal después de una transformación logit. Este enfoque puede ser ineficiente e inadecuado debido a la distribución asimétrica de la respuesta aún después de la transformación. Un modelo de regresión con distribución beta es una buena alternativa para modelar este tipo de datos debido a la flexibilidad de esta distribución.

El enfoque bayesiano del modelo propuesto ofrece varias ventajas sobre su contraparte frecuentista. Primera, la comparación de curvas entre tratamientos a través del tiempo no necesita ajustes por multiplicidad. Segunda, el enfoque bayesiano es más conveniente en términos computacionales debido a que no usa rutinas de optimización basadas en derivadas tales como Newton-Raphson o gradiente conjugado.

Estas rutinas son particularmente problemáticas en modelos que involucran dimensiones altas. Tercera, a diferencia del enfoque frecuentista que usa métodos para aproximar los errores estándar de funciones de parámetros, por ejemplo, el método delta, en el enfoque bayesiano este cálculo de errores estándar se lleva a cabo fácilmente usando la distribución posterior.

Por medio de simulaciones se muestra que el método propuesto estima los parámetros del modelo ajustado adecuadamente. Las simulaciones e implementación del modelo propuesto se llevaron a cabo en el software bayesiano JAGS vía el paquete R2jags de R usando Cadenas de Markov de Monte Carlo (MCMC).

Finalmente, con base a los resultados obtenidos mediante las simulaciones el método propuesto se aplicó al conjunto de datos reales que corresponde a progreso en la severidad de la Sigatoka negra en cultivos de banano en Isabela, Puerto Rico. El modelo ajustado permite estimar las curvas de progreso de la Sigatoka satisfactoriamente y llevar a cabo cualquier tipo de comparación entre los tratamientos del estudio. En conclusión, el método propuesto en este trabajo es una muy buena alternativa de modelaje de curva de progreso de una enfermedad, cuya respuesta está en el intervalo $(0, 1)$.

Abstract of Dissertation Presented to the Graduate School
of the University of Puerto Rico in Partial Fulfillment of the
Requirements for the Degree of Master of Sciences

BAYESIAN APPROACH FOR A SEMIPARAMETRIC MIXED MODEL WITH BETA DISTRIBUTION

By

Liz del Rosario Teran Herrera

May 2017

Chair: Pedro A. Torres-Saavedra
Major Department: Mathematical Sciences

In this work we propose a semi-parametric mixed model with beta distribution with a Bayesian approach using Monte Carlo algorithms via Markov Chains (MCMC). In the proposed model, smoothing techniques are incorporated for modeling curves with more flexibility. Existing approaches to model variable responses that are constrained in the $(0, 1)$ range use models with normal response after a logit transformation. This approach may be inefficient and inadequate because of the asymmetric distribution of the response even after transformation. A regression model with beta distribution is a good alternative to model this type of data due to the flexibility of this distribution.

The Bayesian approach of the proposed model offers several advantages over its frequentist counterpart. First, the comparison of curves between treatments across time does not need adjustments for multiplicity. Second, the Bayesian approach is more convenient computationally because it does not use optimization routines based on derivatives such as Newton-Raphson or gradient descent. These routines could problematic particularly in models involving high dimensions. Third, unlike

the frequestist approach that uses methods to approximate the standard errors of functions of parameters, for instance, delta method, in the Bayesian approach this calculation of standard errors is easily performed using the posterior distribution.

Using simulations we show that the Bayesian approach estimates the parameters of the proposed model adequately. The simulations and implementation of the proposed model were done in the Bayesian JAGS software via the R2jags package of R using Monte Carlo Markov Chains (MCMC).

Finally, based on the results obtained by simulation, the proposed method was applied to a real dataset of severity of black Sigatoka in banana crops in Isabela, Puerto Rico. The fitted model allowed us to estimate progress curves for Sigatoka satisfactorily and to perform any kind of treatment comparison in the study. In conclusion, the proposed method is a good alternative to model curves of disease progress, since its response variable lies on the interval $(0, 1)$.

Dedico este trabajo a Dios, a mi familia y a ti Alcy que me apoyaste en todo momento, tus palabras de aliento no me dejaron desfallecer.

Agradecimientos

Quiero agradecer al Dr. Pedro A. Torres Saavedra por todas sus enseñanzas, orientación y apoyo durante la elaboración de esta investigación.

Al Dr. Raúl E. Macchiavelli agradezco por proporcionar las ideas iniciales en el problema de investigación, particularmente en el modelo frecuentista realizado en conjunto con Adriana Calvo (Calvo, 2015).

Al Dr. José A. Chavarría Carvajal le agradezco por facilitar el acceso a los datos para la aplicación de la tesis.

Al Centro de Tecnologías de Información por facilitarnos el acceso sus servidores.

A Alcibiades Bustillo por su incondicional estímulo y apoyo en este proyecto.

Al *Departamento de Ciencias Matemáticas* por brindarnos su apoyo para la realización de este proyecto.

A mis amigos gracias por acompañarme en todos estos años, por sus consejos y ayuda.

Los datos de aplicación en este proyecto provinieron del proyecto “Practice for the Control of Black Sigatoka in Puerto Rico” financiado por el Departamento de Agricultura de Puerto Rico a través de la Estación Experimental Agricola (UPRM), proyecto Z-FIDA01.

Este trabajo usó Ciencia Extrema e Ingeniería del Medio Ambiente (XSEDE por sus siglas en inglés) con financiamiento de la Fundación Nacional para la Ciencia con número de concesión ACI-1053575.

Índice general

página

Índice de tablas

Tabla página

Índice de figuras

Figura página

LISTA DE ABREVIATURAS

GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
MCMC	Markov Chain Monte Carlo
JAGS	Just Another Gibbs Sampler
BUGS	Bayesian Inference Using Gibbs sampling
PSRF	Potencial Scale Reduction Factor
DIC	Deviance Information Criterion
EAIC	Expected Akaike Information Criterion
EBIC	Expected Bayesian (or Schwarz) Information Criterion
MCE	Monte Carlo Error
iid	independiente e idénticamente distribuida
AN	Asintóticamente normal

Capítulo 1
INTRODUCCIÓN

Los modelos lineales son herramientas para explorar relaciones entre la variable respuesta y covariables de interés. Los modelos lineales clásicos, particularmente aquellos que asumen residuales normalmente distribuidos con varianza homogénea, tienen una estructura en la media que consiste de coeficientes fijos. Sin embargo, hay casos donde se asumen que estos coeficientes son aleatorios. Por ejemplo, en estudios donde hay presencia de medidas repetidas de la misma unidad de observación. Este es el caso de estudios longitudinales, que se caracterizan por recolectar información del mismo sujeto a través del tiempo.

Cuando se presentan observaciones correlacionadas y se asume distribución normal en la variable respuesta, una herramienta útil para modelar este tipo de datos son los modelos lineales mixtos (MLM). En estos modelos se incluyen tanto efectos fijos como efectos aleatorios para modelar la media. Debido a la inclusión de efectos aleatorios en la estructura de la media, estos últimos sirven para modelar estructuras de correlación entre las observaciones.

De otro lado, los modelos lineales generalizados mixtos (MLGM) nos permiten modelar respuestas no normales e incluir tanto efectos fijos como aleatorios. En escenarios donde la variable respuesta de interés son proporciones, tasas, porcentajes o probabilidades, Ferrari and Cribari-Neto (2004) propusieron un modelo de regresión beta como opción para modelar variables respuestas continuas en el intervalo $(0, 1)$. Los mismos autores sugirieron una reparametrización de la densidad de la

distribución beta clásica para dejarla en función de dos parámetros: media y dispersión. Esta reparametrización permite formular un modelo lineal generalizado usando la distribución beta. Figueroa-Zúñiga et al. (2013) extendieron el modelo anterior realizando inferencias bayesianas a un modelo de regresión beta con efectos mixtos usando Cadenas de Markov de Monte Carlo (MCMC, por sus siglas en inglés). El modelo propuesto por Figueroa-Zúñiga et al. (2013) es un modelo paramétrico. En el caso de datos longitudinales para curvas de progreso de enfermedad un modelo paramétrico podría no ser el más adecuado para describir las relaciones complejas entre la variable de interés y las covariables tales como el tiempo. Por lo tanto, una alternativa a un modelo paramétrico es un modelo semiparamétrico bajo el mismo enfoque bayesiano. Es decir, un modelo que incluya una función suave para describir el progreso de la enfermedad a través del tiempo y términos adicionales para los factores de diseño del estudio. En este trabajo proponemos modelar una función de la media de la variable respuesta la cuál cae en el intervalo $(0, 1)$ usando una función suave a través del tiempo con técnicas de suavizamiento sobre los datos, por ejemplo, *B-splines* o bases truncadas. El modelo propuesto es una extensión del modelo propuesto por Figueroa-Zúñiga et al. (2013) usando técnicas de suavizamiento para modelar las partes fija y aleatoria del modelo. Esto permitirá modelar de manera más flexible curvas de progreso de enfermedad tanto para las curvas promedio por grupos (promedio poblacional) como para las curvas de individuos (sujeto-específica). Es decir, expandimos la idea de efectos aleatorios a curvas aleatorias.

Inferencias en un MLGM pueden realizarse usando dos alternativas generales: un enfoque bayesiano o frecuentista. Una de las ventajas del enfoque bayesiano sobre el enfoque frecuentista es su capacidad de adoptar información de estudios previos, que se incorpora en el análisis de distribuciones previas de los parámetros. Esta información previa puede ayudar a mejorar la inferencia de los parámetros considerando su distribución posterior. Además, en la inferencia bayesiana la incertidumbre

de estimar tanto las curvas sujeto-específico como curvas promedio poblacionales se incorpora cuando se comparan las curvas de dos o más grupos a través del tiempo. También, la forma cómo se realiza la inferencia bayesiana usando distribuciones posteriores hace que sea más conveniente comparar curvas a través del tiempo sin necesidad de hacer ajustes requeridos como en el contexto frecuentista (por ejemplo, ajuste por multiplicidad en prueba de hipótesis). El enfoque bayesiano es más conveniente en términos computacionales debido a que no usa rutinas de optimización basadas en derivadas tales como Newton-Raphson o gradiente conjugado. Estas rutinas son particularmente problemáticas en modelos que involucran dimensiones altas. Además, a diferencia del enfoque frecuentista que usa métodos para aproximar los errores estándar de funciones de parámetros, por ejemplo, el método delta, en el enfoque bayesiano este cálculo de errores estándar se lleva a cabo fácilmente usando la distribución posterior.

El modelo propuesto en este trabajo será aplicado a un conjunto de datos de la severidad para enfermedad de la Sigatoka negra en cultivos de banano en Isabela, Puerto Rico.

1.1. Objetivos

Desarrollar un modelo semiparamétrico mixto de regresión beta usando un enfoque bayesiano para el análisis de datos longitudinales.

Objetivos específicos

1. Describir el modelo de regresión beta mixto cuando la función de la media y las curvas sujeto-específico se puedan modelar usando curvas a través del tiempo.
2. Comparar el modelo propuesto con otros modelos alternativos.
3. Evaluar el desempeño del modelo propuesto para diferentes escenarios y distribuciones previas.
4. Aplicar el modelo propuesto a un conjunto de datos reales sobre severidad de enfermedad la Sigatoka negra en cultivos de banano en Puerto Rico.

Capítulo 2
REVISIÓN DE LITERATURA

En este capítulo se realizará una revisión de literatura sobres los tópicos relacionados con el enfoque de este trabajo y que serán aplicados en los siguientes capítulos en el desarrollo de la metodología propuesta y aplicación.

2.1. Modelos Lineales Generalizados

Un modelo lineal generalizado (MLG) es una extensión de un modelo de regresión lineal ordinario. Un MLG sirve para modelar variables respuestas que no siguen una distribución normal. El modelo lineal clásico con respuesta distribuida normalmente también se considera un MLG; como un caso particular. En este tipo de modelos se asume que la variable respuesta es un miembro de la familia exponencial de distribuciones con dispersión. Una variable Y se dice que pertenece a la familia exponencial de distribuciones con dispersión si su función de densidad de probabilidad $f(y;\theta)$ puede expresarse como

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}.$$

donde θ representa el parámetro de localización, ϕ el parámetro de dispersión, $b(\cdot)$, $a(\cdot)$ y $c(\cdot)$ son funciones conocidas cuya forma depende de una distribución en particular. Este concepto de MLG fue introducido por Nelder y Wedderburn(1972). Un MLG se define usando tres componentes:

1. Un componente aleatorio correspondiente a la distribución de la variable respuesta Y. En general, se asumen observaciones independientes $(y_1, \ldots, y_n)$ que tienen una distribución de probabilidad que pertenece a la familia exponencial.
2. Un componente sistemático que involucra las covariables usadas en el predictor lineal, esto es,

$$\eta_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ip} = \boldsymbol{x_i}^T \boldsymbol{\beta}.$$

   donde $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^T$ es un vector de covariables $(1 \times p)$ y $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ es un vector de parametros de $(p \times 1)$,
3. Una función de enlace $g(\cdot)$ que se encarga de linealizar la relación entre la media de la variable respuesta y las covariables

$$g(\mu_i) = g[E(y_i)] = \eta_i = \boldsymbol{x_i}^T \boldsymbol{\beta}.$$

   La función de enlace $g(\cdot)$ es una función monótona y diferenciable. Ejemplos de funciones son el logaritmo natural para datos de conteos y la función logit $= \log \frac{\mu}{1-\mu}$ para datos binomiales.

2.2. Distribución Beta

Cuando la variable de interés es continua y se encuentra acotada en el intervalo $(0, 1)$, por ejemplo proporciones, porcentajes y tasas, una opción para analizar este tipo de datos es la distribución beta. La distribución beta es una distribución altamente flexible, permitiendo acomodar densidades unimodales y bimodales variando la severidad del sesgo. Esta flexibilidad es importante en el caso particular donde se quiera estimar distribuciones altamente sesgadas. Además, los modelos se ajustan en la escala original, lo cual hace más sencilla la interpretación. En la literatura, usualmente este tipo de datos son modelados desde diferentes enfoques tales como modelos de regresión lineal después de aplicar transformaciones a la variable respuesta. Sin embargo, esta metodología tiene algunas desventajas dado que en muchos casos los parámetros de la regresión nos son directamente interpretables en la escala

original de la variable respuesta. Además, un modelo lineal con la variable respuesta transformada puede no satisfacer los supuestos de normalidad y varianza constante debido a la asimetría de la respuesta en la escala original. Este podría ser el caso de las tasas de enfermedad las cuales podrían exhibir distribuciones altamente simétricas al comienzo y al final del estudio.

2.2.1. Definición

La función de densidad de una variable aleatoria Y que sigue una distribución beta con parámetros p y q está dada por:

$$f(y|p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{1-q},\ 0 < y < 1, \tag{2.1}$$

donde $p,\ q > 0$ y $\Gamma(.)$ denota la función gama. La media y la varianza de la distribución beta están expresadas respectivamente por

$$E(y) = \frac{p}{p+q}, \tag{2.2}$$

$$Var(y) = \frac{pq}{(p+q)^2(p+q+1)}. \tag{2.3}$$

La Figura 2–1 muestra un ejemplo de la función de densidad beta utilizando diferentes valores de los parámetros (p, q).

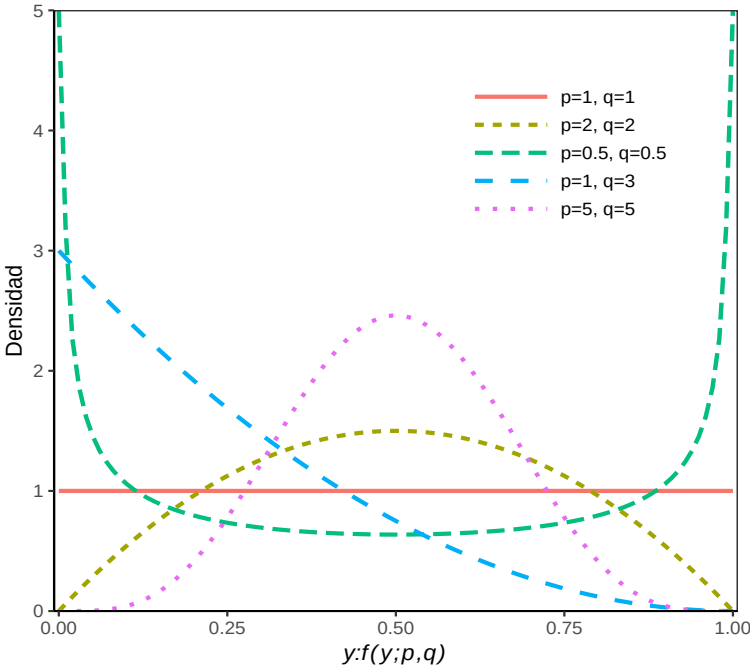

Figura 2–1: *Función de distribución beta*

La distribución beta se caracteriza por tener dos parámetros de forma, $p, q > 0$. Para usar esta distribución en el contexto de modelos lineales generalizados una simple transformación algebraica de estos parámetros define la distribución beta en términos de los parámetros de la media y escala o precisión. Dentro de la literatura de modelos de regresión beta encontramos varios autores. (Paolino, 2001) describe el uso de la distribución beta y sus ventajas sobre las estimaciones normales cuando los datos son proporciones; Kieschnick and Mccullough (2003) comparan el desempeño de la regresión beta para proporciones (ej. investigaciones en economía y finanzas) concluyendo que es la mejor opción para modelar este tipo de datos. Ferrari and Cribari-Neto (2004) proponen un modelo de regresión donde la respuesta tiene una distribución beta con dispersión fija y la media de la respuesta es modelada a través de un enlace logit. Esta formulación es introducida en la siguiente Sección.

2.2.2. Re-Parametrización de la densidad beta

La fórmula de la densidad de la distribución 2.1 está de forma estándar pero no permite la formulación de un MLG. Por lo tanto, es necesario re-parametrizar 2.1 en función de un parámetro para la media y otro para la varianza (dispersión). Con el fin de obtener una estructura de regresión para la media de la variable respuesta junto a un parámetro de precisión, Ferrari and Cribari-Neto (2004) propusieron una estructura de regresión para variables respuestas que se distribuyen beta. La densidad de y en la ecuación 2.1 puede ser escrita como

$$f(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu\phi)\phi)} y^{\mu\phi}(1-y)^{(1-\mu)\phi-1},\ 0<y<1. \tag{2.4}$$

donde $\mu \in (0,1)$, $\phi > 0$ y el parámetro ϕ puede ser interpretado como parámetro de precisión.. De las ecuaciones 2.2 y 2.3 la media de la distribución beta esta dada por

$$E(y) = \mu,$$

y su varianza

$$Var(y) = \frac{\mu(1-\mu)}{1+\phi}.$$

La formulación anterior propuesta por Ferrari and Cribari-Neto (2004) se encuentra dentro de un enfoque clásico y utiliza propiedades asintóticas de los estimadores de máxima verosimilitud para realizar inferencias.

La Figura 2–2 muestra diferentes densidades de la distribución beta para valores de (μ, ϕ).

