Evaluación de modelos mixtos para datos longitudinales biológicos:

Aplicación de NBMM, ZINBMM y LMM a la microbiota vaginal y al crecimiento de plantines

Memoria para optar al título profesional de
Ingeniera Civil Matemática

Gabriela Margarita Gutiérrez Bernal

Profesor guía

Cristian Meza Becerra
Instituto de Ingeniería Matemática, Universidad de Valparaíso

Profesora co-guía

Karine Bertin
Instituto de Ingeniería Matemática, Universidad de Valparaíso

Miembros de la comisión

Héctor Olivero Quinteros
Instituto de Ingeniería Matemática, Universidad de Valparaíso

INSTITUTO DE INGENIERÍA MATEMÁTICA, UNIVERSIDAD DE VALPARAÍSO
VALPARAÍSO, 14 DE AGOSTO DEL 2025

Agradecimientos

A mis padres; a mi madre Pamela, por su amor inagotable, por cuidar cada detalle y facilitarme el camino con entrega absoluta; y a mi padre Ricardo, por su esfuerzo constante y por darme incluso lo que no tenía, enseñándome el valor del trabajo y la perseverancia. A ambos, gracias por su sacrificio, dedicación y amor; son los principales responsables de que mis sueños y metas hayan sido posibles.
A mis abuelos, por apoyarme siempre desde sus posibilidades; con sus herramientas y esfuerzos contribuyeron, junto a mis padres, a que este logro fuera posible.
A mi hermana Javiera, que en su silencio me acompañó en innumerables ocasiones, brindándome esas recargas de energía cuando más las necesitaba. Su compañía y apoyo fueron un aliento importante en aquellas largas noches de estudio.
A Marco, por acompañarme y sostenerme en esta última etapa, brindándome su compañía, apoyo y momentos de alegría cuando más los necesitaba. Gracias por su disposición incondicional para ayudarme y por el equilibrio que su presencia aportó en este camino.
A mis amigas de la universidad, por acompañarme en este camino y ser un pilar fundamental durante este proceso. Aprendí mucho de cada una de ustedes, y su compañía junto a sus risas, tanto en lo académico como en lo personal, fueron un apoyo importante para atravesar las distintas etapas de este proceso.
Al profesor Héctor, por todo lo que me enseñó, no solo en el ámbito académico, sino también en lo personal. Fue un gran mentor a lo largo de este proceso, y le agradezco profundamente por su guía constante y por siempre creer en mí.
A mis profesores guías, Cristian y Karine, por acompañarme en este proyecto con compromiso y claridad en sus orientaciones. Fue una experiencia muy valiosa trabajar junto a ustedes, y agradezco que me hayan permitido culminar esta etapa con una vivencia tan grata como enriquecedora.
A todos los profesores que contribuyeron a mi formación académica, por guiarme y prepararme con dedicación, paciencia y el compromiso que pusieron en cada una de sus clases.
A la profesora Romina y a su equipo, por su colaboración y disposición para contribuir al desarrollo de uno de los estudios comprendidos en este trabajo.
Al proyecto FONDECYT Regular N.º 1221373, “Adaptive estimation in non-parametric models”.

Resumen

Este estudio aborda la modelación de dos conjuntos de datos longitudinales biológicos con desafíos estadísticos complementarios. El primer estudio, que llamaremos **Estudio I** analiza la dinámica temporal de la microbiota vaginal en mujeres gestantes y no gestantes mediante modelos mixtos binomiales negativos (NBMM) y su extensión inflada en ceros (ZINBMM), estructuras que permiten manejar simultáneamente sobredispersión, inflación de ceros y correlación intra-sujeto. El segundo estudio, que llamaremos **Estudio II** evalúa el crecimiento de *Quillaja saponaria* en vivero a partir de siete mediciones sucesivas del diámetro del cuello (DAC), utilizando datos reales obtenidos en colaboración con el equipo académico de la Escuela de Ingeniería Ambiental de la Universidad de Valparaíso. Estos datos fueron recolectados en el marco de un proyecto de investigación aplicado en el Jardín Botánico Nacional de Viña del Mar. Se comparan diversos modelos mixtos lineales (LMM) jerárquicos con efectos aleatorios y covariables ambientales acumuladas, con el objetivo de describir con precisión la variabilidad interindividual en el crecimiento de los plantines.

En ambos casos se sigue un flujo reproducible de limpieza de datos, exploración gráfica, ajuste realizado en `R` y selección mediante AIC, BIC y pruebas de razón de verosimilitud. Los resultados confirman que los NBMM/ZINBMM capturan adecuadamente los patrones de abundancia bacteriana asociados al estado gestacional, mientras que los LMM con pendientes aleatorias y temperatura acumulada describen con mayor precisión la variación inter-planta.

El trabajo aporta (i) una guía metodológica unificada para elegir modelos mixtos según la naturaleza de la respuesta, (ii) códigos reproducibles en R que integran paquetes especializados, y (iii) recomendaciones sobre diseño temporal y tratamiento de ceros para futuros estudios de microbiota y crecimiento vegetal.

Índice general

Capítulo 1

Introducción general

Los **datos longitudinales** corresponden a mediciones repetidas realizadas sobre las mismas unidades de estudio (personas, animales, plantas u otros organismos) a lo largo del tiempo. A diferencia de los estudios transversales, que únicamente permiten observar el estado del proceso en un punto fijo, los datos longitudinales permiten seguir la evolución de un proceso y describir sus trayectorias temporales. Esto los convierte en una herramienta clave para entender dinámicas biológicas y evaluar cambios intra e inter-individuos en el tiempo.

El estudio de datos longitudinales se ha convertido en una piedra angular de la investigación aplicada en ciencias de la vida, ya que permite describir y predecir trayectorias temporales en procesos biológicos que difícilmente podrían captarse con diseños transversales. Cuando las mediciones se repiten en el mismo individuo —sea una persona, un animal o una planta— surgen dependencias seriales y heterogeneidades entre sujetos que exigen modelos estadísticos capaces de representar simultáneamente la dinámica temporal y la variabilidad individual. En particular, las herramientas de *modelos mixtos* y sus extensiones para conteos dispersos han demostrado ser esenciales para abordar este tipo de datos, ofreciendo un marco unificado que combina efectos fijos (tendencias poblacionales) y efectos aleatorios (variabilidad específica).