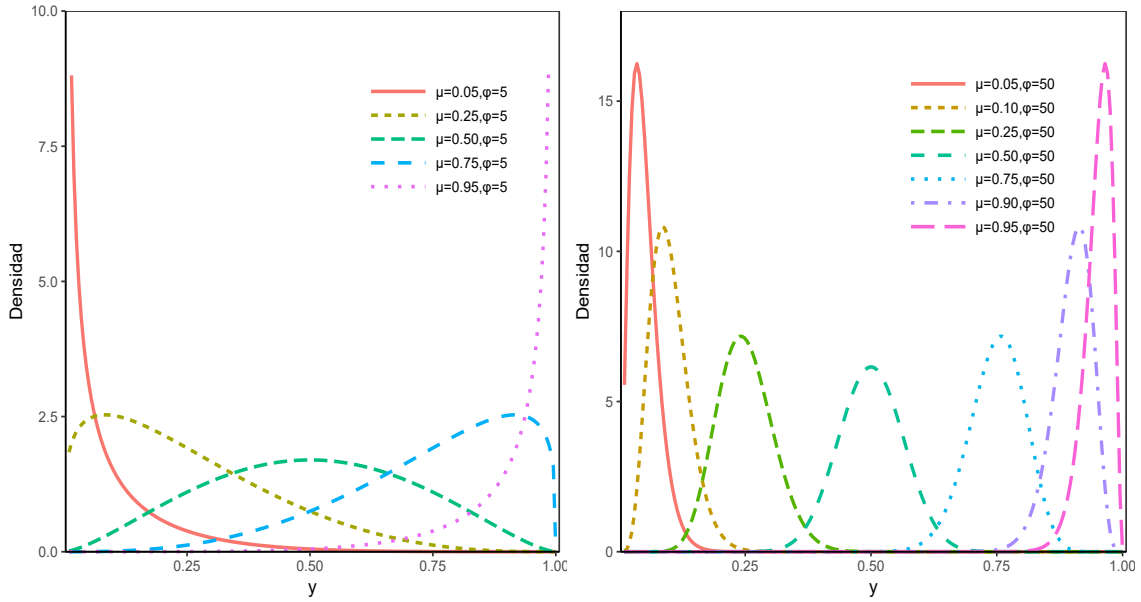

Figura 2–2: *Función de densidad beta para diferentes valores de los parámetros* (μ, ϕ).

2.2.3. Regresión Beta

Supongamos que tenemos las variables aleatorias independientes $y_1, \cdots, y_n$ tales que $y_i \sim Beta(\mu_i, \phi)$, $i = 1, \cdots, n$. Se define un modelo de regresión beta con parámetro de dispersión fijo ϕ como

$$g(\mu_i) = \boldsymbol{x_i^T \beta} = \eta_i = \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, \tag{2.5}$$

donde $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^T$ es un vector de covariables $(1 \times p)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T \in \mathbb{R}^k$ es un vector de coeficientes de regresión desconocidos y η_i es el predictor lineal. La función $g(\cdot)$ es una función de enlace estrictamente monótona y doblemente diferenciable que va del intervalo (0,1) a los números reales ($\mathbb{R}$). La Tabla 2–1 muestra las funciones de enlace más comunes para diferentes tipos de datos tales como conteos conteo y datos binarios.

Tabla 2–1: *Funciones de enlace más comunes para diferentes tipo de observaciones*

Función de enlace	Ecuación del modelo
Logit	$\log(\frac{\mu_i}{1-\mu_i}) = x_i^T\beta$
Probit	$\Phi^{-1}(\mu_i) = x_i^T\beta$
Log-Log Complementario	$\ln(-\ln(1-\mu_i)) = x_i^T\beta$
Log	$\log(\mu_i) = x_i^T\beta$

En la Tabla anterior $\Phi^{-1}(\cdot)$ es la función de distribución acumulada de una variable aleatoria con distribución normal estándar. El modelo presentado en la ecuación 2.5 puede ser escrito utilizando la función de enlace logit :

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu i}\right) = \boldsymbol{x_i^T\beta} \tag{2.6}$$

Si resolvemos para μ_i en la ecuación 2.6, el resultado es el siguiente:

$$\mu_i = \frac{\exp(\boldsymbol{x_i^T\beta})}{1+\exp(\boldsymbol{x_i^T\beta})} = \text{logit}^{-1}(\boldsymbol{x_i^T\beta}). \tag{2.7}$$

La ecuación anterior muestra una expresión para la variable respuesta y_i en función de las covariables y los coeficientes de la regresión.

2.3. Modelos Lineales Generalizados Mixtos

Los modelos lineales generalizados mixtos (MLGM) son modelos lineales generalizados con efectos aleatorios y fijos. Estos modelos son útiles cuando se tienen conjuntos de datos cuyas observaciones están correlacionadas. Por ejemplo, medidas repetidas para un mismo individuo a través del tiempo. En en contexto de MLGM se asume que dado los efectos aleatorios $\boldsymbol{b_i}$ de tamaño $(q \times 1)$, las variables aleatorias y_i, $i = 1, \ldots, n$, son condicionalmente independientes y siguen una distribución de la familia exponencial con dispersión, esto es

$$y_i|\boldsymbol{b_i} \sim f(y_i|\boldsymbol{b_i}, \phi).$$

En el modelo condicional anterior se define el predictor lineal η_i como la combinación lineal de los efectos aleatorios y fijos. La media condicional de y_i dado $\boldsymbol{b_i}$ se relaciona a las covariables como se sigue

$$\eta_i = g[E(y_i|\boldsymbol{b_i})] = g(\tilde{\mu}_i) = \boldsymbol{x_i}^T\boldsymbol{\beta} + \boldsymbol{z_i}^T\boldsymbol{b_i},$$

donde $\boldsymbol{z_i} = (z_{i1}, \ldots, z_{iq})^T$ es el vector de covariables de dimensión $(q \times 1)$ para los efectos aleatorios.

2.4. Modelo de Regresión Beta Mixto

En la subsección 2.2.2 se introdujo un modelo de regresión beta el cual no considera posibles dependencias como las inducidas por múltiples medidas del mismo sujeto a través del tiempo. En Figueroa-Zúñiga et al. (2013) se define un modelo de regresión beta mixto de la siguiente forma. Sea $y_1, \ldots, y_n$ un vector de variables aleatorias independientes $y_{ij} = (y_{i1}, \ldots, y_{im})$ con la siguiente distribución $y_{ij}|\boldsymbol{b_i} \sim beta(\mu_{ij}\phi, (1-\mu_{ij}\phi))$ con $i = 1 \ldots, n$, $j = 1, \ldots, m$.

$$g\{E(y_i|\boldsymbol{b_i})\} = \boldsymbol{x_{ij}}^T\boldsymbol{\beta} + \boldsymbol{z_{ij}}^T\boldsymbol{b_i},$$

$$\ln\left\{\frac{\mu_{ij}}{1-\mu_{ij}}\right\} = \eta_{ij} = \boldsymbol{x_{ij}}^T\boldsymbol{\beta} + \boldsymbol{z_{ij}}^T\boldsymbol{b_i}. \tag{2.8}$$

De lo anterior,

$$\mu_{ij} = \frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})} = \frac{\exp(\boldsymbol{x_{ij}}^T\boldsymbol{\beta} + \boldsymbol{z_{ij}}^T\boldsymbol{b_i})}{1+\exp(\boldsymbol{x_{ij}}^T\boldsymbol{\beta} + \boldsymbol{z_{ij}}^T\boldsymbol{b_i})},$$

donde $\boldsymbol{b_i} = (b_{i1}, \ldots, b_{iq})^T$ es el vector de efectos aleatorios de dimensión $(q \times 1)$. El modelo anterior fue ajustado dentro de un enfoque bayesiano en el cual se compararán diferentes distribuciones previas clásicas en la literatura con la previa propuesta por estos mismos autores para el parámetro de precisión ϕ en la formulación del modelo.

2.5. Regresión Semiparamétrica

Cuando los datos de estudio son medidas para un individuo repetidas en el tiempo, uno de los objetivos es modelar la variable respuesta a través del tiempo y el efecto de otras covariables en la variable respuesta. Estudios en agricultura que involucran el modelaje de curvas de progreso de enfermedad suelen enfrentar problemas para capturar la forma del progreso de la enfermedad usando formas paramétricas en la media a través del tiempo. Por lo tanto, una forma de abordar el problema es recurrir a técnicas no paramétricas que capturen el progreso promedio de la enfermedad a través del tiempo.

Los modelos de regresión semiparamétrico son técnicas estadísticas que combinan componentes paramétricos y no paramétricos. Es decir, permiten al modelo ser más flexible para modelar relaciones longitudinales más complejas y que los datos reflejen esta relación mediante una curva de suavizado no paramétrica junto a las ventajas de un modelo paramétrico. Dentro de las técnicas no paramétricas de suavizamiento existentes encontramos: *B-splines* (Boor, 1976), kernel, *splines* suavizados, regresión polinomial (Ruppert et al., 2003). Los MLGM semiparamétricos ofrecen flexibilidad para el ajuste de datos longitudinales. Es decir, los datos observados determinan la forma funcional de la asociación entre variables predictoras y la variable respuesta. Verbyla et al. (1999), Zeger and Diggle (1994) propusieron modelos semiparamétricos mixtos para datos longitudinales utilizando diferentes técnicas de suavizamiento.

2.6. Splines

Los *splines* son una forma atractiva de modelamiento de curvas de regresión, fueron implementados por Wahba (1990). Los splines se definen como una curva suave definida a trozos mediante polinomios los cuales se unen en puntos extremos llamados nodos. Son una herramienta útil debido ya que permiten el manejo de relaciones no lineales complejas. Dentro de las formas de calcular la base para la

curva de regresión encontramos bases *B-splines* (Boor, 1976, Dierckx, 1993) y bases de polinomios truncados (Ruppert et al., 2003).

Sea $k_1 < k_2 < k_3 < \ldots < k_n$ nodos. Luego un *spline* de grado $p \geq 0$ es una función $S(x)$ con $p-1$ derivadas continuas tales que:

$$S(x) = \begin{cases} P_0(x), & x < k_1, \\ P_i(x), & k_i \leq x < k_{i+1}; i = 1, 2, \ldots, n-1 \\ P_n(x), & x \geq k_n, \end{cases} \tag{2.9}$$

donde en cada intervalo $[k_i, k_{i+1}]$, $i = 1, 2, \ldots, n-1$, $P_i(x)$ es un polinomio de grado p.

2.6.1. B-splines

Dentro de las formas para calcular las bases para la curva de regresión encontramos las bases *B-splines*. En la literatura se encuentran referencias en Boor (1976) y Dierckx (1993). Un *B-spline* está formado por trozos de polinomios conectados entre sí a través de nodos. Las bases *B-splines* de grado p se definen a través de la siguiente relación de recurrencia:

$$B_{i,p}(t) = \frac{t - t_i}{t_{i+p} - t_i} B_{i,p-1}(t) + \frac{(t_{i+p+1} - t)}{t_{i+p+1} - t_{i+1}} B_{i+1,p-1}(t) \tag{2.10}$$

donde p es el orden del polinomio, t_i corresponde a la secuencia de nodos para $i = 0, \cdots, n+p$. También se define $\beta_{i,0}(t)$ que corresponde a la base de *B-splines* de orden 0.

$$B_{i,0}(t) = \begin{cases} 1, & t_i \leq t \leq t_{i+1}, \\ 0, & \text{en otro caso.} \end{cases} \tag{2.11}$$

En general, un *B-spline* de grado p tiene las siguientes propiedades (Eilers and Marx, 1996):

- Se construye a partir de $p+1$ piezas de polinomio de orden p.

- Los polinomios se unen a través de p nodos internos.
- En los puntos de unión las derivadas hasta el orden $p-1$ son continuas.
- El *B-spline* es positivo en el dominio expandido por $p+2$ y 0 en el resto.
- Excepto en los extremos, se solapa con $2p$ trozos de polinomios de sus vecinos.
- Para cada valor de x en el dominio, $p+1$ *B-splines* son no nulos.

En la Figura 2–3 se muestra un ejemplo de bases de *B-splines* lineales la cual está formado por dos trozos de polinomios de grado 1 que se unen en un nodo y bases de *B-splines* cúbicos compuestos por cuatro trozos de polinomios cúbicos. Todas las bases están definidas en el intervalo $[0, 1]$.

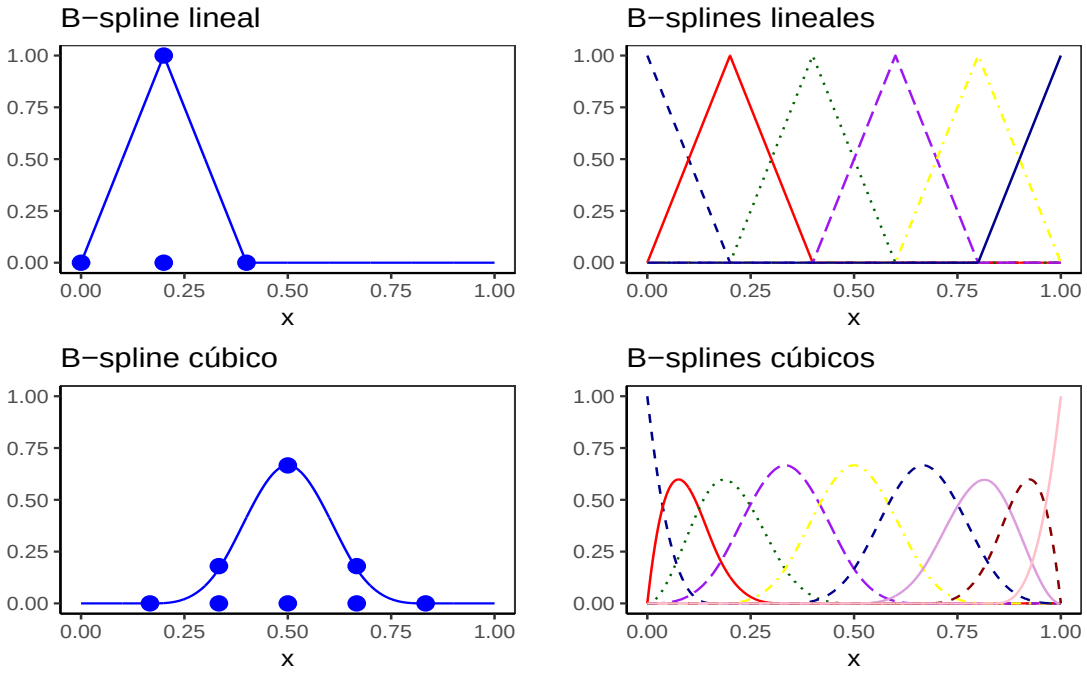

Figura 2–3: *Bases de B-splines en el intervalo* $(0, 1)$ *de orden 1 y 3. La posición de los nodos es indicada por los puntos sólidos.*

Una función suave $f(t)$ se puede expandir como una serie infinita usando las bases *B-splines* $\phi_k(t)$:

$$f(t) = \sum_{k=1}^{\infty} \beta_k \phi_k(t).$$

En la práctica se puede asumir que la función de interés es suave y se puede escribir usando un número finito de bases ϕ_k

$$f(t) = \sum_{k=1}^{K} \beta_k \phi_k(t),$$

donde β_k es el coeficiente para la base ϕ_k y K corresponde al número de bases. Para la selección de las funciones bases del modelo de estudio en este trabajo se considerarán bases *B-splines*. Las bases *B-splines* son fáciles de construir. Además, tienen ventajas numéricas y prácticas comparadas con otras bases, por ejemplo, polinomios truncados.

2.7. Modelo de regresión beta mixto con enfoque bayesiano

Un enfoque bayesiano del modelo de regresión beta con efectos fijos ha sido estudiado previamente en la literatura. Branscum et al. (2007) presentaron un enfoque bayesiano usando MCMC para el ajuste de un modelo semiparamétrico de regresión beta usando *splines* penalizados el cual fue aplicado a dos conjuntos de datos uno de ellos corresponde a la fiebre aftosa del ganado y el otro a datos de gastos del hogar en ciudades grandes de USA. Figueroa-Zúñiga et al. (2013) extendieron la regresión beta incluyendo efectos aleatorios y un enfoque bayesiano a través del *Gibbs Sampling* abordado por (Branscum et al., 2007). Figueroa-Zúñiga et al. (2013) presentaron un estudio de simulación basado en diferentes combinaciones para la media y la dispersión llevado a cabo a través del software WinBUGS. Crainiceanu et al. (2005) presentaron un análisis no paramétrico bayesiano usando regresión spline penalizada haciendo uso del software WinBUGS.

2.7.1. Análisis Bayesiano

La estadística bayesiana difiere de la estadística clásica en el uso de distribuciones previas para los parámetros del modelo. Estas distribuciones caracterizan el conocimiento acerca de los valores de los parámetros antes de la colección de los datos. Es decir, en la estadística bayesiana todos los parámetros son considerados variables aleatorias.

En general, el análisis bayesiano consiste en formular un modelo de probabilidad para los datos y decidir sobre una distribución previa para los datos que cuantificará la incertidumbre en los valores de los parámetros desconocidos antes que los datos sean observados. Una vez observados los datos se construye la función de verosimilitud basada en los datos junto al modelo formulado inicialmente.

El paradigma bayesiano se establece como

$$\pi(\theta|y) = \frac{f(\theta)f(y|\theta)}{\int f(y|\theta)f(\theta)d\theta}, \tag{2.12}$$

donde $\int f(y|\theta)f(\theta)d(\theta) = f(y)$ es la verosimilitud marginal de los datos, $f(\theta)$ es la distribución previa de θ y $f(y|\theta)$ es la verosimilitud de los datos dado θ. La ecuación 2.12 puede ser escrita de forma proporcional

$$\pi(\theta|y) \propto f(\theta)f(y|\theta). \tag{2.13}$$

La distribución previa representa un elemento clave en la inferencia bayesiana. Esta expresa conocimientos previos acerca de los parámetros desconocidos de nuestro modelo. Este conocimiento previo combinado con la distribución de probabilidad de los datos tiene como resultado la distribución posterior $\pi(\theta|y)$.

La selección de la distribución previa se divide en dos categorías: distribuciones previas no informativas e informativas. La primera corresponde a previas que expresan información vaga acerca de los parámetros. En otras palabra, $f(\theta)$ es no informativa si tiene un impacto mínimo en la distribución posterior de θ. La segunda categoría resume evidencia acerca de los parámetros y a menudo tiene un impacto considerable en los resultados. Esta evidencia puede provenir de opiniones de expertos, de la literatura o de experimentos previos (Syversveen, 1998).

Uno de los principales objetivos del análisis estadístico bayesiano es obtener la distribución posterior de los parámetros del modelo. La distribución posterior representa todo el conocimiento acerca de los parámetros después que los datos han sido

observados. Dentro del enfoque bayesiano se usarán métodos Monte Carlo con Cadenas de Markov (MCMC) por sus siglas en inglés para muestrear de la distribución posterior.

2.8. Métodos de cadenas de Markov Monte Carlo (MCMC)

Los métodos MCMC fueron introducidos en la física con el artículo de Metropolis et al. (1953). Estos métodos son una alternativa útil para aproximaciones analíticas e integraciones numéricas de modelo complicados. Los métodos MCMC están basadas en generar muestras de una densidad posterior arbitraria y a partir de estas muestras aproximarse a la distribución posterior de interés. Bajos condiciones regulares, la cadena de Markov eventualmente converge a la distribución objetivo llamada estacionaria o de equilibrio, que en nuestro caso es la distribución posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

Luego, los MCMC generan muestras dependientes de la distribución de destino, posteriormente las inferencias se llevarán a cabo utilizando la distribución empírica de la muestra. Uno de los métodos existentes para muestrear de la distribución posterior es el *Gibbs Sampling*.

2.8.1. Gibbs Sampling

Entre las técnicas MCMC el *Gibbs Sampling* es uno de los métodos más utilizado, fue descrito por Geman y Geman (1984). El *Gibbs sampling* es un caso especial del algoritmo Metropolis-Hasting en el cual la distribución propuesta está condicionada a los componentes individuales de un vector de parámetros. En general, este algoritmo es de mucha utilidad cuando la distribución conjunta del parámetro no se conoce de manera explícita pero la distribución condicional de cada parámetro dado los otros parámetros es conocida. En el artículo de Casella and George (1992) encontramos una introducción detallada de este método.