La presente investigación se sitúa en ese contexto metodológico y aporta dos aplicaciones contrastantes pero complementarias:

1. El análisis longitudinal de la **microbiota vaginal** en mujeres gestantes y no gestantes, donde los conteos bacterianos presentan sobredispersión y abundancia de ceros.

2. El modelado del **crecimiento de plantines** de *Quillaja saponaria* bajo cuatro tratamientos bioestimulantes, con registros sucesivos de diámetro y altura junto a covariables ambientales acumuladas.

En ambos casos se persigue un mismo hilo conductor: identificar la estructura estadística que describe mejor la evolución temporal del fenómeno y, a partir de ella, derivar inferencias biológicamente interpretables y/o predicciones confiables. Sin embargo, las particularidades de cada conjunto de datos —inflación de ceros en la microbiota y posible no linealidad en el crecimiento de los plantines— obligan a escoger estrategias de modelación diferenciadas. En el caso del estudio de la microbiota, se emplean

modelos de conteo adaptados a la sobredispersión y a la presencia de ceros excesivos, tales como el Modelo Mixto Binomial Negativo (NBMM) y su extensión con Inflación de Ceros (ZINBMM). En este contexto, la *inflación de ceros* hace referencia a que la frecuencia de ceros observada en los datos es mucho mayor de la que podría explicarse únicamente por un modelo de conteo estándar (Poisson o binomial negativo). Es importante señalar que ambos modelos forman parte del marco general de los Modelos Mixtos Lineales Generalizados (GLMM), los cuales permiten abordar variables de respuesta no gaussianas incorporando simultáneamente efectos fijos y aleatorios. En el caso del crecimiento de plantines, se considera la aplicación de Modelos Mixtos Lineales (LMM) y modelos mixtos no lineales (NLMM) para capturar la dinámica temporal y las posibles curvas de crecimiento no lineales. Esta diversidad metodológica permite discutir críticamente los alcances y limitaciones de cada enfoque en función de la naturaleza de los datos analizados.

Este capítulo introductorio se organiza del siguiente modo. En la Sección §1.1 se detalla la relevancia de modelar datos longitudinales biológicas y los desafíos estadísticos que plantean. La Sección §1.2 enuncia los objetivos globales de la investigación, tanto metodológicos como aplicados. Finalmente, la Sección §1.3 expone la arquitectura del documento, indicando el contenido de cada capítulo y la lógica de su encadenamiento.

1.1. Motivación: importancia de modelar datos longitudinales biológicos

Los fenómenos biológicos raramente permanecen estáticos en el tiempo. Desde los cambios poblacionales de microorganismos que colonizan un ecosistema hasta el crecimiento de tejidos vegetales sometidos a tratamientos silvícolas, los procesos vitales se caracterizan por una evolución continua que sólo puede capturarse mediante observaciones repetidas. Sin embargo, la simple acumulación de datos longitudinales no garantiza conocimiento: es indispensable dotarse de un marco analítico que *explique*, *cuantifique* y *prediga* dichas trayectorias con rigurosidad estadística.

Complejidad inherente a los datos biológicos.

- **Estructura jerárquica**. Las mediciones tomadas en un mismo individuo (mujer o planta) están naturalmente correlacionadas; ignorar esta dependencia conduce a varianzas subestimadas y a inferencias engañosas.

- **Variabilidad entre sujetos**. Diferencias genéticas, ambientales o de manejo provocan que cada unidad experimente su propia *trayectoria de base*. Los modelos mixtos permiten describir esta heterogeneidad mediante *efectos aleatorios* que conviven con tendencias poblacionales (efectos fijos).

- **Sobredispersión y ceros excesivos**. En microbiota, los conteos suelen presentar varianzas mayores que la media y una abundancia de ceros que reflejan ausencias reales o límites de detección.

Necesidad de métodos longitudinales avanzados. Los modelos lineales mixtos (LMM), los modelos binomiales negativos (NBMM) y sus extensiones infladas en ceros (ZINBMM) proporcionan un marco flexible que:

1. ajusta simultáneamente *tendencias globales* y *desvíos individuales*;
2. incorpora covariables dinámicas (tiempo de gestación, temperatura acumulada, tratamiento experimental);
3. permite *predicciones condicionales por individuo*, es decir, proyecciones de la trayectoria futura de un sujeto específico condicionadas a sus observaciones previas. Esto resulta esencial para el *monitoreo personalizado* (por ejemplo, seguir la abundancia de ciertas bacterias en grupos de mujeres gestantes y no gestantes y contrastar sus diferencias en el tiempo) o para el *manejo adaptativo* (al permitir estudiar la curva de crecimiento de cada plantín sometido a un tratamiento bioestimulante y comprender cómo evoluciona su respuesta a lo largo del tiempo).

En síntesis, modelar datos longitudinales biológicas no sólo aporta conocimiento básico sobre los mecanismos de cambio, sino que habilita herramientas predictivas críticas para la salud pública, la producción agrícola y la conservación del medioambiente. Esta investigación se sitúa en ese cruce, explorando y comparando modelos mixtos capaces de capturar la complejidad inherente a dos conjuntos de datos representativos: la microbiota vaginal humana y el crecimiento de plantines de *Quillaja saponaria*.

1.2. Objetivos globales

El propósito central de este trabajo es integrar, en un marco metodológico coherente, las herramientas estadísticas necesarias para modelar datos biológicos longitudinales que presentan dos características contrastantes: (i) *conteos sobre-dispersos e inflados en ceros*, propios de la microbiota vaginal, y (ii) *mediciones continuas correlacionadas*, típicas del crecimiento de plantines. Bajo esta premisa se plantean los siguientes objetivos:

- **Objetivo general.**
  Diseñar, implementar y evaluar una estrategia unificada de modelación longitudinal basada en *modelos mixtos* que sea capaz de:
  (a) capturar la estructura de correlación intra-sujeto o intra-planta,
  (b) acomodar la sobredispersión y la inflación de ceros cuando corresponda, y
  (c) ofrecer criterios comparables de ajuste y predicción.

- **Objetivos específicos.**
  1. *Desarrollar una implementación reproducible* en `R` que combine los paquetes `NBZIMM`, `nlme` y `lme4`, documentando los flujos de trabajo de pre-procesamiento, ajuste, diagnóstico y visualización.

2. *Aplicar la metodología* a dos estudios de caso:
   (I) la evolución longitudinal de la microbiota vaginal en mujeres gestantes y no gestantes, y
   (II) el crecimiento de *Quillaja saponaria* bajo cuatro tratamientos de biofertilización.
3. *Comparar rigurosamente* los modelos propuestos mediante criterios de información (AIC, BIC), y pruebas de razón de verosimilitud, para determinar la opción con mejor capacidad predictiva en cada contexto.
4. *Extraer conclusiones biológicas y prácticas* que vinculen los hallazgos estadísticos con:
   (a) la estabilidad temporal de taxones protectores durante el estado gestacional, y
   (b) la influencia de variables ambientales y tratamientos en la dinámica del diámetro del cuello de los plantines.

1.3. Estructura

Este manuscrito se organiza en seis capítulos. A continuación se ofrece una visión global de cada capítulo, destacando su contenido principal y la relación con los objetivos planteados.

1. **Marco teórico y estado del arte**. Revisa la literatura sobre datos longitudinales del microbioma y del crecimiento vegetal, así como los enfoques de modelación empleados (LMM, NLMM, GLMM). Se enfatizan los retos estadísticos específicos—sobredispersión, inflación de ceros y correlación intra–sujeto—y la evidencia previa que motiva los estudios empíricos de la investigación.

2. **Metodología general**. Detalla el flujo de trabajo utilizado: limpieza, exploración, ajuste, validación y visualización de modelos. Se documentan las herramientas de `RStudio` (`nlme`, `lme4`, `NBZIMM`) y los criterios de comparación (AIC, BIC, diagnósticos gráficos, ...).

3. **Estudio I – Microbiota vaginal**. Presenta los objetivos, el diseño de muestreo y la estructura de los datos 16S; expone la formulación, el ajuste y el diagnóstico de los modelos NBMM y ZINBMM; muestra los resultados con mapas de calor y discute las conclusiones parciales sobre la estabilidad microbiana durante la gestación.

4. **Estudio II – Crecimiento longitudinal de plantines**. Describe el experimento de biofertilización, la depuración de la base de datos y la exploración inicial de trayectorias y supervivencia. Se comparan ocho especificaciones de LMM y se discuten sus limitaciones y alcances en ausencia de evidencia para modelos no lineales mixtos.

5. **Conclusiones generales**. Integra los hallazgos prácticos, responde a los objetivos globales de la investigación y plantea desafíos pendientes y líneas de trabajo futuro.

Capítulo 2

Modelos mixtos aplicados a datos longitudinales biológicos

El análisis estadístico de datos longitudinales en biología ha cobrado gran relevancia en los últimos años, impulsado por el desarrollo de técnicas de medición de alta resolución y por la creciente disponibilidad de datos observacionales complejos. En particular, estudios que monitorean a lo largo del tiempo la composición microbiana en humanos o el crecimiento de especies vegetales requieren herramientas metodológicas que capturen tanto la estructura temporal de las observaciones como la heterogeneidad entre individuos o unidades experimentales.

Este capítulo presenta los fundamentos teóricos que sustentan el enfoque adoptado en esta investigación, articulando dos ejes centrales: por un lado, la caracterización de los **datos longitudinales** propios del microbioma vaginal y del crecimiento de plántulas; por otro, la revisión de **modelos estadísticos mixtos** que permiten abordar de forma robusta los desafíos inherentes a este tipo de datos. Se detallan a continuación las propiedades estadísticas distintivas de cada dominio, así como los retos comunes que justifican el uso de estructuras jerárquicas con efectos aleatorios y covariables dinámicas. A partir de esta base, se introducen los modelos mixtos lineales (LMM), no lineales (NLMM), generalizados (GLMM) y binomiales negativos (NBMM) con su extensión inflados en ceros (ZINBMM), describiendo su formulación, supuestos, métodos de estimación y aplicaciones en contextos similares.

2.1. Datos longitudinales del microbioma y del crecimiento vegetal

El hilo conductor de esta investigación es el análisis de *datos longitudinales* generados en dos dominios biológicos—la **microbiota vaginal humana** y el **crecimiento de plántulas en ambientes controlados**—que comparten una estructura jerárquica de observaciones repetidas en el tiempo. Sea

$$\mathbf{Y}_i = \left(Y_{i1}, \ldots, Y_{in_i}\right)^\top, \qquad i = 1, \ldots, n,$$

el vector de respuestas observadas en la unidad experimental i (ya sea una mujer o una planta),

correspondientes a los tiempos de medición $t_{ij} \subset \mathbb{R}_{\geq 0}$, con $j = 1, \ldots, n_i$. Cada componente Y_{ij} puede representar

$$Y_{ij} \begin{cases} \in \mathbb{N}_0, & \text{conteos 16S de un taxón bacteriano,} \\ \in \mathbb{R}, & \text{altura o diámetro de cuello de una plántula.} \end{cases} \tag{2.1}$$

En el caso del estudio de la microbiota, los datos de conteo utilizados provienen de la secuenciación del gen **16S rRNA**, un componente presente en todas las bacterias que permite su identificación. Este gen contiene regiones que son muy similares entre especies (conservadas), pero también otras que varían (hipervariables), lo que permite distinguir distintos tipos de bacterias. Mediante técnicas modernas de secuenciación, como la secuenciación de las regiones V1–V3 del gen 16S, es posible estimar la abundancia relativa de distintas bacterias en una muestra, sin necesidad de cultivarlas en laboratorio. Esta estrategia fue utilizada para caracterizar la composición de la microbiota vaginal en mujeres gestantes y no gestantes, a lo largo del tiempo (Romero et al., 2014).