El algoritmo Gibbs se define a partir del siguiente algoritmo:

1. Definir valores iniciales arbitrarios para los parámetros de interés $\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)}$.
2. Muestrear un nuevo valor de :

$$\theta_1^{(1)} \sim \pi(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_K^{(0)}, y);$$

3.

$$\theta_2^{(1)} \sim \pi(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_K^{(0)}, y);$$

$$\ldots$$

$$\theta_K^{(1)} \sim \pi(\theta_K|\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{K-1}^{(1)}, y).$$

4. Ir al paso dos.

Repetir los pasos 2 al 4 hasta obtener uns muestra de tamaño n.

2.9. Criterios para la comparación de modelos

A continuación, se presenta un conjunto de métodos para comparar y seleccionar modelos bayesianos. Spiegelhalter et al. (2002) exponen los criterios para medir la complejidad de un modelo con enfoque bayesiano.

Sea $\{\theta^{(i)}, \cdots, \theta^{(T)}\}$ donde θ^i es el $i-ésimo$ valor simulado del parámetro de un total de T iteraciones. Silva and Lopes (2008), definen las siguientes aproximaciones de media posterior del desvio y la media posterior de θ como

$$Dbar \approx \frac{1}{T}\sum_{i=1}^{T} D(\theta^i) \quad y \quad Dhat \approx \frac{1}{T}\sum_{i=1}^{T} \theta^i.$$

El ajuste de un modelo puede ser resumido utilizando la devianza la cual se define como

$$D(\theta) = -2\ln(L(\theta))$$

donde $\boldsymbol{\theta}$ es el vector de parámetros desconocido incluido en el modelo, $L(\theta)$ es la función de verosimilitud de las observaciones. Además, se hace uso del DIC, EAIC introducidos por Brooks et al. (2002) y EBIC propuesto por Cowles and Carlin (1996). Los criterios anteriores están basados en la media posterior de la devianza $E[D(\theta)]$ y pueden ser estimados usando MCMC las siguientes expresiones,

$$DIC = Dbar + p_D = 2Dbar - Dhat, \tag{2.14}$$

$$EAIC = Dbar + 2p, \tag{2.15}$$

$$EBIC = Dbar + p\ln(N), \tag{2.16}$$

$$p_D = Dbar - Dhat = E[D(\hat{\theta})] - D[E(\hat{\theta}),] \tag{2.17}$$

donde p es el número de parámetros del modelo, N es el total de observaciones y p_D es el número de parámetros efectivo en el modelo. Entre más pequeño el valor de los criterios mejor el ajuste del modelo.

2.10. Diagnósticos de convergencia

De la teoría expuesta en la Sección 2.8 se espera que las cadenas de Markov eventualmente tiendan a una distribución estacionaria, la cual es la distribución objetivo. Dentro de la teoría para el diagnóstico de la convergencia de una cadena de Markov encontramos los métodos informales y formales. Los métodos informales se refieren a técnicas gráficas mientras que los criterios formales se refieren a pruebas univariadas y multivariadas de evaluación de las cadenas de Markov.

Entre las técnicas más populares de evaluación de convergencia de estos modelos se encuentran el diagnóstico de Geweke (1992) y Gelman and Rubin (1992). Brooks et al. (2002) realizan una revisión donde comparan los métodos anteriores y llegan a la conclusión que no se puede afirmar cual de ellos es más eficiente y se deben utilizar con precaución.

2.10.1. Métodos gráficos

Una de las formas de monitorear la convergencia y mezcla de las cadenas de Markov es el análisis gráfico de la trayectoria de la cadena y densidad posterior. Lo anterior nos proporciona indicios iniciales sobre la convergencia. También son útiles para detectar saltos periódicos. Dentro de los métodos de inspección visual encontramos la función de autocorrelación, gráficos de trayectorias (traceplots), histogramas y la media suavizada (running mean plots) el cual es un gráfico de las iteraciones versus la media de las muestras en cada iteración. Los métodos informales fueron introducidos por Gelman and Rubin (1992). A partir de estos gráficos se puede verificar si los diferentes parámetros se estabilizaron en algún valor sin mostrar tendencias estacionarias.

2.10.2. Criterio de Gelman y Rubin

Gelman and Rubin (1992) plantean un criterio de convergencia basado en el análisis de varianza. La idea general es comparar si la varianza entre las cadenas es mayor que la varianza dentro de las cadenas con diferentes puntos iniciales para cada parámetro. Desvíos grandes entre cadenas indican no convergencia. Para desarrollar este criterio es necesario $m \geq 2$ cadenas. Sea θ el parámetro de interés, $\{\theta_j\}^{[i]}$ es la $i-ésima$ de las n iteraciones de θ en la cadena m, se definen las siguientes medidas:

- Varianza dentro de las cadenas

$$W = \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_{(j)}^{[i]} - \bar{\theta}_{(j)})^2.$$

  donde $\bar{\theta}_j$ representa la media de las observaciones de la cadena m, $j = 1, \ldots, m$.

- Varianza entre cadenas

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\theta}_{(j)} - \bar{\theta})^2$$

  donde $\bar{\theta}$ representa el promedio de todos estos promedios.

- Varianza estimada

$$\hat{V}(\theta) = \frac{n-1}{n}W + \frac{1}{n}B$$

La varianza $V(\theta)$ puede ser estimada como la media ponderada de B y W. Si todas las cadenas m han alcanzado la distribución objetivo la varianza posterior estimada será muy cercana a la varianza dentro de las cadenas W. Por lo anterior, se espera que la razón $\dfrac{\hat{V}}{W}$ sea cercano a 1. Si esto no ocurre la varianza estimada $\hat{V}$ será subestimada.

Para detectar este problema se calcula la raíz cuadrada de la razón, la cual se conoce como factor de reducción de escala potencial (PSRF, por sus siglas en inglés). Este valor será interpretado como un factor de diagnóstico de convergencia.

$$\hat{R} = \sqrt{\frac{\hat{V}(\theta)}{W}}$$

Valores grandes PSRF sugieren que la varianza entre las cadenas es mayor que la varianza dentro de las cadenas, sugiriendo que es necesario aumentar el número de simulaciones. Brooks and Gelman (1998) sugieren que un valor PSRF que se encuentre menor a 1.2 para todos los parámetros del modelo es "aceptable" y se podría concluir que las m cadenas convergen. Por lo tanto, la distribución de interés fue encontrada, las cadenas han "olvidado" sus valores iniciales y en el gráfico para todas las cadenas es indistinguible.

2.10.3. Criterio de Geweke

Geweke (1992) propuso un diagnóstico de convergencia de Cadenas de Markov basado en probar la igualdad de medias de la primera parte de la cadena después del periodo de calentamiento (*burn-in*) y la última parte de la cadena (generalmente, el 10 % de la primera parte y el último 50 %). Si la distribución es estacionaria se espera que la media de la primera parte será similar a la segunda.

La estadística de esta prueba está dada por

$$z_N = \frac{\bar{\theta}_1 - \bar{\theta}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim AN(0,1) \tag{2.18}$$

Luego por el teorema central del límite a medida que aumenta en número iteraciones la distribución de la estadística Z se aproxima a una normal estándar bajos la hipótesis de convergencia de la cadena. De la ecuación 2.18 se sigue que $\bar{\theta}_1 = \sum_{i=1}^{n_1} \theta^i/n_1$ representa la media de las n_1 primeras observaciones de la cadena, $\bar{\theta}_2 = \sum_{i=1}^{n_2} \theta^i/n_2$ la media de las n_2 observaciones de la cadena, S_1^2y S_2^2 representan estimadores de la varianza $\hat{\theta}$ en la parte inicial y final de la cadena. Valores pequeños de la estadística z_N no significan que exista convergencia. Pero valores grandes de z_N indican falta de convergencia. Si el valor del $p - value$ obtenido es mayor que el nivel de significancia prefijado, entonces no existe evidencia contra la convergencia de los parámetros.

2.10.4. Criterio de Heidelberger y Welch

La prueba de convergencia de Heidelberger and Welch (1981) usa el estadístico Cramér-Von Mises el cual consta de dos partes. La primera parte consiste en probar la hipótesis nula que la cadena de Markov proviene de una distribución estacionaria y el tamaño de muestra es adecuado para estimar la media con precisión. Si la cadena no llega a la estacionalidad se descarta el 10 % de las iteraciones y la prueba es repetida nuevamente; si no se descarta otro 10 % de las iteraciones iniciales. Este proceso es repetido hasta un máximo de cinco veces. A partir de la sexta vez se considera que la prueba ha fallado e indica que hay que aumentar el número de iteraciones. Si se cumple la etapa anterior, se usa la porción de la cadena que paso la prueba y se aplica el criterio *"half-width"*. Este criterio es empleado para verificar si la media estimada está siendo calculada con precisión utilizando la parte de la cadena que paso la prueba de estacionaridad. Para verificar que cumple con lo anterior se

calcula la siguiente proporción

$$\frac{\text{límite superior} - \text{límite inferior}}{2} \tag{2.19}$$

Los resultados de la prueba pasan la prueba si la proporción de la ecuación (2.19) es menor que 0.1 con un nivel de confianza del 95 %. En este caso se concluye que la media fue calculada con precisión. Si esto no ocurre la cadena considerada no es lo suficientemente grande.

Capítulo 3
METODOLOGÍA

En este capítulo se describe un modelo de regresión semiparamétrico mixto con distribución beta propuesto desde un enfoque bayesiano. En la Sección 3.1 se define el modelo bayesiano propuesto y a su vez se especifican las distribuciones previas para cada uno de los parámetros en el modelo. En la Sección 3.3 se evalúa el desempeño del modelo propuesto a través de simulaciones.

3.1. Modelo propuesto

Uno de los objetivos de este trabajo es modelar de forma más flexible curvas de progreso de enfermedad en cultivos. Para tal propósito, se propone un modelo semiparamétrico mixto con distribución beta para modelar la relación entre la variable respuesta (severidad) y el tiempo, permitiendo términos adicionales para explicar variabilidad en la respuesta. El modelo propuesto modela la función media individual y poblacional usando funciones suaves con respecto al tiempo t. Además, se hará uso del muestreo MCMC, en particular la técnica *Gibbs sampling*, para realizar aproximaciones de la distribución posterior de los parámetros del modelo.

Sea Y_{ij} la variable respuesta definida en el intervalo $(0, 1)$ para el individuo $i = 1, \cdots, n$ con medidas repetidas $j = 1, \cdots, m$ en el tiempo t_{ij}. Se asume que la variable tiene la siguiente distribución $Y_{ij}|i \overset{ind}{\sim} beta[\mu_{ij}\phi, (1 - \mu_{ij})\phi]$. Es decir, Y_{ij}'s son condicionalmente independientes dado el sujeto i.

Se define la siguiente ecuación para la media condicional de la variable respuesta

$$\text{logit}\{E(Y_{ij}|i)\} = \text{logit}(\mu_{ij}) = \underbrace{f(t_{ij}) + f_g(t_{ij})}_{\text{parte fija}} + \underbrace{f_i(t_{ij})}_{\text{parte aleatoria.}} \tag{3.1}$$

En la ecuación (3.1) $f(t)$ representa el grupo de referencia, $f_g(t)$ las desviaciones del grupo g a la curva de referencia y $f_i(t)$ corresponde a las desviaciones de las curvas sujeto-específico de la curva promedio de su respectivo grupo. Las funciones anteriores serán modeladas en términos de bases *B-splines* discutidos en la Sección 2.6.1. Las funciones son modeladas como

$$f(t) = \sum_{k=1}^{K_1} a_k z_k(t), \;\; f_g(t) = \sum_{l=1}^{K_1} b_l^g z_l^g(t), \;\; f_i(t) = \sum_{s=1}^{K_2} c_{is} z_s^i(t) \tag{3.2}$$

donde $\boldsymbol{a} = (a_1, \cdots, a_{K_1})^T$ y $\boldsymbol{b^g} = (b_1^g, \cdots, b_{K_1}^g)^T$ son vectores de efectos fijos de tamaño $(K_1 \times 1)$ respectivamente, $\boldsymbol{c_i} = (c_{i1}, \cdots, c_{iK_2})^T$ representa un vector de efectos aleatorios de tamaño $(K_2 \times 1)$ y típicamente son asumidos independientes y normalmente distribuidos $c_{is} \overset{iid}{\sim} N(0, \sigma_c^2)$. Los $z_k(t)$ representan la $k - ésima$ base *B-spline* cúbica evaluada en el tiempo t, $z_l^g(t)$ y $z_s^i(t)$ son las bases *B-splines* cúbicas correspondiente al grupo $f_g(\cdot)$ y curvas sujeto-específico $f_i(\cdot)$, respectivamente. En el modelo K_1 representan el número de nodos de la parte fija y K_2 del componente aleatorio. Reescribiendo la ecuación (3.1) del modelo

$$\text{logit}(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \eta_{ij} = f(t_{ij}) + f_{g(i)}(t_{ij}) + f_i(t_{ij}). \tag{3.3}$$

Si resolvemos para μ_{ij} en la ecuación (3.3), la media condicional de Y_{ij} dado el sujeto i está dada por

$$\mu_{ij} = \frac{\exp\left[f(t_{ij}) + f_{g(i)}(t_{ij}) + f_i(t_{ij})\right]}{1 + \exp\left[f(t_{ij}) + f_{g(i)}(t_{ij}) + f_i(t_{ij})\right]} \tag{3.4}$$

La función de verosimilitud condicional del modelo descrito en ecuación (3.4) está dada por:

$$\begin{aligned}\mathcal{L}_c = \prod_{i=1}^{n}\prod_{j=1}^{m} f(y_{ij}|i,\mu_{ij},\phi) =& [\Gamma(\phi)]^n \Big\{ \prod_{i=1}^{n}\prod_{j=1}^{m} \Big[\Gamma\big(\tfrac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})}\phi\big)\Big] \Big\} \\ &\times \Big\{ \prod_{i=1}^{n}\prod_{j=1}^{m} \Gamma\Big[\phi\Big(1-\tfrac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})}\Big)\Big] \Big\}^{-1} \\ &\times \prod_{i=1}^{n}\prod_{j=1}^{m} \Big\{ y_{ij}^{\left(\frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})}\phi-1\right)} (1-y_{ij})^{\left[\phi\left(1-\frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})}\right)-1\right]} \Big\}\end{aligned} \tag{3.5}$$

Dentro del enfoque frecuentista, la ecuación (3.5) se usa para calcular la función de verosimilitud marginal después de integrar los efectos aleatorios. La función de verosimilitud marginal se maximiza para encontrar el estimador de máxima verosimilitud de los parámetros $\boldsymbol{a}$, $\boldsymbol{b}^g$, σ_c^2 y ϕ del modelo. Este modelo involucra $(K_1 g + 2)$ parámetros efectivos. Computacionalmente, esto implicaría aproximaciones y optimizaciones que pueden ser complicadas para el modelo propuesto. Por lo tanto, en este trabajo optamos por un enfoque bayesiano como una alternativa para estimar los parámetros en el modelo de regresión semiparamétrico mixto propuesto. En el enfoque bayesiano es necesario especificar distribuciones previas para los parámetros $\boldsymbol{a}$, $\boldsymbol{b}^g$, $\boldsymbol{c_i}$, σ_c^2 y ϕ. En el modelo bayesiano se deben estimar $(K_1 g + nK_2 + 2)$ parámetros. Una vez que las distribuciones previas han sido especificadas la distribución posterior puede ser encontrada. En la siguiente Sección se plantearon las distribuciones previas para así implementar MCMC, en particular, *Gibbs sampling*.

3.1.1. Distribuciones Previas

A continuación se define las distribuciones previas de los parámetros descritos en modelo (3.1) para así completar la especificación del modelo bayesiano. Dentro del marco de teoría de modelos mixtos bayesianos las distribuciones previas de los parámetros son cruciales para el modelaje. En consecuencia, se usarán distribuciones previas no informativas tanto para los efectos fijos como los efectos aleatorios

sugeridas por Crainiceanu et al. (2005).

Las distribuciones previas consideradas en el modelo (3.1) propuesto son las siguientes

$$\begin{cases} a_k \sim N(0, 10^{-6}), k = 1, \cdots, K_1, \\ b_l^g \sim N(0, 10^{-6}),\ g = 1, \cdots, G,\ l = 1, \ldots, K_1, \\ c_{is} \sim N(0, \tau_c),\ i = 1, \cdots, n,\ s = 1, \ldots, K_2, \\ \tau_c \sim Gamma(10^{-6}, 10^{-6}). \end{cases}$$

donde τ_c representa el parámetros de precisión, $\tau_c = \sigma_c^{-2}$, para los coeficientes de las curvas sujeto-específico. Los coeficientes a_k's, b_l^g's y c_{is}'s corresponden al vector de coeficientes de las bases *B-splines* para efectos fijos y aleatorios, respectivamente. Además, se asume que a_k, b_l^g, c_{is} y τ_c son independientes.

En la formulación del modelo (3.1) el parámetro ϕ representa la precisión la cual se considerará constante sobre las observaciones para este modelo. Usualmente, en la literatura bayesiana una distribución gamma inversa es usada para el parámetro de precisión ϕ, luego $\phi \sim IG(\epsilon, \epsilon)$, para un valor pequeño de ϵ. Hobert and Casella (1996) demuestran que la falta de conocimiento de los efectos fijos se puede afrontar asignando una distribución normal con media cero y varianza muy grande o recomienda el uso de la distribución gamma. Gelman (2006) sugiere una distribución $\phi = U^2$ donde $U \sim Uniforme(0, a)$ argumentando que esta previa es menos informativa. Siguiendo esta misma línea, Figueroa-Zúñiga et al. (2013) proponen una previa para el parámetro ϕ con $\phi = (aB)^2$, donde $B \sim Beta(1+\epsilon, 1+\epsilon)$ con $\epsilon = 0.1$.

3.2. Ajuste del modelo usando MCMC

Sea $\boldsymbol{y_i} = (y_{i1}, \cdots, y_{im})^T$ el vector de observaciones en la cual cada componente toma valores en el intervalo $(0, 1)$, $\boldsymbol{\gamma} = (\boldsymbol{a}^T, \boldsymbol{b^1}^T, \cdots, \boldsymbol{b^g}^T, \boldsymbol{c_1}^T, \cdots, \boldsymbol{c_i}^T)^T$ representa el vector de todos los parámetros desconocidos y $\boldsymbol{\zeta} = (\phi, \tau_c)$ representa el vector de precisión de todo el modelo. Para la especificación del modelo se asumirá que las

distribuciones previas de $\boldsymbol{\zeta}$ y $\boldsymbol{\gamma}$ son independientes, esto es

$$f(\boldsymbol{\zeta})f(\boldsymbol{\gamma}) = f(\tau_c, \phi)f(\boldsymbol{\gamma}) = f(\tau_c)f(\phi)f(\boldsymbol{\gamma})$$

En la Sección 3.1.1 se asignó distribuciones previas gausianas para los parámetros del predictor lineal y distribuciones no Gausianas a los parámetros de precisión. Para obtener la estimaciones de los parámetros bajo el modelo descrito en (3.3) y siguiendo el teorema de Bayes la distribución posterior conjunta $\pi(\boldsymbol{\gamma}, \boldsymbol{\zeta}|\boldsymbol{y})$ está dada por

$$\begin{aligned} \pi(\boldsymbol{\gamma}, \boldsymbol{\zeta}|\boldsymbol{y}) &\propto \Big\{ \prod_{i=1}^{n} \prod_{j=1}^{m} f(y_{ij}, c_i|\boldsymbol{\gamma}, \boldsymbol{\zeta}) \Big\} f(\boldsymbol{\zeta})f(\boldsymbol{\gamma}) \\ &\propto \Big\{ \prod_{i=1}^{n} \prod_{j=1}^{m} f(y_{ij}|\boldsymbol{\gamma}, \boldsymbol{\zeta}) \Big\} \Big\{ \prod_{i=1}^{n} f(c_i|\tau_c) \Big\} f(\tau_c)f(\phi)f(\boldsymbol{a})f(\boldsymbol{b}). \end{aligned} \tag{3.6}$$

donde $\boldsymbol{b} = (\boldsymbol{b^1}^T, \cdots, \boldsymbol{b^g}^T)$. La distribución posterior conjunta de la ecuación (3.6) es proporcional al producto de las distribuciones previas de todos los parámetros con la función de verosimilitud completa. Cabe notar que en la ecuación descrita en (3.6) las distribuciones condicionales no tienen una forma cerrada.