2.2. Desafíos estadísticos en microbiota y crecimiento vegetal

Los conjuntos de datos analizados en esta investigación presentan una estructura longitudinal con observaciones repetidas en individuos (ya sean mujeres o plantas) a lo largo del tiempo. En biología, un *taxón* (plural: *taxones*) corresponde a una unidad de clasificación que agrupa organismos con características comunes, como una especie, un género o una familia. En el caso de la microbiota vaginal humana, los datos corresponden a perfiles microbianos obtenidos mediante secuenciación de amplicones 16S, que permiten estimar la abundancia relativa o absoluta de diferentes taxones bacterianos para cada participante, en distintos momentos del tiempo (Romero et al., 2014). Estas abundancias pueden expresarse en distintos niveles taxonómicos—como filo, clase, orden, familia, género o especie—según el grado de resolución requerido por el análisis. Las observaciones se estructuran comúnmente en lo que se conoce como tablas OTU (Operational Taxonomic Units) o ASV (Amplicon Sequence Variants), matrices que contienen el número de lecturas asociadas a cada taxón en cada muestra, y que constituyen la base para el procesamiento estadístico posterior.

Este tipo de datos plantea varios desafíos metodológicos. En primer lugar, se trata de *conteos discretos y sobredispersos*, cuya varianza suele exceder la media, lo que hace que la distribución de Poisson simple resulte inadecuada (Chen and Li, 2016). En segundo lugar, existe una considerable *inflación de ceros* en las tablas OTU/ASV, atribuible tanto a la ausencia biológica real de ciertos taxones como a limitaciones técnicas de detección. Estas características justifican el uso de modelos con componentes inflados en ceros, como los modelos binomiales negativos cero-inflados (ZINBMM), que han sido utilizados exitosamente en estudios recientes (Zhang et al., 2020; Zhang and Yi, 2020). Además, las abundancias relativas cumplen con una restricción de suma constante, lo que induce correlaciones espurias entre taxones: un aumento en la proporción de un grupo implica una disminución en otro, aun si no existe una relación biológica directa entre ellos (fenómeno conocido como *composicionalidad*)(Zhang et al., 2018). Finalmente, la estructura longitudinal del estudio implica que cada participante es medida en múltiples ocasiones, lo que genera *correlación intra-sujeto* y requiere especificar estructuras de

covarianza adecuadas para evitar inferencias sesgadas.

Por otra parte, los datos relacionados con el crecimiento de *Quillaja saponaria* en vivero también presentan una naturaleza longitudinal, con mediciones repetidas de altura y diámetro del cuello (DAC) en distintas fechas para cada plántula. Estos datos exhiben trayectorias no lineales en función del tiempo: el crecimiento sigue patrones sigmoides o leyes de potencias, en las que la *tasa de crecimiento relativo* (RGR) disminuye conforme aumenta el tamaño, como ha sido reportado en estudios de crecimiento vegetal (Turnbull et al., 2012). También se presenta *heterocedasticidad*, ya que la varianza tiende a incrementarse con la biomasa acumulada. Para abordar esto, se han propuesto transformaciones logarítmicas o funciones de varianza explícitas en los modelos de regresión no lineal (Ritz and Streibig, 2008). Además, se deben considerar *efectos aleatorios jerárquicos*, dado que las plantas se encuentran anidadas en tratamientos experimentales específicos y pueden estar expuestas a distintas condiciones ambientales. Otro aspecto relevante es el *diseño temporal del muestreo*: la literatura sugiere que resulta más eficaz realizar más mediciones a lo largo del tiempo con menos réplicas por fecha, que concentrar muchas observaciones en pocas fechas, sobre todo en las etapas iniciales del crecimiento donde los cambios son más pronunciados (Turnbull et al., 2012).

A pesar de tratarse de dos sistemas biológicos muy diferentes — por un lado, la **microbiota vaginal humana**, y por otro, el **crecimiento de plántulas de *Quillaja saponaria*** en vivero — ambos comparten características estadísticas estructurales que permiten abordar su análisis con un enfoque metodológico común. En particular, en ambos contextos se trabaja con datos longitudinales, donde se realizan mediciones repetidas sobre los mismos individuos a lo largo del tiempo, lo que motiva el uso de *modelos mixtos jerárquicos*.

Estos modelos permiten: (i) modelar la *correlación temporal* entre observaciones sucesivas de un mismo sujeto (mujer o planta), (ii) capturar la *heterogeneidad interindividual* mediante interceptos y pendientes aleatorias, y (iii) combinar la influencia de *covariables fijas y aleatorias* de manera integrada. En su forma general, un modelo mixto se representa mediante la estructura:

$$g(\mathbb{E}[Y_{ij} \mid \mathbf{b}_i]) = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \mathbf{b}_i,$$

donde $g(\cdot)$ es una función de enlace adecuada al tipo de respuesta, $\mathbf{x}_{ij}$ y $\mathbf{z}_{ij}$ son vectores de covariables fijas y aleatorias respectivamente, y los efectos aleatorios $\mathbf{b}_i$ se modelan como realizaciones de una normal multivariada centrada, $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$.