La aproximación del modelo propuesto fueron realizadas a través del uso del software JAGS (Plummer, 2003) el cual resuelve este problema al multiplicar las distribuciones previas de todos los parámetros con la función de verosimilitud completa y luego toma muestras de las distribuciones posteriores a través del algoritmo iterativo *Gibbs Sampling*. Así, en lugar de obtener una fórmula exacta para la distribución posterior, JAGS devuelve muestras de ella. Una vez se especificó el modelo y las distribuciones previas para los parámetros se llevó a cabo estudios de simulación para evaluar el desempeño del modelo propuesto bajo diferentes escenarios.

3.3. Simulaciones

En esta Sección se incluyen los resultados al implementar la metodología propuesta en la Sección 3.1 a un conjunto de datos simulados con el objetivo de evaluar

el desempeño del modelo propuesto. Además, se describe cada uno de los escenarios donde se evaluó el modelo y análisis de los resultados obtenidos del estudio de simulación.

3.4. Descripción del estudio

Para estudiar el desempeño del modelo propuesto con enfoque bayesiano en la ecuación 3.1 se generaron datos simulados que corresponden a curvas de observaciones a través del tiempo que poseen una estructura similar a la base de datos de la aplicación. En el **Algoritmo** 1 se presenta un esquema general de cómo se simularon los datos. La línea 1 se generan valores de la covariable tiempo t a partir de una secuencia cuyo tamaño corresponde al número de medidas repetidas m por sujeto. En la línea 2 se fijan los parámetros asociados a los coeficientes del polinomio de grado 6 para el proceso de simulación los cuales son: $\beta_0 = -1.794722, \beta_1 = 0.503973, \beta_2 = -0.08281949, \beta_3 = 0.002590216, \beta 4 = 0.0002224817, \beta_5 = -1.419026e-05, \beta_6 = 2.178633e-07$. Los valores de estos parámetros se obtuvieron a través del ajuste polinomial al conjunto de datos de la aplicación. Se escogió un polinomio de grado seis después de establecer que este grado es el que mejor describe los datos de la aplicación.

De la línea 4 a la 6 se generan los valores necesarios para crear las curvas sujeto-específico, en la línea 7 se define la curva sinusoidal. Luego, la ecuación del modelo verdadero se define de la siguiente forma

$$\text{logit}\{E(Y_{ij}|i)\} = P(t_{ij}) + S_i(t_{ij}), \tag{3.7}$$

donde $P(t)$ corresponde al polinomio de grado seis y S_i corresponde a las curvas aleatorias (Djeundje and Currie, 2011). El modelo anterior es definido para un solo grupo con desviaciones (curvas) aleatorias para cada individuo. Finalmente, en la línea 9 se generan variables aleatorias que siguen una distribución beta.

Algoritmo 1 ***Generar Datos***

Entrada: n, m, ϕ, σ_c^2

Salida: M es una matriz de datos

1: Crear la variable fija tiempo $t = 0, \cdots, m$

2: Calcular $P = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 + \beta_5 t^5 + \beta_6 t^6$

3: **para** $i = 1 : n$ **hacer**

4: $U_i \sim Uniforme(0, 1)$

5: $V_i \sim Uniforme(0, 1)$

6: $Z_i \sim Normal(0, \sigma_c^2)$

7: Calcular $S_{it} = Z_i \times sin(\pi U_i \frac{t}{(m/2)} + 2\pi V_i)$

8: Calcular la media $\mu_{it} = \dfrac{\exp(P + S_{it})}{1 + \exp(P + S_{it})}$

9: Generar valores para $y_{it} \sim Beta(\mu_{it}, \phi)$

10: $M \longleftarrow y_{it}$

11: **fin para**

12: **devolver** M

Se generaron curvas de enfermedad para diferentes tamaños de muestra y precisión clasificadas en un solo tratamiento. La Figura 3–1 presenta un ejemplo típico de un conjunto de datos simulados.

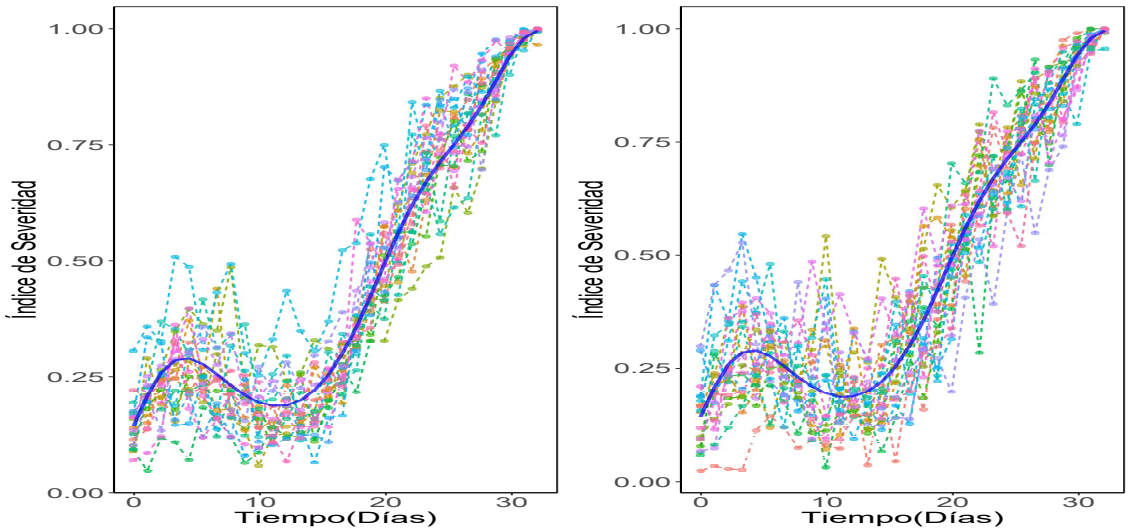

Figura 3–1: *Ejemplo de datos simulados para* $n = 20, m = 30, \sigma_c^2 = (0.5)^2$ *variando* $\phi = 60$ *(izquierda),* $\phi = 40$ *(derecha) . Las líneas entre-cortadas representan los datos observados de cada individuo y la línea sólida representa la curva asociada al polinomio del cual se generaron los datos.*

3.5. Descripción de los escenarios

Las simulaciones son consideradas bajo diferentes escenarios dado que nos permitirán medir el impacto del tamaño de la muestra y variabilidad de los datos en el desempeño del método propuesto para ajustar el modelo de interés. A continuación, se presenta un resumen de los escenarios considerados y variables a cambiar en cada una de las simulaciones. Para cada escenario se definen las siguientes variables:

1. Número de individuos n y medidas repetidas por individuo m.
2. Desviación estándar de las curvas aleatorias para los datos simulados σ_c.
3. Parámetro de precisión ϕ.

Se consideraron dos niveles para el tamaño de muestra (m, n) y dos niveles para la desviación estándar y parámetro de dispersión (σ_c, ϕ).

1. $n = 20$, $m = 30$
   a.) $\phi = 60, \sigma_c^2 = (0.5)^2$
   b.) $\phi = 50, \sigma_c^2 = (0.8)^2$
2. $n = 30$, $m = 40$
   a.) $\phi = 60, \sigma_c^2 = (0.5)^2$
   b.) $\phi = 50, \sigma_c^2 = (0.8)^2$

Con el objetivo de evaluar el desempeño del método propuesto se consideraron diferentes valores para el tamaño de muestra (n) y valores cercanos a nuestro caso de aplicación. También se consideró poca y bastante variabilidad de las curvas aleatorias para los datos simulados.

3.5.1. Descripción de distribuciones previas

Se realizó un análisis de sensitividad para evaluar el desempeño del modelo usando diferentes especificaciones de distribuciones previas para el parámetro de precisión ϕ. Este análisis se realiza con el objetivo de medir que tan susceptible es el modelo bayesiano propuesto a cambios en el parámetro de dispersión de la distribución beta utilizada. Las distribuciones previas que se utilizaron para este

análisis se basan en recomendaciones hechas en Gelman (2006) y un resumen de ellas es el siguiente:

(i) $\phi = U^2$, con $U \sim U(0, 50)$

(ii) $\phi \sim IG(\epsilon, \epsilon)$, con $\epsilon = 0.001$

(ii) $\phi = (50B)^2$, donde $B \sim Beta(1 + \epsilon, 1 + \epsilon)$, $\epsilon = 0.1$.

En el estudio de simulación no se consideraron diferentes números de nodos. Se fijó la combinación de nodos $K_1 = 10$ para la parte fija y $K_2 = 3$ para la parte aleatoria. La escogencia de esta combinación se realizó por medio de pruebas previas donde se ajustó el modelo para diferentes cantidades de nodos en la parte fija y aleatoria. La combinación con mejor ajuste fue seleccionada utilizando el criterio DIC y p_D.

3.5.2. Implementación

La implementación del muestreo bayesiano para el estudio de simulación se hizo usando los métodos de simulación MCMC, en específico *Gibbs sampling*. Para ajustar el modelo propuesto se utilizó el software JAGS, (Plummer, 2003) el cual es llamado por R (R Core Team, 2017) a través de R2jags (Su and Yajima, 2015). Este paquete permite ajustar modelos de JAGS en R junto al paquete `rjags`. Además del uso de los paquetes antes mencionados se utilizaron `ggmcmc`, `lattice` y `coda` para diagnósticos gráficos y numéricos del modelo propuesto. Un esquema general de la rutina del ajuste del modelo se puede visualizar en el siguiente algoritmo.

Algoritmo 2 (Ajuste del Modelo)

Entrada: $N, n, m, \phi, \sigma_c^2$

Salida: M

1: **para** $i = 0 : N$ **hacer**
2: Generar datos usando **Algoritmo** 1
3: Definir *n.iter*, *n.burn*, *n.thin*
4: Definir verosimilitud y previas del modelo
5: Ajustar modelo `JAGS` para cada escenario
6: Calcular DIC y P_D
7: Calcular $MADE$
8: **fin para**
9: **devolver** M

En el **Algoritmo** 2 se presenta el seudocódigo del ajuste del modelo. En la línea 2 se generan los datos simulados con distribución beta. En la línea 3 se establece el número de interacciones (*n.iter*), período de calentamiento (*n.burn*) y el salto entre iteraciones (*n.thin*) que realizará el algoritmo MCMC. En la línea 4 se define la verosimilitud y distribuciones previas de los parámetros del modelo propuesto en (3.1). De la línea 6 al 7 se calcula los estadísticos DIC, p_D y la medida de desempeño MADE.

3.5.3. Modelos ajustados

Para comparar diferentes métodos de estimar el modelo 3.7 se ajustaron los siguientes modelos:

1. Modelo beta

   En este modelo asumimos una distribución condicional beta con

$$\text{logit}\{E(Y_{ij}|i)\} = \text{logit}(\mu_{ij}) = f(t) + f_i(t). \tag{3.8}$$

   Las funciones anteriores serán modeladas en términos de bases *B-splines* como sigue

$$f(t) = \sum_{k=1}^{K_1} a_k z_k(t), \;\; f_i(t) = \sum_{s=1}^{K_2} c_{is} z_s^i(t).$$

Note que estamos asumiendo la misma distribución del cual se generaron los datos pero el polinomio verdadero $P(t)$ y las curvas aleatorias $S_i(t)$ son modeladas usando *B-splines*. Este modelo anterior se ajusta para evaluar si el método estima los parámetros(curvas) en el modelo adecuadamente. Para el ajuste del modelo beta se utilizó la transformación inversa descrita en la ecuación (2.7) para obtener los valores de la media en la escala original.

2. Modelo normal con transformación logit

Sea Y_{ij} la respuesta en el intervalo $(0,1)$. Defina $r_{ij} = \text{logit}(Y_{ij})$. Usando la transformación logit inversa se tiene $Y_{ij} = \text{logit}^{-1}(r_{ij})$, En este caso asumimos un modelo lineal normal después de la transformación logit

$$r_{ij} = g(t_{ij}) + g_i(t_{ij}) + \epsilon_{ij}, \tag{3.9}$$

donde $E(r_{ij}|g_i(t_{ij}) \equiv 0) = g(t_{ij})$, $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ y $\tau_\epsilon = \sigma_\epsilon^{-2}$ corresponde el parámetro de dispersión al que se le asignó las distribuciones previas descritas en 3.5.1. Las funciones $g(t)$ y $g_i(t)$ se modelan en términos de bases *B-splines* definidas a continuación

$$g(t) = \sum_{k=1}^{K_1} e_k w_k(t), \;\; g_i(t) = \sum_{s=1}^{K_2} p_{is} w_s^i(t). \tag{3.10}$$

Para efectos de comparación del modelo 2 con el modelo 1 en la escala original $(0,1)$ se realizó el siguiente ajuste basado en expansión de Taylor de grado dos:

$$E(Y_{ij}) \cong E\{\text{logit}^{-1}[\hat{g}(t_{ij})]\} + \frac{\hat{\sigma_\epsilon}^2}{2} \cdot \frac{\exp[\hat{g}(t_{ij})] - \exp[\hat{g}(t_{ij})]^3}{\{1 + \exp[\hat{g}(t_{ij})]\}}. \tag{3.11}$$

Este modelo se incluye para evaluar el efecto de usar un modelo lineal normal transformando la respuesta en la escala $(0,1)$. Más detalles de la aproximación (3.11) se discuten en el Apéndice B.1.

3. Modelo polinomio de grados seis

El polinomio de grados seis se ajusta para tener un punto de comparación dado que los datos fueron generados con una media siguiendo el polinomio de grado seis. Se define

$$\text{logit}\{E(Y_{ij}|i)\} = \text{logit}(\mu_{ij}) = Q(t_{ij}) + f_i(t_{ij}) \tag{3.12}$$

donde $Q(t)$ corresponde al polinomio de grado seis del cual son generaron los datos para el estudio de simulaciones y $f_i(t)$ las curvas aleatorias las cuales se siguen modelando usando *B-splines* y se definen

$$Q(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^4 + \beta_5 t^5 + \beta_6 t^6, \;\; f_i(t) = \sum_{s=1}^{K_2} c_{is} z_s^i(t). \tag{3.13}$$

Este modelo sirve como punto de referencia para el desempeño de los *B-splines* dado que la forma funcional de la media es la forma verdadera de donde se generaron los datos.

3.5.4. Consideraciones computacionales

Con respecto al número de iteraciones necesarias para que una muestra de la cadena de Markov se aproxime lo suficiente a una muestra de la distribución objetivo dependerá del modelo. Para dar solución a esto se realizaron pruebas previas donde se consideraron diferentes números de iteraciones desde 1000 hasta 20000 iteraciones. Además, se varió el número de calentamiento y salto de iteración. Se encontraron los mejores resultados cuando el número de iteraciones es igual a 12000, un período de calentamiento de 2000 y un salto de iteración de 10 usando los criterios de convergencia numéricos y gráficos discutidos en la Sección 2.10. Una vez obtenidas las simulaciones se monitoreó la convergencia de las cadenas del modelo propuesto bajos los diferentes escenarios. En conjunto se realizó un análisis de sensitividad con respecto a las distribuciones previas usadas para el parámetro de dispersión ϕ.

Algunas simulaciones de este estudio fueron realizadas en un PC de 4.0 GB de memoria, 2.50 GHz CPU, procesador Intel Core i3-310M y en promedio cada

simulación tardaba 15 minutos para tamaños de muestra pequeños. Dado que el modelo propuesto presenta cierta complejidad computacional y en algunos casos las simulaciones no culminaban de manera satisfactoria para tamaños de muestra grande, se hizo uso de las computadoras de XSEDE donde cada nodo de esta supercomputadora tiene un procesador Xeon E5 y memoria de 32GB. Luego, el proceso de simulaciones se realizó de forma continua haciendo uso de las computadoras de XSEDE.

3.6. Medidas de desempeño del método de estimación

Para evaluación del desempeño de los modelos propuestos en la Sección 3.5.3 se requiere el uso de una medida que mida la cercanía entre curvas. Por lo tanto, proponemos usar la *error absoluto integrado medio* (MIAE, por sus siglas en inglés) para evaluar el desempeño de los modelos propuestos. Supongamos que tenemos una curva verdadera $\xi(t)$ y su estimado $\hat{\xi}(t)$. Una medida robusta del sesgo y variabilidad del estimador $\hat{\xi}(t)$ se define como

$$MIAE = E\Big\{ \int |\hat{\xi}(t) - \xi(t)|\, dt \Big\}. \tag{3.14}$$

Para comparar los tres modelos ajustados usamos una versión empírica del MIAE la cual compara las curvas sobre los puntos observados a través del tiempo. Esta medida se conoce como *desviación media del error absoluto* (MADE, por sus siglas en inglés) y se define como

$$MADE(\hat{\xi}) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mid \hat{\xi}(t_{ij}) - \xi(t_{ij}) \mid \tag{3.15}$$

La curva verdadera $\xi(t)$ en nuestros escenarios de simulación corresponde al polinomio de grado seis $P(t)$. Los respectivos estimadores $\hat{\xi}(t)$ de esta curva típica $P(t)$ para cada modelo ajustado son: $\hat{f}(t)$, $\text{logit}^{-1}[g(t)]$ con el ajuste propuesto en 3.11 y $\hat{Q}(t)$ para el modelo asumiendo la forma polinomial. Los valores de $\hat{\xi}(t)$ para cada modelo están en la escala $(0, 1)$ y corresponden a las medias posteriores de $\hat{\xi}(t)$ en

cada modelo. Se calculó el MADE para cada simulación en los distintos escenarios. Un modelo con menor MADE corresponde a un modelo cuya curva estimada es más próxima a la curva real.

Otra medida reportada en la literatura es el *error cuadrático integrado medio* (MISE, por sus siglas en inglés), (Marron and Wand, 1992)

$$MISE = E\Big\{ \int [\hat{\xi}(t) - \xi(t)]^2 \, dt \Big\}. \tag{3.16}$$

La medida de desempeño MADE propuesta es más robusta a valores atípicos debido a la norma L_1 usada.

3.7. Resultados

Los resultados de las muestras MCMC obtenidas aplicando *Gibbs Sampling* descrito en la Sección 2.8.1 se presentan a continuación. En el estudio se simulación se generaron dos cadenas paralelas cada una de 12.000 iteraciones Monte Carlo tomando un período de calentamiento (*burn-in*) de 2.000 y un salto entre iteraciones (*thinning*) de 10. Este último, tiene como objetivo reducir la autocorrelación entre las cadenas y mejorar la convergencia.

3.7.1. Resultados de convergencia

Para analizar la convergencia de las cadenas y estimaciones respectivas de los parámetros en el modelo propuesto se utilizaron prueba de diagnósticos de convergencias numéricas como: la versión multivariada del diagnóstico de convergencia de Gelman y Rubin's propuesta por Brooks and Gelman (1998), Gelman and Rubin (1992),Geweke (1992) y el diagnóstico de Heidelberg y Welch's obtenidos con el paquete `ggmcmc` en `R-project` y pruebas gráficas como: gráfico de traza, de autocorrelación y densidad entre otros discutidos en la Sección 2.10.

Las gráficas de las densidades y series de tiempo posteriores obtenidas en la simulaciones se presentan en la Figura 3–2 para una de las combinaciones de los escenarios propuestos. En general, se puede considerar una buena mezcla y una traza uniforme lo cual nos da indicios de estacionaridad. En todos los escenarios de simulación las cadenas partieron de diferentes puntos iniciales.

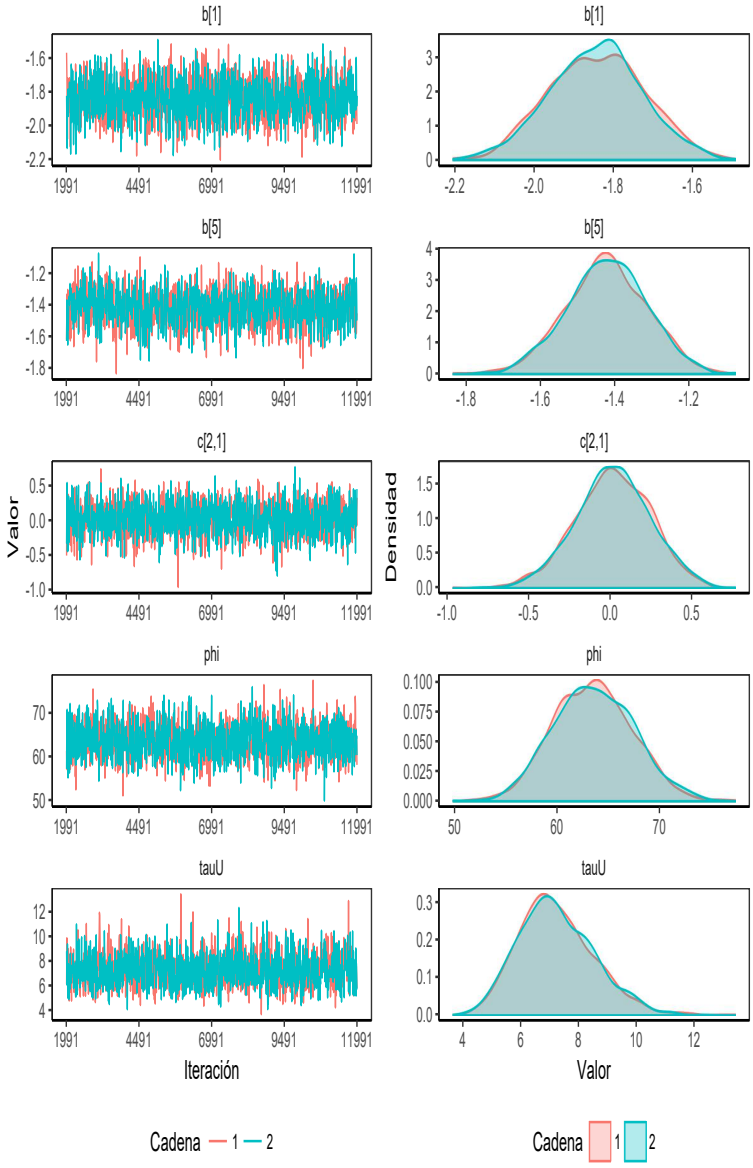

Figura 3–2: *Gráfico de las trazas y densidades para 2 cadenas para algunos de los parámetros del modelo con datos simulados, donde b*[·] *corresponde a los coeficientes de la parte fija y c*[·] *a los coeficientes de la parte aleatoria.*

La Figura 3–3 muestra los gráficos de autocorrelación y media suavizada. Estas indican que la autocorrelación decrece rápidamente, disminuyendo la dependencia asociada a cada cadena.

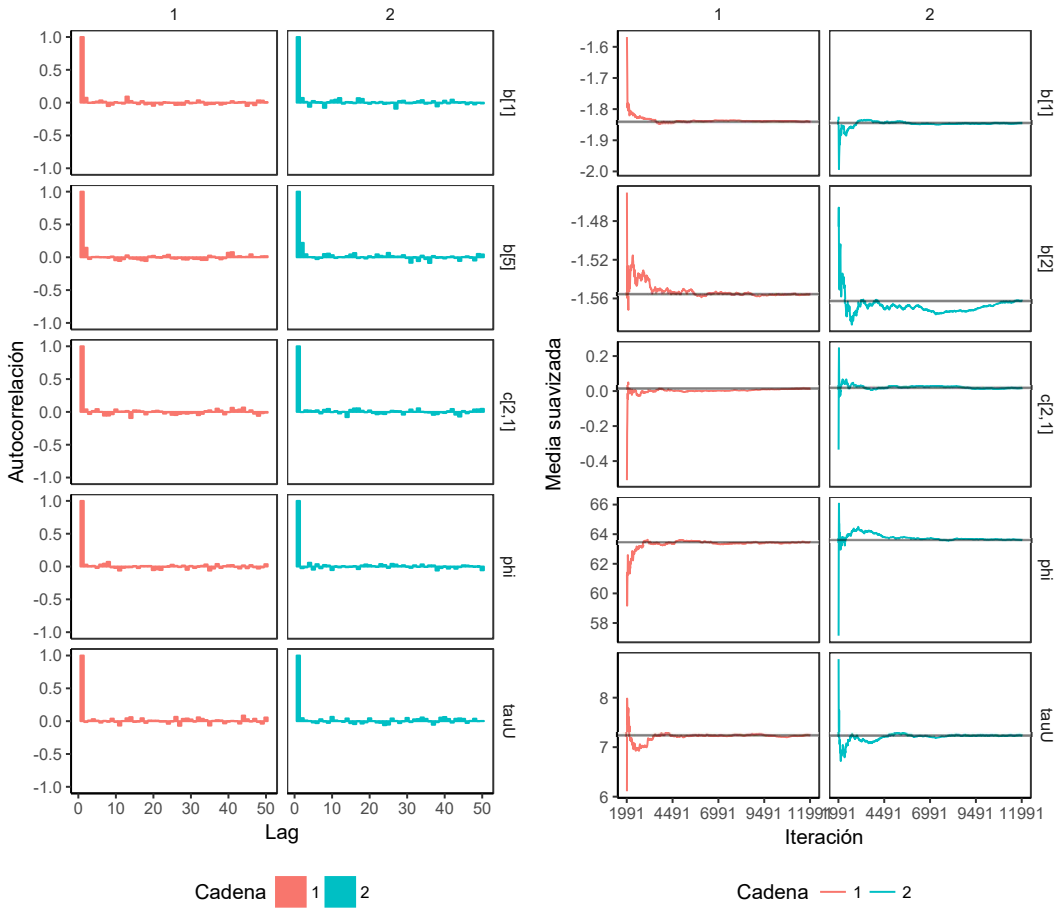

Figura 3–3: *Gráfico de autocorrelación y media suavizada de las cadenas MCMC para algunos de los parámetros del modelo con datos simulados.*

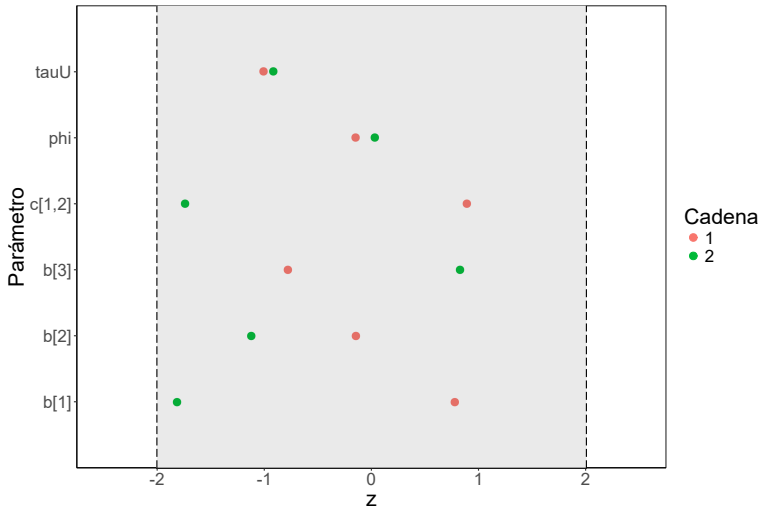

Figura 3–4: *Diagnóstico de convergencia de Geweke.*

La Figura 3–4 muestra la representación gráfica de los Z-scores del diagnóstico de Geweke. En ella se puede observar que la mayoría de los Z-scores se encuentran dentro del intervalo esperado. Esto nos da indicio de convergencia en los parámetros del modelo.

Tabla 3–1: *Resumen de diagnósticos de convergencia Gelman, Geweke, Heidelberg y Welch de algunos parámetros para datos simulados.*

Prueba	Prueba estadística					
	b_1	b_8	$c_{10,2}$	$c_{17,1}$	ϕ	τ_u
Gelman						
Estimador Puntual	1.001	1.000	1.000	1.001	0.99	0.99
$\hat{R}$						1
Geweke						
Cadena 1	1.26206	0.50864	-0.61468	-1.01619	-0.06922	-1.68454
Cadena 2	-0.3915	-0.6705	-0.5558	1.4281	2.4844	-0.3348
Heidel						
Cadena 1						
Stationarity test	pasó	pasó	pasó	pasó	pasó	pasó
Star iteration	1	1	1	1	1	1
P-value	0.506	0.288	0.617	0.846	0.894	0.707
Halfwidth test	pasó	pasó	pasó	pasó	pasó	pasó
Media	-1.841	-0.603	0.200	0.1703	63.427	7.240
Halfwidth	0.008	0.008	0.015	0.013	0.24	0.082
Cadena 2						
Stationarity test	pasó	pasó	pasó	pasó	pasó	pasó
Star iteration	1	1	1	1	1	1
P-value	0.45	0.175	0.748	0.607	0.350	0.890
Halfwidth test	pasó	pasó	pasó	pasó	pasó	pasó
Media	-1.844	-0.607	0.198	0.180	63.592	7.327
Halfwidth	0.007	0.008	0.015	0.014	0.24	0.084

La Tabla 3–1 presenta los resultados obtenidos para los criterios de convergencia para las cadenas. Por el criterio de Geweke, los resultados para los parámetros en el modelo pasaron la prueba de estacionaridad con un nivel de significancia escogido

de 0.05. Para el criterio de Gelman el factor $\hat{R}$ (estimador de reducción potencial de escala) en todos los casos es igual a 1 lo cual indica que las cadenas simuladas se han solapado. Es decir, pertenecen a la distribución estacionaria. También se puede notar que los parámetros no fueron rechazados por el diagnóstico de Heidelberger y Welch. Además, pasaron la prueba Half-Width. En la Figura 3–5 se muestra el gráfico del mejor ajuste del modelo de regresión semiparamétrico mixto de acuerdo a los resultados obtenidos el cual tiene distribución previa $\phi = (50B)^2$, donde $B \sim Beta(1+\epsilon, 1+\epsilon)$ con $\epsilon = ,1$ para un caso particular de datos simulados.

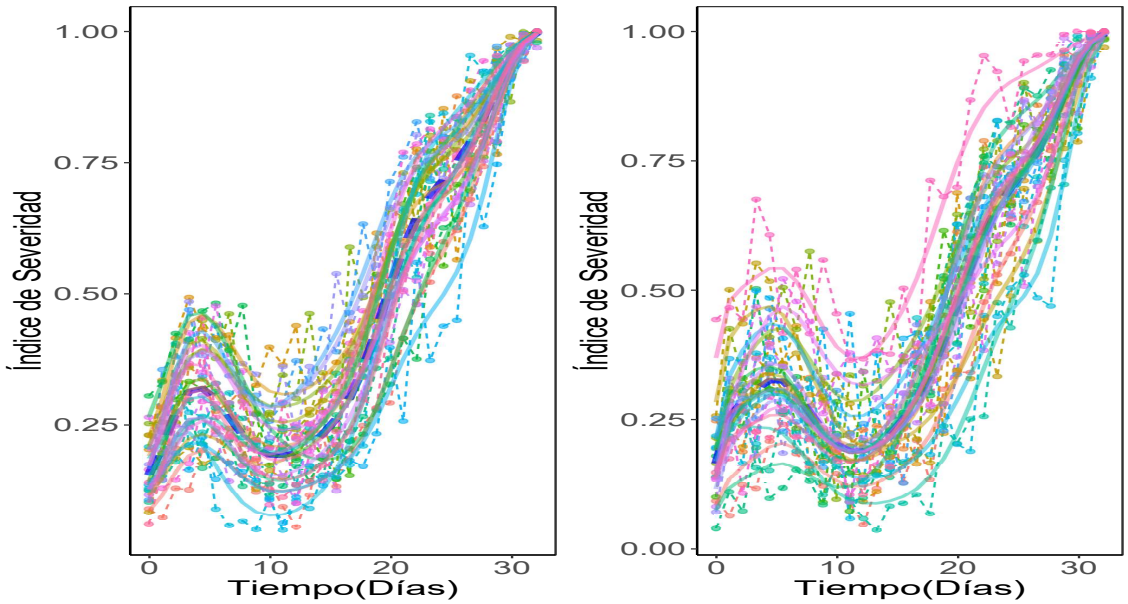

Figura 3–5: *Ejemplo de datos simulados y ajuste del modelo propuesto para* $n = 20, m = 30, \sigma_c^2 = (0.5)^2$ *y* $\phi = 60$ *(izquierda),* $\phi = 40$ *(derecha). La curva estimada por el modelo es denotada por la línea azul entre-cortada y las líneas sólidas corresponden a las curvas ajustadas por sujeto. Las líneas entre-cortadas corresponden a los datos observados.*

3.7.2. Resultados del análisis de sensitividad

A continuación, se presenta los resultados obtenidos del análisis de sensitividad para el modelo semiparamétrico mixto con distribución beta. Este análisis tiene como objetivo de evaluar la variación de los resultados del modelo propuesto bajo el uso de diferentes distribuciones previas para el parámetro de dispersión ϕ para así establecer qué tan sensibles son los resultados posteriores a variaciones de la distribución previa.

Tabla 3–2: *Análisis de sensitividad para la especificación de las previas del parámetro de precisión ϕ para el modelo con distribución beta.*

Escenarios	Distribuciones Previas	Estadístico	
		DIC	p_D
1.(a)		-1936.86	142.484
1.(b)	(i)	-1858.28	150.85
2.(a)		-3893.22	199.25
2.(b)		-3762.57	213.40
1.(a)		-1934.05	142.76
2.(b)	(ii)	-1874.78	155.20
1.(a)		-3898.03	197.345
2.(b)		-3766.407	211.293
1.(a)		-1938.82	143.23
2(b)	(iii)	-1874.81	150.82
1.(a)		-3892.08	197.65
2.(b)		-3768.73	211.53

En la Tabla 3–2 se reportan los estadísticos DIC y p_D para el modelo propuesto. Estos valores corresponde al modelo ajustado con diferentes distribuciones previas para ϕ y escenarios mencionado en 3.5. Se puede observar que para las diferentes distribuciones previas se conduce a un DIC similar en cada escenario. El modelo (iii) muestra un ajuste mejor al tener menor valor de DIC y número efectivo de parámetros p_D.

3.7.3. Resultados MADE

En la Tabla 3–3 se muestran los resultados obtenidos al usar la medida de desempeño MADE discutida en la Sección 3.6. Después de realizar 1000 simulaciones se comparan los resultados del MADE para los diferentes modelos ajustados a los datos simulados, esto es mencionado en la Sección 3.5.3. Como se puede observar los valores de MADE son más pequeños a medida que se aumenta el tamaño de la muestra, mientras que se hacen más grandes cuando la varianza de los datos aumenta. Los valores del MADE para el modelo beta presentan una mejor estimación con respecto a la distribución normal después de la transformación logit y el polinomio. También se puede notar que el valor del MADE para la variable respuesta que sigue una distribución normal con transformación logit son más grandes comparados con los otros dos casos. Es decir, las curvas estimadas están más alejadas de la curva verdadera y presentan más variabilidad a través de las simulaciones.

Tabla 3–3: *Comparación del MADE. Entradas sin paréntesis corresponden a la media y en paréntesis al error estándar de Monte Carlo.Los valores fueron multiplicados por* 100 *para evitar número muy pequeños.*

Parámetros	previas	$n = 20, m = 30$			$n = 30, m = 40$		
		Beta	Polinomio	Logit	Beta	Polinomio	Logit
$\phi = 60, \sigma_c^2 = (0{,}5)^2$	(i)	1.226 (0.039)	1.151 (0.036)	1.502 (0.047)	0.986 (0.032)	0.947 (0.030)	1.360 (0.043)
	(ii)	1.246 (0.039)	1.165 (0.037)	1.454 (0.060)	0.968 (0.040)	0.950 (0.038)	1.405 (0.054)
	(iii)	1.1214 (0.038)	1.141 (0.038)	1.489 (0.049)	0.967 (0.031)	0.937 (0.039)	1.403 (0.052)
$\phi = 50, \sigma_c^2 = (0{,}8)^2$	(i)	1.91 (0.057)	1.757 (0.056)	2.248 (0.078)	1.511 (0.054)	1.430 (0.054)	2.005 (0.086)
	(ii)	1.842(0.058)	1.782 (0.060)	2.197 (0.076)	1.474 (0.047)	1.441 (0.059)	2.008 (0.098)
	(iii)	1.804 (0.057)	1.754 (0.059)	2.277(0.072)	1.494 (0.047)	1.454 (0.064)	2.003 (0.098)

La Figura 3–6 presenta la distribución de los valores para la medida MADE para los modelos discutidos en la Sección 3.5.3 para un escenario en particular y distribuciones previas para el parámetro ϕ consideradas en el análisis de simulación.

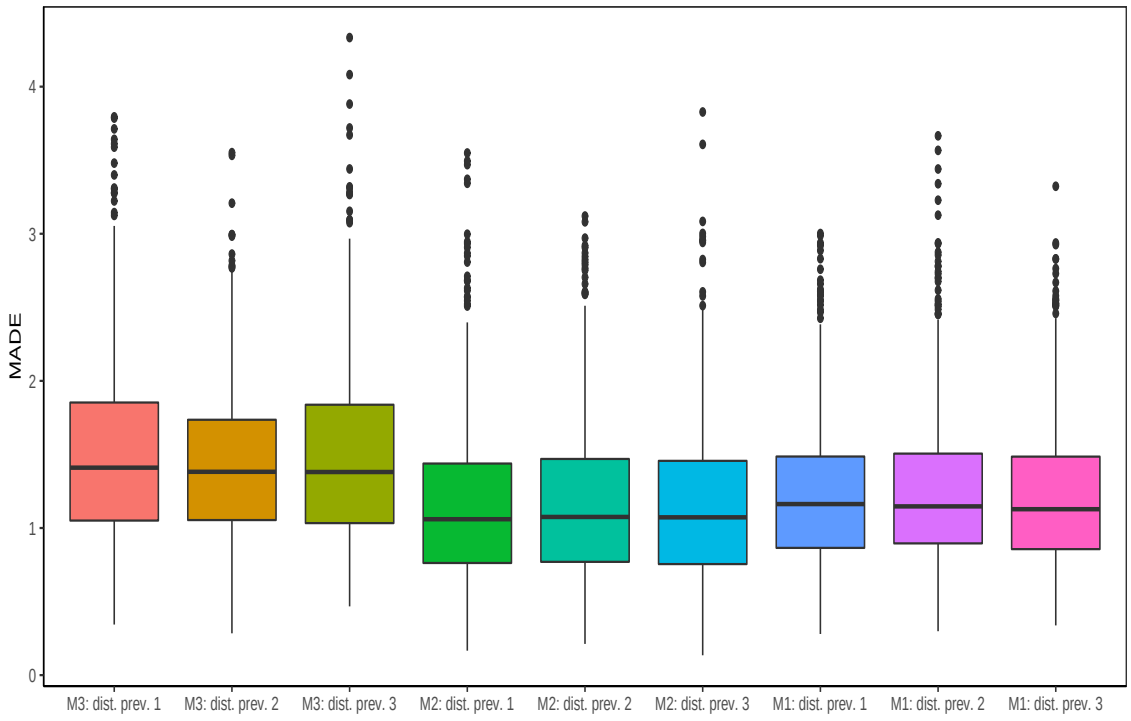

Figura 3–6: *Boxplot para la valores de la medida MADE de cada uno de los modelo propuestos en 3.5.3 para un escenario con* $n = 20, n = 30, \phi = 60$ *y las diferentes distribuciones previas propuestas.*

3.8. Conclusiones del estudio de simulación

Dentro de las conclusiones del estudio de simulación se obtuvo que el modelo semiparamétrico mixto con distribución beta tiene un mejor desempeño con respecto al modelo normal con transformación logit y el modelo polinomial. Lo anterior, se sustenta con los resultados obtenidos por la medida de desempeño MADE al presentar valores pequeños cercanos a cero. Es decir, las curvas estimadas son más próximas a la curva real. Con respecto al análisis de sensitividad, se puede observar que el modelo propuesto no presenta mucha variabilidad después de considerar diferentes distribuciones previas para el parámetro de precisión ϕ. Aun así, el modelo con distribución previa $\phi = (50B)^2$, donde $B \sim Beta(1+\epsilon, 1+\epsilon)$ con $\epsilon = 0.1$ fue el modelo seleccionado por tener el DIC más pequeño comparado con los otros dos modelos. También, se evaluó el modelo usando los diagnósticos de convergencia y se concluye que la convergencia de las cadenas es alcanzada y tienen un buen

comportamiento en todos los escenarios propuestos para el estudio de simulación. Esta última conclusión es corroborada por diferentes pruebas de convergencia tanto gráficas como numéricas discutidas en la Sección 2.10. Finalmente, se concluye que método propuesto es adecuado y presenta buenos resultados en términos del ajuste del modelo de acuerdo a los resultados obtenido en el estudio de simulación al comparar el método propuesto para diferentes modelos.