En síntesis, tanto los datos de microbiota como los de crecimiento vegetal comparten una estructura longitudinal y jerárquica que no puede analizarse de manera adecuada con modelos estadísticos simples. Esta situación motiva el uso de *modelos mixtos*, que permiten representar de forma simultánea la evolución en el tiempo y las diferencias entre individuos. Dependiendo de la naturaleza de la variable de interés, es necesario recurrir a distintas extensiones: en el caso de respuestas continuas como la altura o el diámetro de los plantines, se utilizan **modelos lineales mixtos** (LMM) o sus versiones **no lineales** (NLMM) cuando la trayectoria sigue curvas de crecimiento más complejas; en el caso de datos de conteo como los de microbiota, se requieren **modelos mixtos generalizados** (GLMM) que incorporan distribuciones distintas de la normal, como la *binomial negativa*, y en presencia de muchos ceros, versiones **infladas en ceros** (ZINBMM). De este modo, cada tipo de modelo se ajusta a las particularidades estadísticas del conjunto de datos, ofreciendo un marco coherente y flexible para su

análisis.

2.3. Modelos mixtos lineales (LMM)

2.3.1. Motivación y formulación

Los **modelos mixtos lineales** (LMM) extienden el modelo lineal clásico para situaciones en las que las observaciones no son independientes, sino que están *agrupadas*, como ocurre en estudios con mediciones repetidas en el mismo individuo o unidad experimental. Al introducir *efectos aleatorios*, estos modelos permiten capturar explícitamente la correlación intra-grupo, diferenciando la variabilidad atribuible a diferencias entre grupos (entre-sujetos) de aquella presente dentro de cada grupo (intra-sujetos) (Correa Morales and Salazar Uribe, 2016).

En términos generales, un LMM se compone de tres elementos fundamentales. En primer lugar, los *efectos fijos* $\boldsymbol{\beta}$, que representan parámetros de interés poblacional y describen la tendencia promedio de la respuesta en función de covariables como tratamiento, tiempo o temperatura. En segundo lugar, los *efectos aleatorios* $\mathbf{b}_i$, que capturan la desviación específica de cada unidad i respecto de dicha tendencia poblacional. Modelar interceptos y/o pendientes aleatorias permite que cada individuo tenga su propia línea o trayectoria. Finalmente, los *errores* ε_{ij} recogen la variabilidad residual no explicada por los efectos anteriores.

Sea $i = 1, \ldots, n$ el índice de la unidad experimental (sujeto o planta) y $j = 1, \ldots, n_i$ el índice de medición en el tiempo t_{ij}. La formulación básica del LMM a nivel individual se expresa como:

$$Y_{ij} \;=\; \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} \;+\; \mathbf{z}_{ij}^{\top}\mathbf{b}_i \;+\; \varepsilon_{ij}, \qquad \mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D}), \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \tag{2.2}$$

Aquí:

- Y_{ij} es la respuesta observada para la unidad i en el tiempo t_{ij},
- $\mathbf{x}_{ij} \in \mathbb{R}^p$ es el vector de covariables asociadas a los efectos fijos,
- $\mathbf{z}_{ij} \in \mathbb{R}^q$ es el vector de covariables asociadas a los efectos aleatorios,
- $\boldsymbol{\beta} \in \mathbb{R}^p$ es el vector de coeficientes de efectos fijos,
- $\mathbf{b}_i \in \mathbb{R}^q$ es el vector de efectos aleatorios específicos de la unidad i, con matriz de covarianza $\mathbf{D} \in \mathbb{R}^{q \times q}$,
- ε_{ij} es el error aleatorio asociado a la observación ij, con varianza σ^2.

Al considerar el conjunto completo de observaciones (con $N = \sum_{i=1}^{n} n_i$), el modelo puede escribirse en forma matricial como:

$$\mathbf{Y} \;=\; \mathbf{X}\boldsymbol{\beta} \;+\; \mathbf{Z}\mathbf{b} \;+\; \boldsymbol{\varepsilon}, \qquad \mathbf{b} \sim \mathcal{N}\big(\mathbf{0},\, \mathbf{I}_n \otimes \mathbf{D}\big), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_N). \tag{2.3}$$

Donde:

- $\mathbf{Y} \in \mathbb{R}^N$: vector columna que concatena todas las respuestas observadas,

- $\mathbf{X} \in \mathbb{R}^{N \times p}$: matriz de diseño para los efectos fijos,
- $\mathbf{Z} \in \mathbb{R}^{N \times nq}$: matriz de diseño para los efectos aleatorios (estructurada por bloques de cada individuo),
- $\mathbf{b} \in \mathbb{R}^{nq}$: vector que agrupa los efectos aleatorios de todos los individuos,
- $\boldsymbol{\varepsilon} \in \mathbb{R}^{N}$: vector de errores aleatorios,
- $\mathbf{I}_n \otimes \mathbf{D}$: matriz de covarianza en bloque para los efectos aleatorios; el símbolo $\otimes$ denota el producto de Kronecker,
- $\mathbf{I}_N$: matriz identidad de orden N,
- σ^2: varianza residual común a todas las observaciones.

El uso del producto de Kronecker $\mathbf{I}_n \otimes \mathbf{D}$ expresa que cada unidad experimental i tiene su propio efecto aleatorio $\mathbf{b}_i \sim \mathcal{N}(0, \mathbf{D})$, y que los efectos aleatorios de distintas unidades son independientes entre sí (Correa Morales and Salazar Uribe, 2016).

De este modo, se tiene que $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, siendo la matriz de varianza-covarianza marginal

$$\mathbf{V} = \mathbf{Z}(\mathbf{I}_n \otimes \mathbf{D})\mathbf{Z}^\top + \sigma^2 \mathbf{I}_N.$$

Aquí, $\mathbf{V} \in \mathbb{R}^{N \times N}$ representa la matriz de varianza-covarianza marginal de las observaciones, que incorpora tanto la estructura de dependencia entre mediciones de una misma unidad, como la varianza residual común. Este modelo asume que los errores $\boldsymbol{\varepsilon}$ son independientes entre sí y también independientes de los efectos aleatorios $\mathbf{b}$, lo cual es una hipótesis esencial para que las estimaciones de máxima verosimilitud sean consistentes.