Capítulo 4
APLICACIONES: ESTUDIO DE SEVERIDAD DE ENFERMEDAD EN CULTIVO DE BANANO EN PUERTO RICO

En este capítulo se aplicará el modelo bayesiano propuesto al conjunto de datos de la aplicación los cuales corresponden a estudios de severidad de la enfermedad Sigatoka negra en cultivos de banano en Puerto Rico.

4.1. Enfermedad de la Sigatoka Negra

La Sigatoka negra es una enfermedad causada por el hongo *Mycosphaerella fijensis* Morelet. Esta enfermedad se caracteriza por un deterioro severo del área foliar de la planta, disminuyendo su capacidad fotosintética lo cual afecta el crecimiento tanto en la planta como en los racimos y frutos, en comparación con plantas sanas.(Álvarez, 2013, Marengo, 2010). En la Figura 4–1 se pueden observar la imagen de plantas infectadas con la enfermedad.

Figura 4–1: *Plantas con síntomas característicos de la enfermedad de Sigatoka negra.* (Marengo, 2010)

4.2. Descripción de los datos

Los datos utilizados en este trabajo provienen del proyecto "Practice for the Control of Black Sigatoka in Puerto Rico"(Proyecto Z-FIDA01) a cargo del Dr. José A. Chavarría Carvajal, Estación Experimental Agrícola. Se realizó una siembra de banano (*Musa acuminata, AAA cv."Grand Naine"*) por su importancia económica en Puerto Rico y su sensibilidad a la enfermedad de la Sigatoka Negra.

El diseño experimental usado en el conjunto de datos objetivo fue el de parcelas divididas con tres replicaciones por factor químico, en cada parcela experimental se realizó la siembra de 24 plantas de banano las cuales se distribuyeron en cuatro hileras de 6 plantas cada una. Con el objetivo de comparar diferentes prácticas de cultivo para controlar la Sigatoka negra, se midió el índice de severidad de las tres plantas centrales de cada hilera. Las plantas de los extremos de cada hilera fueron utilizadas como barrera con el objetivo de evitar la contaminación entre tratamientos. El estudio contó con dos factores de interés:

1. Factor Químico
   - Ausencia
   - Presencia
2. Factor Cultural
   - Tratamiento 1: Desfoliación Mecánica
   - Tratamiento 2: No Desfoliación Mecánica
   - Tratamiento 3: Deshije
   - Tratamiento 4: No Deshije

Para cada combinación de factor químico con factor cultural se obtuvo la información de 9 plantas, las cuales fueron evaluadas a lo largo de 39 semanas, generando una base de datos con un total de 2807 observaciones. Es importante señalar que la base de datos muestra datos faltantes debido a la pérdida de información en algunas

plantas a lo largo del proceso. Por lo tanto, se cuenta con 72 curvas de progreso cada una con aproximadamente 39 observaciones.

4.2.1. Índice de Severidad

Para medir el desarrollo de la enfermedad su utilizó el método de Stover modificado por Gauhl (1994). Este método estima visualmente al área total de la hoja cubierta por los síntomas de la enfermedad en plantas próximas a la floración sin necesidad de bajar la hoja. De acuerdo a esta escala de severidad se clasifica en uno de los siete grados.

Tabla 4–1: *Grados de severidad de la enfermedad de la Sigatoka negra según la escala de Stover Gauhl*

Grado	Porcentaje del área foliar afectada
0	0 %
1	Menos del 1 %
2	1 % - 5 %
3	6 % - 15 %
4	16 %-33 %
5	34 %-50 %
6	51 %-100 %

En la Figura 4–2 se observa un ejemplo donde se visualiza los siete grados en que se puede clasificar el daño de la hoja.

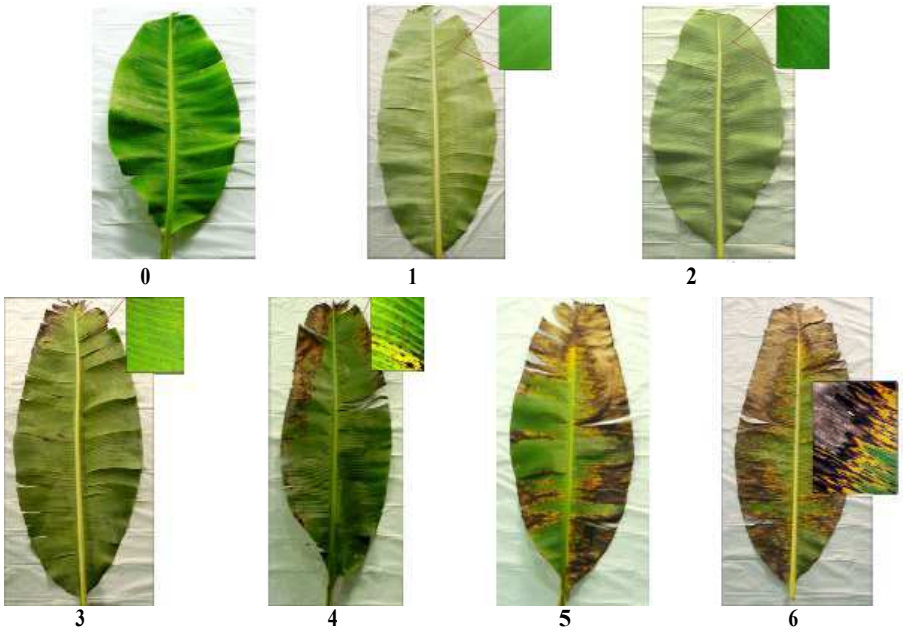

Figura 4–2: *Grados de severidad de Sigatoka negra según la escala de Stover-Gauhl.* (Marengo, 2010)

Cada hoja de una planta es medida a través de la escala de Stover-Gauhl, el promedio de todas las hojas da como resultado el índice de severidad de la enfermedad. Para el análisis con el modelo propuesto se utiliza el índice de severidad en la escala $0 - 1$.

4.3. Métodos de Análisis

Los datos se analizaron mediante un modelo semiparamétrico mixto el cual se introdujo en la Sección 4.4. Todos los análisis fueron realizados en el software R usando el paquete R2jags. La Figura 4–3 muestra las curvas de observación del proceso de la enfermedad de la Sigatoka negra a lo largo del tiempo para cada uno de los factores culturales y factor químico. El objetivo principal en este capítulo será realizar la comparación del efecto del químico y control cultural a través del tiempo usando el modelo semiparamétrico mixto bayesiano propuesto.

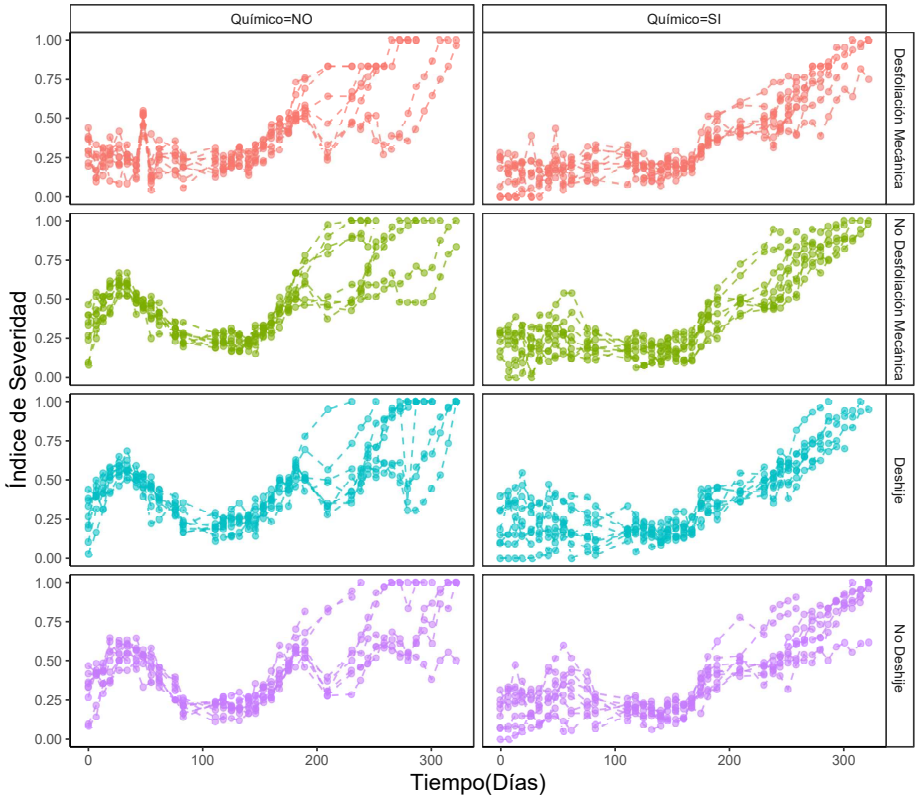

Figura 4–3: *Curvas del proceso de la enfermedad Sigatoka negra en plantas para los diferentes factores.*

4.4. Modelo de Regresión semiparamétrico Mixto

En la formulación del modelo seleccionado para el análisis del índice de severidad se utilizaron bases *B-splines* para la construcción de las funciones asociadas tanto a los efectos fijos como aleatorios del modelo.

Sea IS_{ij} el indice de severidad de la planta i, $i = 1, \cdots, 72$, $j = 1, \cdots, 40$ la ecuación para la media condicional del índice de severidad en la enfermedad de la Sigatoga negra puede ser descrita de la siguiente forma

$$\text{logit}\{E(IS_{ij}|planta_i)\} = \text{logit}(\mu_{ij}) = \underbrace{f(t_{ij}) \ + f_{g(i)}(t_{ij}) + f_{q(i)}(t_{ij})}_{\text{parte fija}} \ + \ \underbrace{f_i(t_{ij})}_{\text{parte aleatoria}} \quad . \tag{4.1}$$

En la ecuación 4.1 $f(t)$ representa el grupo de referencia (no químico, tratamiento 1), f_g(t) las desviaciones del tratamiento $g = 2, 3, 4$ a la curva de referencia, $f_q(t)$ y $f_i(t)$ corresponde a las desviaciones de la curvas de cada planta de la curva

promedio de su respectivo grupo. Las funciones son modeladas como

$$f(t) = \sum_{k=1}^{K_1} a_k z_{tk},\ f_g(t) = \sum_{l=1}^{K_1} b_l^g z_l^g(t),\ f_{q(i)}(t) = \sum_{r=1}^{K_1} d_r^q z_r^q(t),\ f_i(t) = \sum_{s=1}^{K_2} c_{is} z_s^i(t) \quad (4.2)$$

donde $\boldsymbol{a} = (a_1 \cdots, a_{K_1})^T$ corresponde al parámetros de efectos fijos de la curva de referencia, $\boldsymbol{b^g} = (b_1^g \cdots, b_{K_1}^g)^T$ corresponde al parámetro asociado al efecto de tratamiento,$g = 2, 3, 4$, $\boldsymbol{d^q} = (d_1^q \cdots, d_{K_1}^q)^T$ corresponde al parámetro asociado al efecto de químico, $q = 2$, $\boldsymbol{c_i} = (c_{i1} \cdots, c_{iK_2})^T$, $c_{i1} \overset{iid}{\sim} N(0, \tau_c)$ donde τ_c varia por químico, t_{ij} corresponde al tiempo en el cual las medidas fueron tomadas para la $i - ésima$ planta en la $j - ésima$ ocasión, K_1 es el número de nodos para los efectos fijos y K_2 el número de nodos para los efectos aleatorios. El modelo anterior se ajusta asumiendo una distribución beta para la variable respuesta. El modelo propuesto en 4.1 corresponde a un modelo sin interacción químico-tratamiento y se incluye términos asociados con el intercepto debido a que las bases suman 1. Alternativamente se consideró un modelo con interacción químico-tratamiento. Es decir, la forma funcional de la curva para cada control cultural cambia dependiendo del químico. Los resultados de ajuste de este modelo se discuten más adelante.

4.4.1. Distribuciones Previas

A continuación se define las distribuciones previas de los parámetros descritos en modelo 4.1 para así completar la especificación del modelo bayesiano. En consecuencia, se usarán distribuciones previas no informativas sugeridas por Crainiceanu et al. (2005) para los efectos fijos y para los efectos aleatorios.

Las distribuciones previas consideradas en el modelo 4.1 propuesto son las siguientes:

$$\begin{cases} a_k \sim N(0, 10^{-6}), k = 1, \cdots, K_1, \\ b_l^g \sim N(0, 10^{-6}),\ g = 1, \cdots, 4,\ l = 1, \ldots, K_1, \\ d_r^q \sim N(0, 10^{-6}),\ q = 1, 2,\ r = 1, \ldots, K_1, \\ c_{is} \sim N(0, 10^{-6}),\ i = 1, \cdots, n,\ s = 1, \ldots, K_2, \\ \tau_c \sim Gamma(10^{-6}, 10^{-6}). \end{cases}$$

Además, se asume que a_k, b_l^g, d_r^q, c_{is} y τ_c son independientes.

4.4.2. Resultados

En esta subsección se presentan los resultados del análisis para el conjunto de datos de severidad de enfermedades en cultivos de banano en Puerto Rico. En la Sección 4.4 se introdujo el modelo semiparamétrico mixto con distribución beta bases *B-splines* para modelar tanto la parte fija como aleatoria.

Selección del número de nodos

Antes de ajustar el modelo propuesto es necesario fijar el número de nodos tanto para la parte fija y aleatoria del modelo 4.1. Se realizaron pruebas con el modelo propuesto para encontrar la cantidad de nodos óptima. Para esta prueba se utilizó un número menor de iteraciones que las utilizadas en la aplicación.

Tabla 4–2: *Número de nodos óptimo para el ajuste del modelo sin interacción 4.1.*

	Número de nodos		*Estadística*	
Distribuciones Previas	*Fijos*	*Aleatorios*	*DIC*	p_D
	9	3	-6374.439	693.21
	9	4	-6362.043	856.312
$\phi = U^2$, con $U \sim U(0, 50)$	10	3	-6418.473	701.7931
	10	4	-6468.51	845.5104
	11	4	-6428.427	882.9467
	11	5	-6366.173	964.3868
	9	3	-6378.416	688.788
	9	4	-6329.066	893.2229
$\phi \sim IG(\epsilon, \epsilon)$,con $\epsilon = 0{,}001$	10	3	-6392.293	727.2039
	10	4	-6450.432	862.1677
	11	4	-6392.582	919.6847
	11	5	-6301.685	1026.612
	9	3	-6396.421	671.4131
	9	4	-6388.089	838.1296
$\phi = (50B)^2$, donde $B \sim Beta(1 + \epsilon, 1 + \epsilon)$	10	3	-6404.765	714.4499
	10	4	-6474.049	850.698
	11	4	-6434.373	878.3183
	11	5	-6409.857	921.9428

En la Tabla 4–2 se reporta el criterio de información de devianza (DIC, por sus siglas en inglés) y p_D discutido en la Sección 3.5.1 para el ajuste del modelo sin interacción al conjunto de datos de la aplicación con diferentes previas para el parámetro ϕ formulado en 4.2. Se varió el número de nodos para la parte fija y aleatoria. Aun así el modelo con 10 nodos en la parte fija y 4 en la parte aleatoria presenta un ajuste ligeramente mejor al tener menor DIC y un valor pequeño de p_D lo cuál indica un modelo parsimonioso. En particular, el modelo con una distribución beta para distribución previa $\phi = (50B)^2$, donde $B \sim Beta(1+\epsilon, 1+\epsilon)$ con $\epsilon = 0.1$ logra un mejor ajuste. Además, se comparó el modelo propuesto con y sin interacción para evaluar el ajuste al añadir el término de interacción tratamiento-químico y usando la distribución previa con mejor ajuste. En la Tabla 4–3 se puede observar el ajuste del criterio DIC para modelo propuesto con y sin interacción para las combinaciones nodos de la parte fija y aleatoria (10-3) y (10-4).

Tabla 4–3: *Comparación de número de nodos óptimos para el modelo propuesto con y sin interacción.*

Numero de nodos		Modelo con interacción		Modelo sin interacción	
Fijos	Aleatorios	DIC	p_D	DIC	p_D
10	3	-6344.78	756.58	-6404.76	714.44
10	4	-6355.21	933.7	-6474.05	850.69

Una vez especificado el número de nodos óptimos se procedió a ajustar el modelo 4.2 sin interacción a los datos de la aplicación de Sigatoka negra se consideró generar dos cadenas cada una de 70000 iteraciones Monte Carlo, un período de calentamiento (burn-in) de 40000 y un salto entre iteraciones (thinning) de 30. Al final se obtienen 1000 iteraciones para estadísticas y análisis. Adicionalmente, se realizaron los diagnósticos de convergencia gráficos y numéricos.

4.4.3. Diagnósticos de Convergencia

En la Figura 4–4 se puede observar que las cadenas se solapan muy bien, indicando que convergen hacia a la misma área, el gráfico de la derecha muestra la densidad de las probabilidad, las densidades también se solapan, sugiriendo convergencia sugieren que las cadenas para cada parámetro no se correlacionan. La Figura 4–5 nos muestra la función de densidad y traza posterior de algunos parámetros del modelos. Los diagnósticos indican que las cadenas de Markov convergen a su distribución estacionaria.