Ejemplo ilustrativo. Para visualizar el modelo en un caso simple, suponiendo que se estudia el *aumento de altura en plántulas de* Quillaja saponaria a lo largo del tiempo. Considerando tres fechas de medición ($t = 30, 60, 90$ días) y un solo predictor lineal: el tiempo. Un modelo mixto lineal puede escribirse como

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij},$$

donde:

- β_0: altura promedio inicial de las plántulas en la población,
- β_1: incremento promedio de altura por día,
- b_{0i}: desviación específica de la plántula i respecto a la altura inicial promedio,
- b_{1i}: desviación específica de la plántula i respecto a la tasa de crecimiento promedio,
- ε_{ij}: error residual para la observación ij.

Por ejemplo, si los valores estimados fueran $\beta_0 = 5$ cm, $\beta_1 = 0{,}1$ cm/día, y para una plántula particular se obtiene $b_{0i} = +1$ cm y $b_{1i} = -0{,}02$ cm/día, entonces la trayectoria predicha sería:

$$Y_{ij} = (5+1) + (0{,}1 - 0{,}02)t_{ij} + \varepsilon_{ij} = 6 + 0{,}08\, t_{ij} + \varepsilon_{ij}.$$

Esto significa que esta plántula comienza 1 cm por encima del promedio poblacional, pero crece a una velocidad ligeramente menor que el promedio.

2.3.2. Supuestos y propiedades

El modelo mixto lineal se basa en varios supuestos fundamentales que garantizan la validez de las inferencias. En particular, se asume que:

- Los **efectos aleatorios** $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$ son independientes entre sí, con $\mathbf{D}$ una matriz definida positiva.
- Los **errores** $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ son independientes entre sí, con varianza constante $\sigma^2 > 0$, y además son *independientes* de los efectos aleatorios $\mathbf{b}_i$.
- La **estructura del modelo es lineal** respecto de los efectos fijos $\boldsymbol{\beta}$ y los efectos aleatorios $\mathbf{b}_i$, es decir, la respuesta se modela como una combinación lineal de ambos tipos de efectos.

Bajo estos supuestos, los efectos fijos $\boldsymbol{\beta}$ permiten estimar relaciones poblacionales como la diferencia promedio entre tratamientos o la tendencia general en el tiempo. Por otro lado, los *componentes de varianza* $(\mathbf{D}, \sigma^2)$ cuantifican la proporción de variabilidad explicada por la heterogeneidad entre unidades (por ejemplo, entre plantas o sujetos) y por la variación residual.

2.3.3. Estimación y ajuste

La estimación de los parámetros $\boldsymbol{\beta}$, $\mathbf{D}$, σ^2 puede realizarse mediante dos enfoques principales. La *máxima verosimilitud* (ML) se basa en optimizar la **log-verosimilitud marginal** del modelo, es decir, la función de verosimilitud que resulta de integrar los efectos aleatorios fuera del modelo condicional. Formalmente, esta log-verosimilitud se define como:

$$\ell(\boldsymbol{\beta}, \mathbf{D}, \sigma^2) = \log f(\mathbf{Y} \mid \boldsymbol{\beta}, \mathbf{D}, \sigma^2) = \log \int f(\mathbf{Y} \mid \mathbf{b}, \boldsymbol{\beta}, \sigma^2)\, f(\mathbf{b} \mid \mathbf{D})\, d\mathbf{b}.$$

Esta formulación permite comparar modelos que difieren en los efectos fijos, ya que evalúa la probabilidad marginal de los datos observados sin condicionar en valores específicos de los efectos aleatorios.

En cambio, la *verosimilitud restringida* (REML) modifica este enfoque eliminando $p = \dim(\boldsymbol{\beta})$ grados de libertad para corregir el sesgo en la estimación de las varianzas, lo que mejora la precisión de los estimadores de $\mathbf{D}$ y σ^2.

En la práctica, los paquetes `nlme` y `lme4` implementan algoritmos híbridos de optimización que combinan robustez y eficiencia. Estos algoritmos suelen combinar el método EM (Expectation–Maximization)

y el enfoque de Newton–Raphson o Fisher-Scoring. El algoritmo EM (Dempster et al., 1977) es robusto, aunque su convergencia es lineal, mientras que Newton–Raphson/Fisher-Scoring (M.J. et al., 1988) utiliza gradientes y matrices hessianas para lograr una convergencia cuadrática más rápida.

2.3.4. Evaluación y diagnóstico del modelo

Para evaluar la calidad del ajuste en modelos mixtos, es habitual examinar los residuos marginales y condicionales, lo cual permite identificar posibles problemas de linealidad o heterocedasticidad. Asimismo, la comparación entre modelos anidados mediante pruebas de razón de verosimilitud bajo REML constituye una herramienta valiosa para evaluar la estructura de covarianza de los efectos aleatorios.

Por otro lado, cuando el interés está en la selección de covariables fijas o aleatorias, los criterios de información como el **Akaike Information Criterion (AIC)** y el **Bayesian Information Criterion (BIC)** resultan particularmente útiles. Ambos criterios combinan el ajuste del modelo (evaluado a través de la log-verosimilitud) con una penalización por el número de parámetros incluidos. De este modo, no solo se busca maximizar la log-verosimilitud —es decir, obtener un modelo que represente adecuadamente los datos—, sino también evitar la sobreparametrización.

De esta manera, en la práctica, un modelo con valores más bajos de AIC y BIC se interpreta como preferible, ya que indica un equilibrio más adecuado entre calidad de ajuste y parsimonia. Mientras que la log-verosimilitud refleja directamente qué tan bien se ajusta el modelo a los datos (se prefiere que sea lo más alta posible), los criterios de información penalizan la complejidad excesiva y, por tanto, favorecen modelos más simples que logren un buen compromiso entre ajuste y número de parámetros estimados.

2.3.5. Ventajas y limitaciones

Entre las principales ventajas de los LMM se encuentra la posibilidad de realizar *predicciones individuales* a partir de los llamados BLUPs (Best Linear Unbiased Predictors) y la descomposición de la variabilidad total en componentes atribuibles a distintos niveles jerárquicos. Además, los LMM pueden ajustarse incluso en presencia de datos desbalanceados o con observaciones faltantes, sin necesidad de imputación explícita.