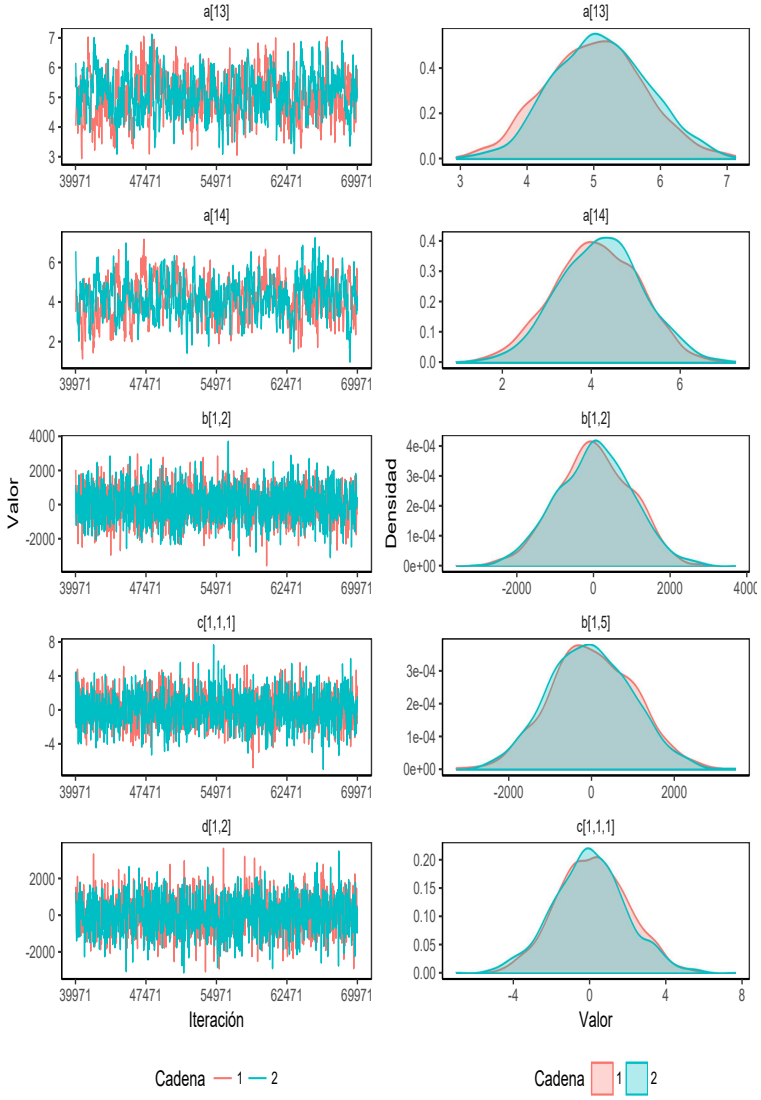

Figura 4–4: *Gráfico de las trazas y densidades para algunos parámetros del modelo con datos reales para 2 cadenas para datos de la aplicación.*

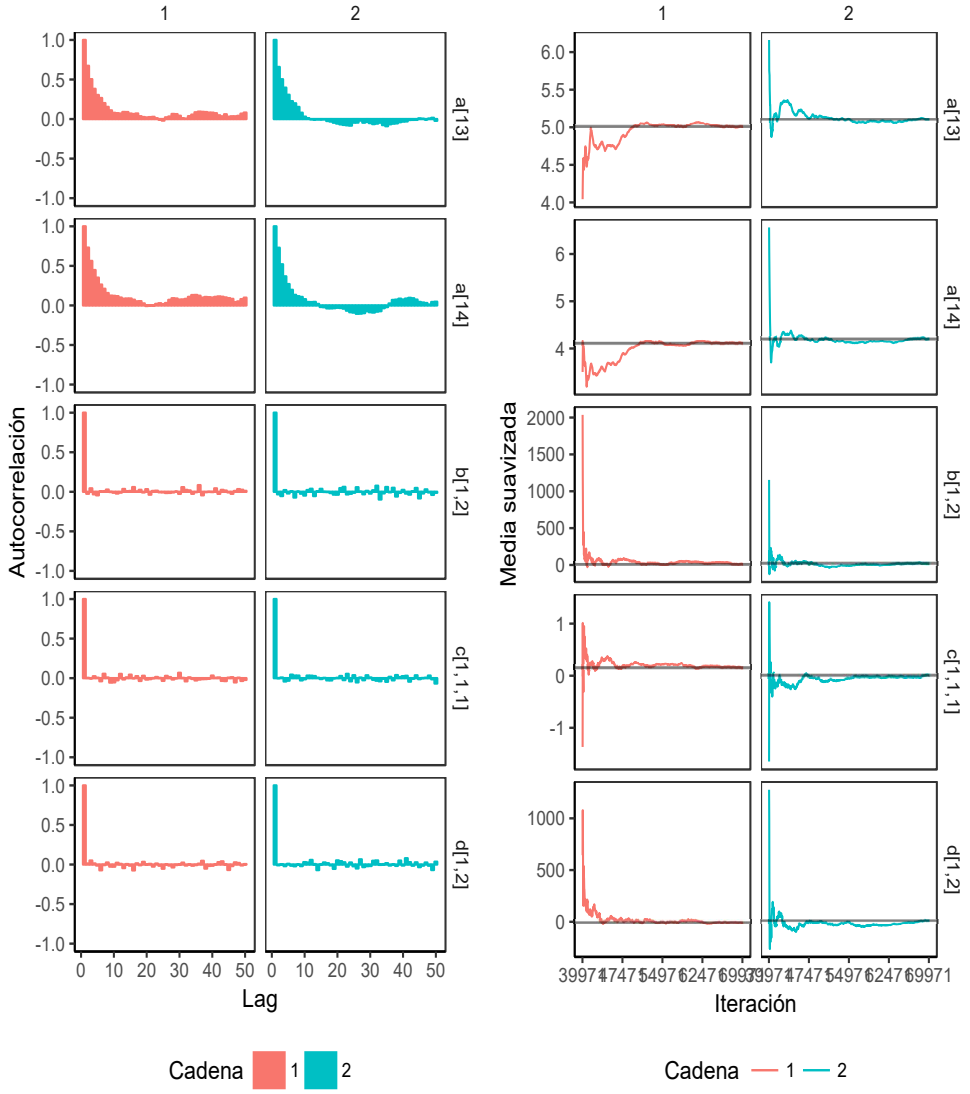

Figura 4–5: *Gráfico de las trazas y densidades para algunos parámetros del modelo con datos reales para 2 cadenas.*

Aparte de los diagnósticos numéricos, se usó el estadístico $\hat{R}$ Gelman como un indicador numérico de convergencia. En la Tabla 4–4 presenta los resultados obtenidos para los criterios de convergencia para las cadenas. Por el criterio de Geweke, los resultados para los parámetros en el modelo pasaron la prueba de estacionaridad con un nivel de significancia escogido de 0.05. Para el criterio de Gelman el factor

$\hat{R}$ (estimador de reducción potencial de escala) en todos los casos es igual a 1 lo cual indica que las cadenas simuladas se han solapado, es decir, pertenecen a la distribución estacionaria. También se puede notar que los parámetros no fueron rechazados por el diagnóstico de Heidelberger y Welch y también pasaron las prueba Half-Width. Gráficos complementarios se encuentran en el Apéndice A.1.

Tabla 4–4: *Resumen de diagnósticos de convergencia Gelman, Geweke, Heidelberg y Welch de algunos parámetros para datos de la aplicación.*

Prueba	Prueba estadística					
	a_{13}	a_{14}	$b_{2,1}$	$c_{1,1,1}$	$d_{1,2}$	ϕ
Gelman						
Estimador Puntual	1.01	1.02	1	1	1	1
$\hat{R}$						1.02
Geweke						
Cadena 1	0.37022	0.23357	-0.11963	0.10625	-0.66979	-0.07943
Cadena 2	0.7784	0.6737	0.6701	0.1168	0.8014	0.3130
Heidel						
Cadena 1						
Stationarity test	pasó	pasó	pasó	pasó	pasó	pasó
Star iteration	1	1	1	1	1	1
P-value	0.585	0.946	0.977	0.950	0.838	0.657
Halfwidth test	pasó	pasó	pasó	no pasó	pasó	pasó
Media	5.0307	4.430	1.1557	0.0205	-2.1582	38.0271
Halfwidth	0.0988	0.1389	0.0727	0.0693	60.0574	0.0851
Cadena 2						
Stationarity test	pasó	pasó	pasó	pasó	pasó	pasó
Star iteration	1	1	1	101	1	1
P-value	0.866	0.6917	0.125	0.432	0.811	0.928
Halfwidth test	pasó	pasó	pasó	no pasó	pasó	pasó
Media	5.1457	4.6460	1.1739	0.0437	-2.1320	38.0374
Halfwidth	0.0903	0.1225	0.0532	0.0736	0.0454	0.0744

4.4.4. Inferencias del Modelo

El modelo ajustado corresponde a un modelo sin interacción, 10 nodos en la parte fija y 4 nodos en la parte aleatoria. En la Tabla 4–5 se presenta los intervalos de credibilidad del 95 % y las inferencias posteriores del parámetro de dispersión ϕ de la distribución beta y de las varianzas de los coeficientes aleatorios de las curvas las cuales son diferentes por químico.

Tabla 4–5: *Estimación posterior de algunos parámetros, medias e intervalo de credibilidad del* 95 %.

Parámetro	Inferencia posterior	
	Media	Intervalo de Credibilidad del 95 %
ϕ	41.80	(39.18, 44.57)
$\sigma^2_{\text{Químico=SI}}$	3.20	(2.64, 3.99)
$\sigma^2_{\text{Químico=NO}}$	1.31	(1.07, 1.63)

La Figura 4–6 muestra la gráfica las curvas individuales observadas y ajustadas para el progreso de la enfermedad de la Sigatoka negra para las 72 plantas correspondientes a cada grupo de tratamiento y factor químico.

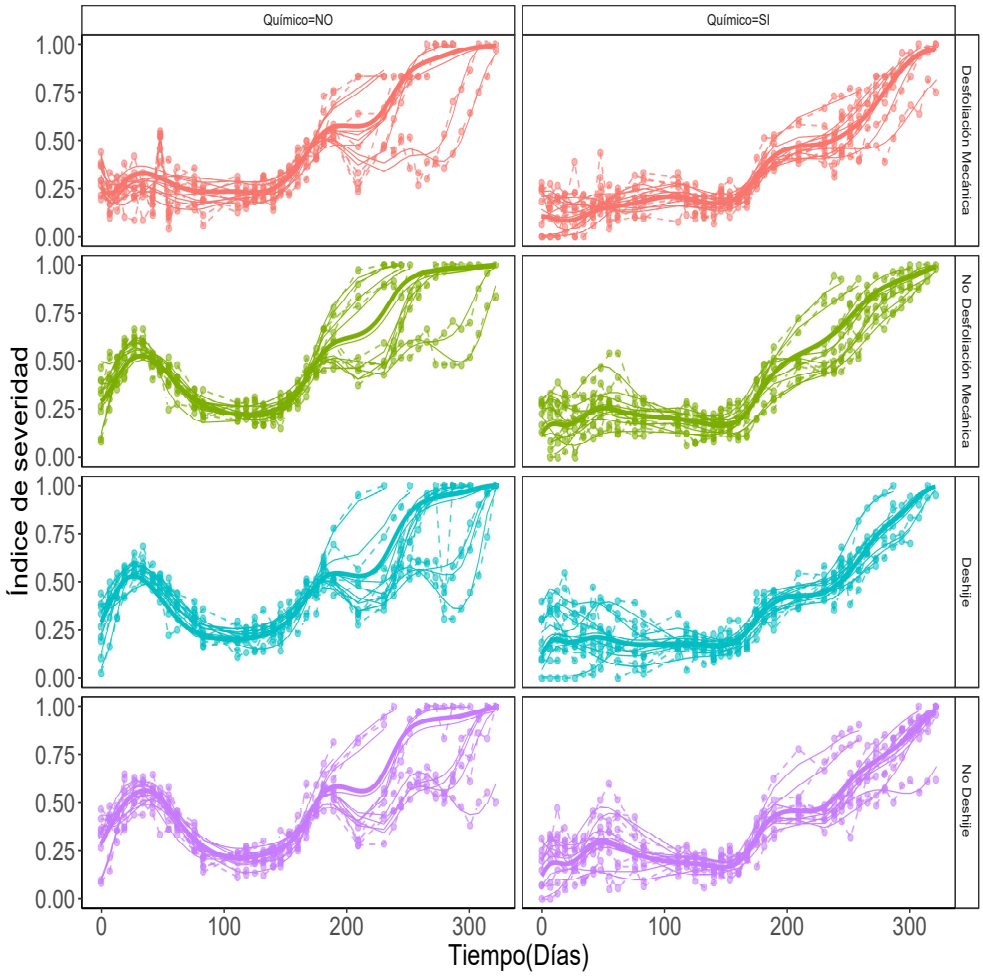

Figura 4–6: *Curvas típicas ajustadas (líneas solidas) y curvas sujeto-específico(líneas solidas delgadas) correspondientes a la media posterior.*

En la Tabla 4–6 se presentan las estimaciones posteriores para la media, mediana e intervalos de credibilidad del 95 % (percentiles 2.5 % y 97.5 %) para la severidad de Sigatoka en los tiempos $t = 0$ y $t = 25$ días. A pesar de que la severidad al comienzo del estudio ($t = 0$ días) no es interesante ya que las primeras plantas empiezan a mostrar síntomas, los resultados se incluyen para ilustrar el alcance del método. Es importante notar que para $t = 25$ días los intervalos de credibilidad para la severidad promedio en todos los tratamientos bajo la presencia de químico están por encima de los intervalos bajo la ausencia de químico. Esto sugiere que a los 25

días la severidad promedio de Sigatoka es estadísticamente inferior cuando se aplica el químico. Este resultado también se ilustra en la Figura 4–8.

Tabla 4–6: *Estimaciones posteriores para la media, mediana e intervalos de credibilidad del 95 % (percentiles 2.5 % y 97.5 %) para la severidad de Sigatoka en los tiempos $t = 0$ y $t = 25$ días usando el modelo de regresión semiparamétrico mixto (4.1).*

Parámetros		Inferencias Posteriores					
		$t =$ 0 días			**$t =$ 25 días**		
Químico	Tratamiento	Media	Mediana	Intervalo de credibilidad del 95 %	Media	Mediana	Intervalo de credibilidad del 95 %
No	Defoliación Mecáica	0.247	0.244	(0.129, 0.459)	0.317	0.313	(0.204, 0.446)
	No Defoliación Mecáica	0.256	0.254	(0.134, 0.452)	0.515	0.515	(0.382, 0.655)
	Deshije	0.292	0.288	(0.173, 0.50)	0.536	0.531	(0.386, 0.676)
	No Deshije	0.283	0.282	(0.157, 0.481)	0.537	0.536	(0.403, 0.668)
Si	Defoliación Mecáica	0.106	0.105	(0.057, 0.200)	0.084	0.082	(0.052, 0.129)
	No Defoliación Mecáica	0.110	0.110	(0.059, 0.208)	0.175	0.171	(0.114, 0.257)
	Deshije	0.123	0.126	(0.073, 0.232)	0.187	0.185	(0.119, 0.272)
	No Deshije	0.124	0.126	(0.064, 0.227)	0.188	0.186	(0.122, 0.267)

Las estimaciones con sus bandas puntuales de 95 % de credibilidad son presentadas en la Figura 4–7.

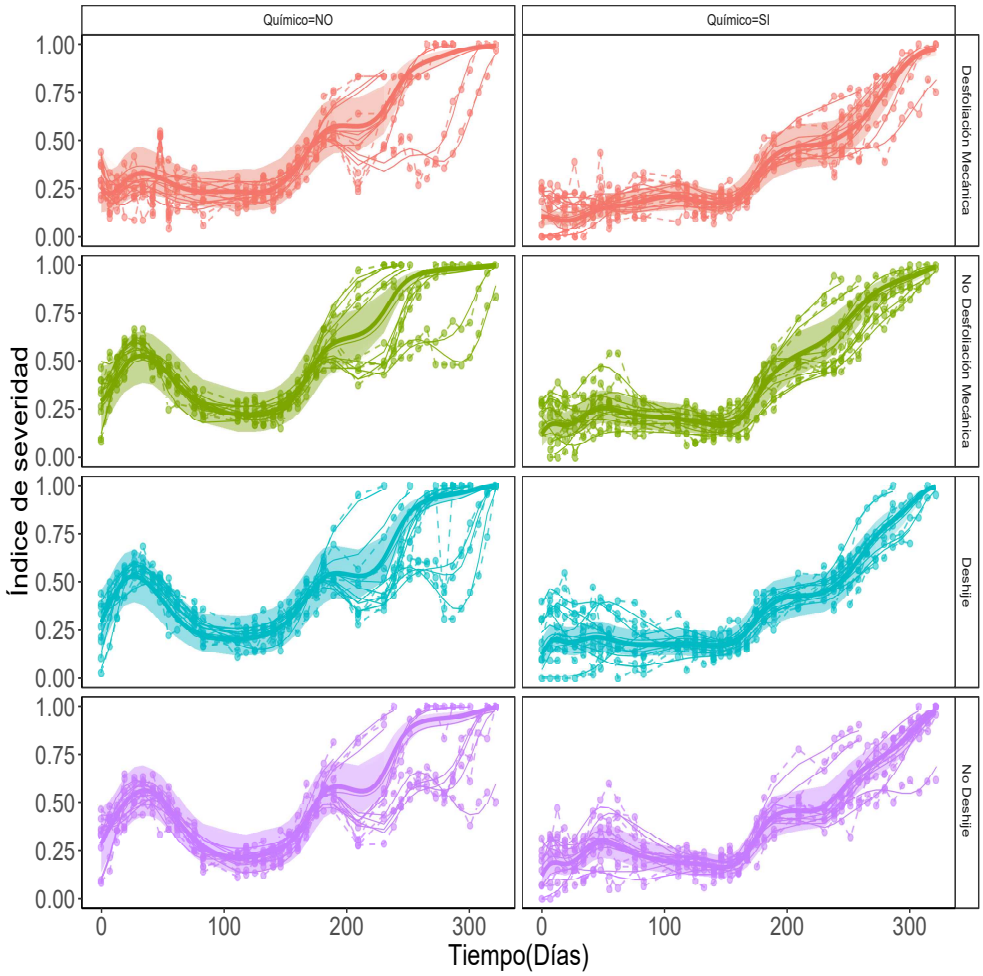

Figura 4–7: *Intervalos puntuales de credibilidad del* 95 % *al conjunto de datos de la aplicación. Las bandas corresponden al percentil* 2.5 % *y* 97.5 % *de las distribución posterior de la severidad promedio de cada grupo.*

En la Figura 4–8 se observa la curva diferencia promedio del índice de severidad de la Sigatoka negra para el efecto de químico y no químico en la escala logit a través del tiempo. En general, se puede notar que hay evidencia que el índice se severidad es menor cuando hay presencia de químico, excepto en el período de 80 a 120 días.

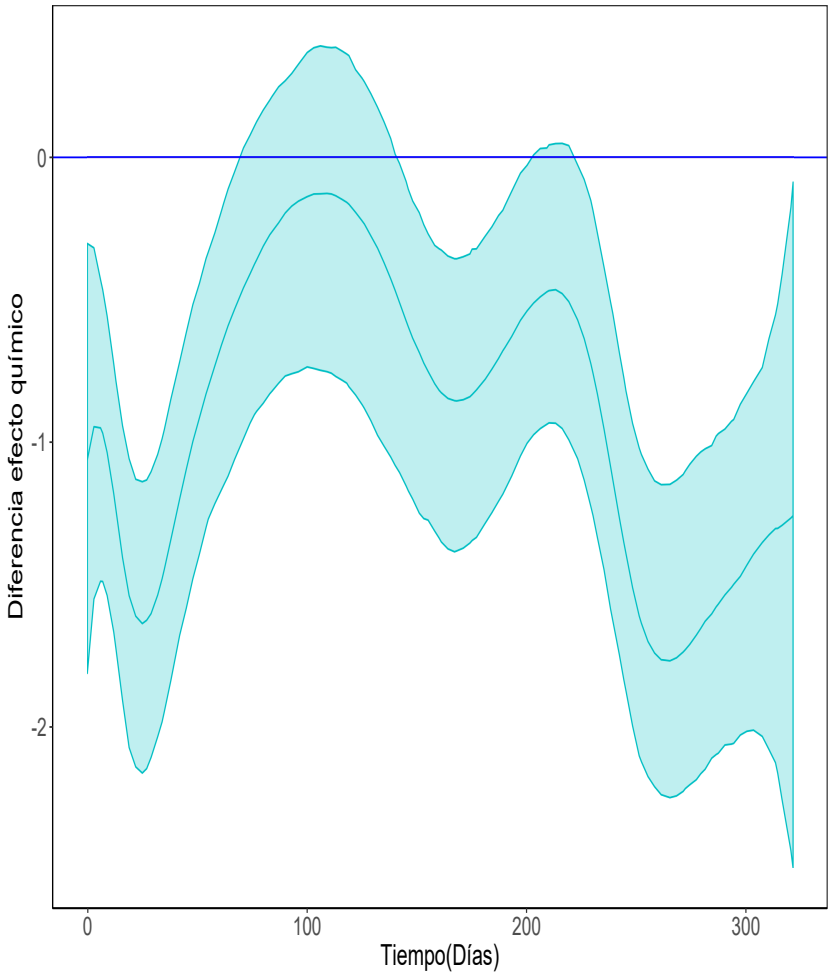

Figura 4–8: *Diferencia de la severidad promedio entre químico y no químico en la escala logit. Intervalos puntuales de credibilidad del* 95 % .