Sin embargo, su validez depende de que se cumpla aproximadamente la suposición de normalidad. En casos donde la variable respuesta corresponde a conteos sobredispersos o se observa una inflación de ceros, como ocurre frecuentemente en estudios de microbiota, los LMM pueden ser inadecuados. En tales contextos, es preferible recurrir a modelos más flexibles como los GLMM, NBMM o ZINBMM, que serán presentados en las Secciones 2.5–2.6.

2.4. Modelos mixtos no lineales (NLMM)

2.4.1. Motivación

Muchos procesos biológicos —por ejemplo, el crecimiento sigmoide de las plántulas o la trayectoria de abundancia relativa de un taxón bacteriano dominante— presentan relaciones *intrínsecamente no*

lineales entre la respuesta y el tiempo (o la dosis, o la temperatura). Cuando además existen mediciones repetidas por unidad experimental, la forma natural de modelar la correlación intra-sujeto es introducir *efectos aleatorios*, dando lugar a los **modelos mixtos no lineales** (Lindstrom and Bates, 1990).

2.4.2. El modelo no lineal clásico

Un modelo se considera no lineal cuando al menos una de las derivadas de la función de esperanza con respecto a los parámetros depende de los propios parámetros (M.J. et al., 1988). Formalmente, se expresa como:

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

donde:

- y_i es la respuesta observada en la unidad i,
- $\mathbf{x}_i \in \mathbb{R}^d$ representa las variables independientes o covariables,
- $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ es una función no lineal en los parámetros $\boldsymbol{\theta} \in \mathbb{R}^p$,
- ε_i es el término de error aleatorio con varianza constante σ^2.

Ejemplo ilustrativo. Es importante aclarar que en los modelos lineales la linealidad se refiere a los coeficientes y no necesariamente a las variables explicativas. Por ejemplo, la siguiente función es cuadrática en x, pero sigue siendo lineal en los parámetros $\beta_0, \beta_1, \beta_2$:

$$E(y \mid x) = \beta_0 + \beta_1 x + \beta_2 x^2 \quad \text{(lineal en los coeficientes)}$$

En contraste, la siguiente expresión es no lineal en los coeficientes:

$$E(y \mid x) = \frac{1}{\beta_0 + \beta_1 x} \quad \text{(no lineal en los coeficientes)}$$

En contextos reales, suele ser necesario relajar el supuesto de homocedasticidad, permitiendo que la varianza del error dependa de las covariables:

$$\mathrm{Var}(\varepsilon_i) = \sigma^2 w(\mathbf{x}_i),$$

donde $w(\cdot)$ es una función conocida o estimable que modela la heterocedasticidad.

2.4.3. Extensión con efectos aleatorios

Cuando se dispone de múltiples mediciones por unidad experimental, se extiende el modelo anterior introduciendo efectos aleatorios. Para la unidad i ($i = 1, \ldots, n$), con n_i observaciones en los tiempos t_{ij}, se postula:

$$y_{ij} = f\big(t_{ij}, \boldsymbol{\phi}_i\big) + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \tag{2.4}$$

donde:

- $f(t, \boldsymbol{\phi}_i)$ es una función no lineal que describe la evolución de la respuesta en función del tiempo t y de los parámetros individuales $\boldsymbol{\phi}_i$,
- $\boldsymbol{\phi}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i$ representa la estructura paramétrica específica del individuo i,
- $\boldsymbol{\beta} \in \mathbb{R}^p$ son los parámetros fijos poblacionales,
- $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{D})$ son los efectos aleatorios del individuo i,
- $\mathbf{A}_i \in \mathbb{R}^{r\times p}$ y $\mathbf{B}_i \in \mathbb{R}^{r\times q}$ son matrices de diseño para los efectos fijos y aleatorios, respectivamente.

En este marco, cada parámetro ϕ_{ir} (por ejemplo: asíntota, tasa de crecimiento o punto de inflexión) puede depender de covariables comunes y de variación individual, permitiendo gran flexibilidad en la modelación de trayectorias específicas.

Agrupando todas las observaciones del conjunto de individuos, con $N = \sum_{i=1}^{n} n_i$, se obtiene:

$$\mathbf{Y} \mid \mathbf{b} \sim \mathcal{N}\left(\boldsymbol{\eta}(\boldsymbol{\phi}),\, \sigma^2\boldsymbol{\Delta}\right), \qquad \mathbf{b} \sim \mathcal{N}\left(\mathbf{0},\, \sigma^2(\mathbf{I}_n \otimes \mathbf{D})\right),$$

donde:

- $\mathbf{Y} \in \mathbb{R}^N$ es el vector de todas las respuestas observadas,
- $\boldsymbol{\phi} = [\boldsymbol{\phi}_1^\top, \ldots, \boldsymbol{\phi}_n^\top]^\top$ es el vector que agrupa los parámetros individuales,
- $\boldsymbol{\eta}(\boldsymbol{\phi}) = [f(t_{ij}, \boldsymbol{\phi}_i)]_{i=1,\ldots,n;\ j=1,\ldots,n_i}$, representa el vector de medias modeladas para todos los datos,
- $\boldsymbol{\Delta}$ es una matriz de covarianza que puede incorporar estructuras heterocedásticas o autocorreladas (por ejemplo, AR(1)),
- $\mathbf{I}_n \otimes \mathbf{D}$ es el producto de Kronecker entre la identidad y la matriz de varianza-covarianza de los efectos aleatorios.

Este marco general permite modelar fenómenos dinámicos complejos con variación entre unidades, como el crecimiento de organismos vivos, la evolución de biomarcadores o trayectorias clínicas longitudinales. Es el enfoque implementado por funciones como `nlme::nlme` y `lme4::nlmer` en R (Turnbull et al., 2012).