Además de lo expuesto, se pueden hacer curvas de diferencia de promedios entre tratamientos. En la Figura 4–9 se observa que no hay evidencia de diferencia entre tratamientos culturales, excepto al comienzo del estudio. Dentro de los aspectos a resaltar del estudio se tiene que el factor más importante es el químico.

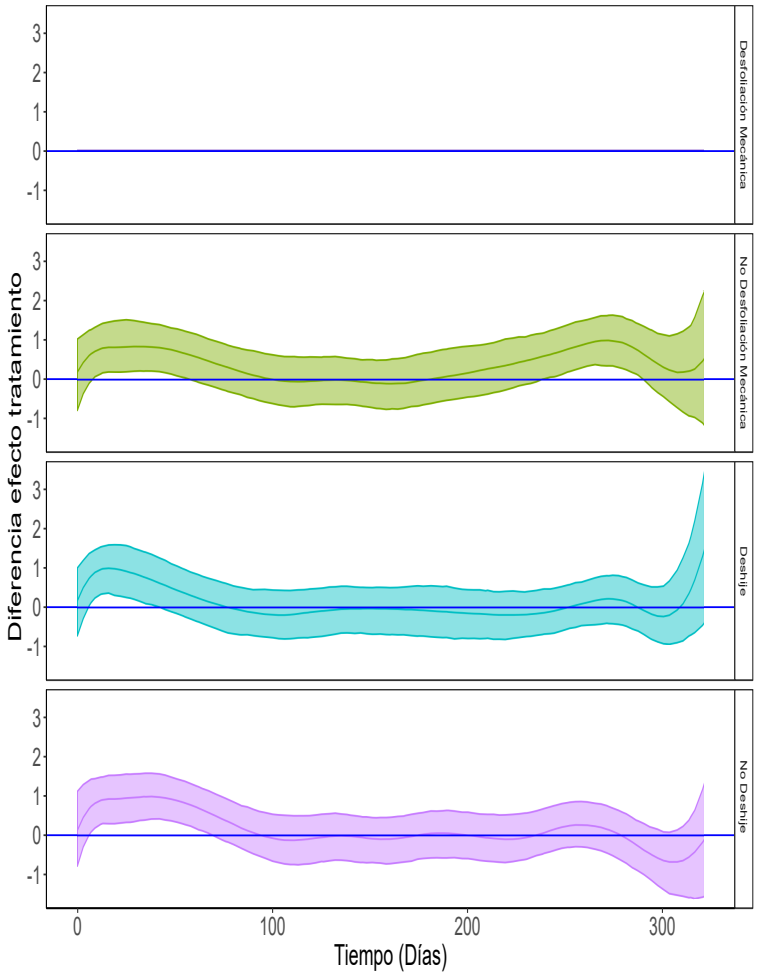

Figura 4–9: *Diferencia de la severidad promedio entre tratamientos culturales en la escala logit. Intervalos puntuales de credibilidad del* 95 % .

Capítulo 5
CONCLUSIONES Y TRABAJOS FUTUROS

5.1. Conclusiones Generales

A través de los resultados obtenidos en el proceso de simulación es posible concluir que los métodos MCMC poseen una alternativa computacional eficiente para un modelo semiparamétrico mixto con distribución beta. Los métodos semiparamétricos modelan formas flexibles para curvas de progreso de enfermedad y pueden ser usados para comparar tratamientos tomando en cuenta la estructura y diseño de los datos. La aplicación del método propuesto ofrece buenos resultados en términos del ajuste del modelo y diagnósticos de convergencia.

5.2. Trabajos Futuros

- En el estudio de simulación y en la aplicación se consideraron bases *B-splines*. Es posible investigar y comparar el ajuste obtenidos usando otro tipo de bases, por ejemplo, bases radiales o truncadas.
- Comparar el modelo de regresión beta semiparamétrico mixto frecuentista con el bayesiano.

APÉNDICES

Apéndice A

A.1. Diagnóstico de convergencia de Gelman

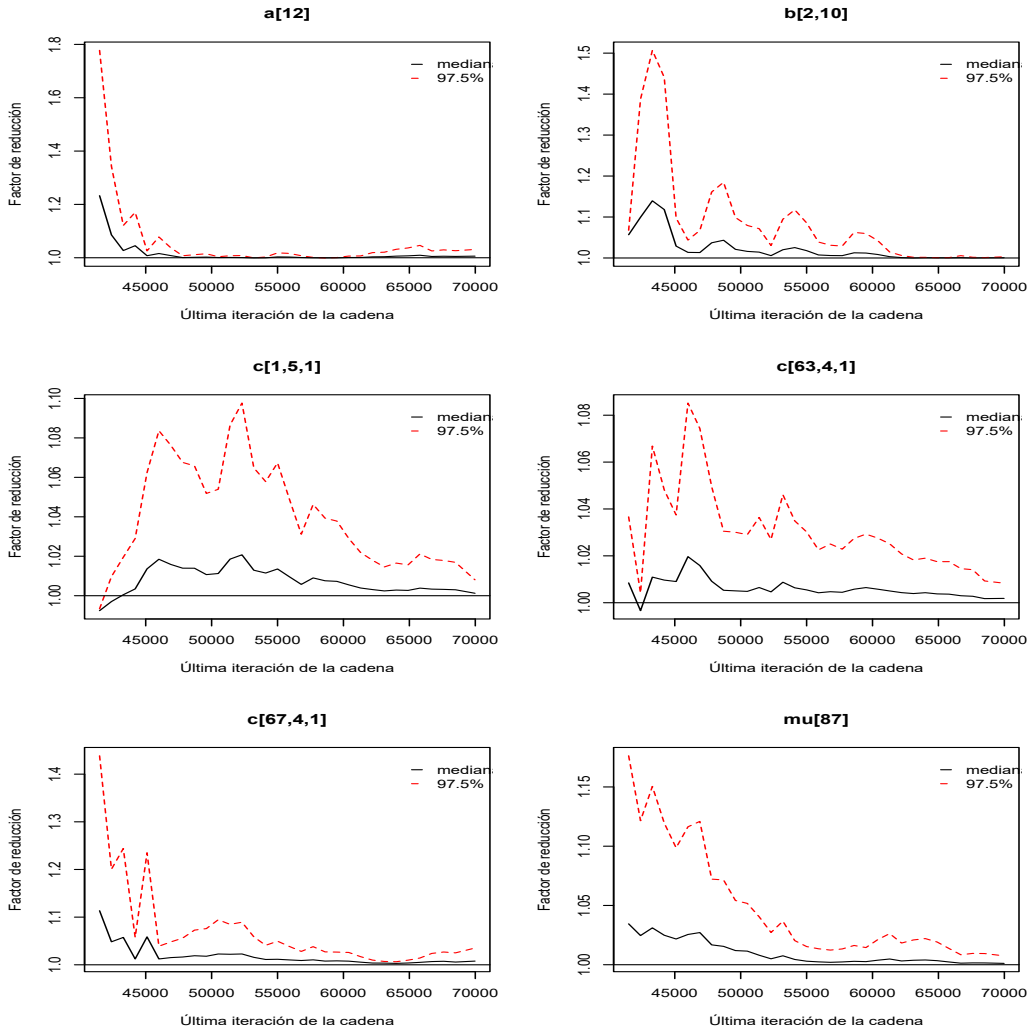

Figura A–1: *Factor de reducción de escala para algunos parámetros del modelo ajustado a la aplicación*

Apéndice B

B.1. Ajuste para el modelo lineal normal después de transformación logit

Definamos la función logit como

$$\text{logit}(t) = \ln\Big(\frac{t}{1-t}\Big), \tag{B.1}$$

y logit inversa de la siguiente manera

$$\text{logit}^{-1}(t) = \frac{\exp(t)}{1+\exp(t)}.$$

La aproximación por series de Taylor de segundo orden de una función $h(t)$ diferenciable centrada en a esta dada por

$$h(t) \approx h(a) + h'(a)(t-a) + \frac{1}{2}h''(a)(t-a)^2. \tag{B.2}$$

Sea $Y_{ij}|i \overset{ind}{\sim} beta[\mu_{ij}\phi, (1-\mu_{ij})\phi]$. Defina $r_{ij} = \text{logit}(Y_{ij})$ donde $Y_{ij} = \text{logit}^{-1}(r_{ij})$. Definamos el modelo lineal normal después de una transformación como

$$\text{logit}(Y_{ij}) = r_{ij} = g(t_{ij}) + g_i(t_{ij}) + \epsilon_{ij} \tag{B.3}$$

donde $E(r_{ij}|g_i(t_{ij}) = 0) = g(t_{ij})$ y $\epsilon_{ij} \overset{iid}{\sim} N(0, \tau_\epsilon)$ y $\tau_\epsilon = \sigma_\epsilon^{-2}$ corresponde el parámetro de precisión. Para que la curva típica verdadera de donde se generaron los datos sea comparable en la escala original $(0, 1)$ con el modelo lineal normal después de una transformación logit (3.5.3) se propone un ajuste utilizando la aproximación por la

expansión de Taylor de segundo orden para $E(Y_{ij}|S_{it} \equiv 0)$. De la ecuación (B.3)

$$E[Y_{ij}|S_{it}(t_{ij}) = 0] = E[\text{logit}^{-1}(r_{ij})|g_i(t_{ij}) = 0]$$

Sea $h(r_{ij}) = \text{logit}^{-1}(r_{ij})$. De acuerdo con (B.2)

$$\begin{aligned} E[Y_{ij}|S_{it}(t_{ij}) = 0] =& E\Big[h[g(t_{ij})] + h'[g(t_{ij})](r_{ij} - g(t_{ij})) + \frac{1}{2}h''[g(t_{ij})]\{r_{ij} - g(t_{ij})\}^2\Big] \\ =& h[g(t_{ij})] + \frac{1}{2}\{h''[g(t_{ij})]\}\Big\{E[r_{ij} - g(t_{ij})]^2|g_i(t_{ij}) \equiv 0\Big\} \\ =& \text{logit}^{-1}\{g(t_{ij})\} + \frac{1}{2}\Big[\frac{\partial^2}{\partial t^2}\text{logit}^{-1}(g(t_{ij}))\Big] \times \sigma_\epsilon^2 \\ \cong& \text{logit}^{-1}\{g(t_{ij})\} + \frac{1}{2} \cdot \frac{\exp[g(t_{ij})] - \exp[g(t_{ij})]^3}{1 + \exp[g(t_{ij})]} \times \sigma_\epsilon^2, \end{aligned} \tag{B.4}$$

donde $E\Big\{[r_{ij} - g(t_{ij})]^2|g_i(t_{ij}) \equiv 0\Big\} = E\{\epsilon_{ij}^2\} = \sigma_\epsilon^2$.

Por lo tanto, para la comparación del modelo lineal normal después de la transformación logit en la escala $(0, 1)$ se utiliza

$$\widehat{E[Y_{ij}]} = \text{logit}^{-1}\{\hat{g}(t_{ij})\} + \frac{1}{2} \cdot \frac{\exp[\hat{g}(t_{ij})] - \exp[\hat{g}(t_{ij})]^3}{\{1 + \exp[\hat{g}(t_{ij})]\}} \times \hat{\sigma_\epsilon}^2.$$

Como el estimador de la media típica verdadera.

Apéndice C

C.1. Aplicación: Modelo de regresión semiparmétrico mixto con distribución beta en JAGS

Esta sección presenta una parte del código BUGS usada para ajustar el modelo de regresión semiparamétrico mixto con distribución beta para los datos de la aplicación.

```
model_simu <- function()
{
for(k in 1:n)
{
#Likelihood function
y[k] ~ dbeta(aa[k],bb[k])
aa[k] <- mu[k] * phi
bb[k] <- (1-mu[k]) * phi

logit(m[k]) <- f[k] + fg.treat[k] + fg.chem[k] #typical mean
logit(mu[k]) <-  f[k] + fg.treat[k] + fg.chem[k] + fi[k] #subject-specific mean
f[k] <- inprod(a[], Z[k,])
fg.treat[k] <- inprod(b[tratamiento[k],], Z[k,])*step(tratamiento[k]-1.5)
fg.chem[k] <- inprod(d[quimiconum[k],], Z[k,])*step(quimiconum[k]-1.5)
fi[k] <- inprod(c[ID.consec[k], , quimiconum[k]], Zr[k,])
```

```
}
#prior definition
for (k in 1:n.basis){ a[k]~dnorm(0, 1.0E-11 ) } (treat=1, chem=1)
for (l in 1:ntreats){
for(m in 1:n.basis){
b[l, m]~dnorm(0, 1.0E-6) }
}

for (l in 1:nchems){
for(m in 1:n.basis){
d[l, m]~dnorm(0, 1.0E-6) }
}

for(ch in 1:nchems){
for (r in 1:nsubjects){
for(s in 1:n.basis.r){
c[r, s, ch]~dnorm(0, tauU[ch]) }
}
}

phi<- (phiinicial*50)*(phiinicial*50)
phiinicial~ dbeta(1.1,1.1)
for(ch in 1:nchems){  tauU[ch] ~ dgamma(1.0E-6, 1.0E-6) }
}
```

Bibliografía

Agresti, A. (2002). *Categorical data analysis*. Wiley-Interscience, 2 edition.

Boor, C. D. (1976). Splines as linear combinations of B-splines. A Survey. pages 1–47.

Branscum, A. J., Johnson, W. O., and Thurmond, M. C. (2007). Bayesian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian & New Zealand Journal of Statistics Aust N Z J Stat*, 49(3):287–301.

Brooks, S., Smith, J., Vehtari, A., Plummer, M., Stone, M., Robert, C., Titterington, D., Nelder, J., Atkinson, A., Dawid, A., Lawson, A., Clark, A., Bernardo, J., Sahu, S., Richardson, S., Green, P., Burnham, K., DeIorio, M., Robert, C., Draper, D., Gelfand, A., Trevisani, M., Hodges, J., Lee, Y., De Luna, X., and Meng, X. (2002). Discussion on the paper by spiegelhalter, best, carlin and van der linde. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64(4):616–639.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Calvo, A. (2015). Comparación de modelos de regresión semiparamétricos mixtos con distribución beta. Master's thesis, Universidad de Puerto Rico Recinto Universitario de Mayagüez.

Carlin, B. P. (1993). *Bayes and Empirical Bayes methods for data analysis*. Chapman & Hall/CRC, Boca Raton, 2nd ed. edition.

Casella, G. and Berger, R. L. (2002). *Statistical inference*. Thomson Learning.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Crainiceanu, C., Ruppert, D., and Wand, M. P. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software J. Stat. Soft.*, 14(14).

Cribari-Neto, F. and Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software J. Stat. Soft.*, 34(2).

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Oxford University Press, Inc., New York, NY, USA.

Djeundje, V. A. B. and Currie, I. D. (2011). Fitting subject-specific curves to grouped longitudinal data. In *58th ISI Congress, Dublin,*, page 6.

Dobson, A. J. (2002). *An introduction to generalized linear models*. Chapman & Hall/CRC.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with b -splines and penalties. *Statistical Science Statist. Sci.*, 11(2):89–121.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Figueroa-Zúñiga, J. I., Arellano-Valle, R. B., and Ferrari, S. L. (2013). Mixed beta regression: A bayesian perspective. *Computational Statistics& Data Analysis*, 61:137–147.

Gauhl, F. (1994). Epidemiology and ecology of black sigatoka (mycosphaerella fijiensis morelet) on plantain and banano (musa spp) in costa rica, central america. Technical report, INIBAP, Montpellier (Francia).

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:515–533.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science Statist. Sci.*, 7(4):457–472.

Geweke, J. (1992). *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments*. University Press.

Harville, D. (1976). Extension of the gauss-markov theorem to include the estimation of random effects. *Ann. Statist.*, 4(2):384–395.

Heidelberger, P. and Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Commun. ACM*, 24(4):233–245.

Hobert, J. P. and Casella, G. (1996). The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. *Journal of the American Statistical Association*, 91(436):1461–1473.

Kieschnick, R. and Mccullough, B. D. (2003). Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *stat modelling Statistical Modelling*, 3(3):193–213.

Kleinman, K. P. and Ibrahim, J. G. (1998). A semi-parametric bayesian approach to generalized linear mixed models. *Statist. Med. Statistics in Medicine*, 17(22):2579–2596.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963.

Álvarez, E., P. A. G. L. y. C. G. (2013). La sigatoka negra en plátano y banano: guía para el reconocimiento y manejo de la enfermedad, aplícado a la agricultura familiar. http://www.fao.org/docrep/019/as089s/ as089s.pdf. 09-17-2016.

Marengo, J. (2010). Epidemiología de la sigatoka negra (*Mycosphaerella fijiensis Morelet*) en una plantilla de guineo en Puerto Rico. *Tesis de Maestría, Universidad de Puerto Rico, Recinto Universitario de Mayagüez.*

Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics*, 20(2):712–736.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):pp. 370–384.

Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, 9(4):325–346.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Plummer, M. (2015). Jags version 4.0. 0 user manual.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press.

Sanchez, J. (2007). Linear models with r and extending the linear model with R. *Journal of Statistical Software J. Stat. Soft.*, 17(Book Review 4).

Searle S., M. C. a. (2001). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, Hoboken, New Jersey.

Silva, R. d. S. and Lopes, H. F. (2008). Copula, marginal distributions and model selection: a Bayesian note. *Statistics and Computing*, 18(3):313–320.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical*

Society: Series B (Statistical Methodology), 64(4):583–639.

Su, Y.-S. and Yajima, M. (2015). R2jags: Using R to Run 'JAGS'.

Syversveen, A. R. (1998). Noninformative bayesian priors. interpretation and problems with construction and applications. *Preprint Statistics*, 3.

Verbeke, G. and Molenberghs, G. (2001). *Linear mixed models for longitudinal data*. Springer.

Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. (1999). The Analysis of Designed Experiments and Longitudinal Data by Using Smoothing Splines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):269–311.

Wang, X.-F. and Li, Y. (2014). Bayesian inferences for beta semiparametric-mixed models to analyze longitudinal neuroimaging data. *Biometrical Journal. Biometrische Zeitschrift*, 56(4):662–677.

Wu, H. and Zhang, J. (2006). *Nonparametric regression methods for longitudinal data analysis:*. Wiley-Interscience.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, 50(3):689–699.

Zeger, S. L. and Karim, M. R. (1991). Generalized Linear Models With Random Effects; A Gibbs Sampling Approach. *Journal of the American Statistical Association*, 86(413):79–86.