2.4.4. Estimación en la práctica

Los parámetros $(\boldsymbol{\beta}, \mathbf{D}, \sigma^2)$ se obtienen por máxima verosimilitud (ML) o verosimilitud restringida (REML). Dada la no linealidad, la optimización se apoya en *aproximaciones iterativas*:

PD–LME (Pseudo-Datos + LME). Propone linealizar f en cada iteración, resolver un LMM y actualizar los parámetros hasta convergencia (Lindstrom and Bates, 1990).

SAEM. Variante estocástica–EM popular en farmacocinética cuando el número de parámetros aleatorios es grande y la verosimilitud presenta varios modos (Correa Morales and Salazar Uribe, 2016).

En `R` estos métodos están implementados en `nlme::nlme`, `saemix` y `nlmer`. En esta memoria solo se presenta el esquema conceptual, pues la densidad temporal de nuestros datos (7 fechas) aún no justifica la complejidad computacional de un NLMM completo; véase la discusión en §5.4.3.

2.5. Modelos mixtos lineales generalizados (GLMM)

2.5.1. Motivación

Cuando la variable respuesta no es continua (p. ej. conteos, proporciones, binaria) pero las observaciones siguen una *estructura longitudinal* o jerárquica, los **modelos mixtos lineales generalizados** permiten combinar:

- la flexibilidad de la *familia exponencial* para la distribución condicional $Y_{ij} \mid \mathbf{b}_i$,
- la descomposición *efectos fijos + aleatorios* que captura la correlación intra-sujeto.

2.5.2. Formulación

Sea $\mathbf{Y} = (Y_1, \ldots, Y_N)^\top$ el vector de variables respuesta observadas, y sea $\mathbf{b} \in \mathbb{R}^q$ un vector de efectos aleatorios. Se asume que, condicionadas en $\mathbf{b}$, las observaciones $Y_i \mid \mathbf{b}$ son independientes y siguen una distribución perteneciente a la familia exponencial (Correa Morales and Salazar Uribe, 2016). La forma general de la función de densidad condicional es:

$$f_{Y_i|\mathbf{b}}(y_i \mid \mathbf{b}, \boldsymbol{\beta}, \phi) = \exp\left\{ \frac{y_i \eta_i - c(\eta_i)}{a(\phi)} + d(y_i, \phi) \right\}, \qquad \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}), \tag{2.5}$$

donde:

- y_i es el valor observado de la respuesta en la unidad i,
- $\eta_i \in \mathbb{R}$ es el **predictor lineal**, definido como:

  $$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \mathbf{b},$$

  siendo $\mathbf{x}_i$ y $\mathbf{z}_i$ las filas i-ésimas de las matrices de diseño $\mathbf{X}$ y $\mathbf{Z}$, respectivamente,
- $\boldsymbol{\beta} \in \mathbb{R}^p$ es el vector de efectos fijos,
- $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$ representa los efectos aleatorios con matriz de varianza-covarianza $\mathbf{D}$,
- ϕ es un **parámetro de dispersión**, no confundir con el parámetro $\boldsymbol{\phi}_i$ usado en la sección de modelos no lineales mixtos; en GLMM, ϕ típicamente controla la varianza condicional (por ejemplo, $\phi = 1$ en Poisson o binomial),

- $a(\phi)$, $c(\eta_i)$, y $d(y_i, \phi)$ son funciones características de la familia exponencial que definen su forma específica.

Entre las distribuciones más comunes dentro de esta familia se encuentran:

- La **normal**, asociada al modelo lineal mixto clásico (LMM),
- La **binomial**, asociada al modelo logístico (respuesta dicotómica),
- La **Poisson**, adecuada para modelar datos de conteo.

En los modelos lineales generalizados mixtos, el predictor lineal, denotado como η, se relaciona directamente con la esperanza condicional de los datos (Meza, 2021).
La relación entre η y la variable de respuesta se establece mediante una función de enlace $g(\cdot)$ y su inversa $h(\cdot)$, tal que:

$$g(\cdot) = \text{función de enlace},$$
$$h(\cdot) = g^{-1}(\cdot) = \text{inversa de la función de enlace}.$$

Por lo tanto, la esperanza de y satisface:

$$g(E[y]) = \eta.$$

De manera equivalente, se tiene que:

$$E[y] = h(\eta) = \mu,$$

donde μ denota la media condicional de la distribución de la variable respuesta. Además, la variable de respuesta también puede expresarse como:

$$y = h(\eta) + \epsilon,$$

donde ϵ representa el término de error.
La elección de la función de enlace depende del tipo de datos que se esté modelando. Existen diferentes funciones de enlace dependiendo de si los datos son continuos, binarios, de conteo, entre otros. A continuación se presenta una tabla (Tabla 2.1) que resume las funciones de enlace más utilizadas y sus correspondientes inversas

Dependiendo de la distribución de la variable respuesta, la función de enlace por defecto puede variar. La siguiente tabla (Tabla 2.2) describe las familias de distribuciones comunes, las funciones de enlace más frecuentemente utilizadas y otras posibles opciones :

Estas tablas sirven como guía para seleccionar $g(\cdot)$ y la familia adecuada al tipo de datos a analizar.

2.5.3. Estimación

La verosimilitud marginal de un GLMM se obtiene integrando los efectos aleatorios fuera de la verosimilitud condicional:

Enlace g	$\eta = g(\mu)$	$\mu = g^{-1}(\eta)$
Identidad	μ	η
Log	$\log \mu$	e^{η}
Inversa	μ^{-1}	η^{-1}
Raíz cuadrada	$\sqrt{\mu}$	η^2
Logit	$\log \frac{\mu}{1-\mu}$	$\frac{1}{1+e^{-\eta}}$
Probit	$\Phi^{-1}(\mu)$	$\Phi(\eta)$
cloglog	$\log\big[-\log(1-\mu)\big]$	$1 - e^{-e^{\eta}}$

Tabla 2.1: Enlaces comunes y sus inversas.

Familia	Enlace por defecto	Soporte Y	Var(Y)	Otros enlaces
Gaussiana	Identidad	$(-\infty, \infty)$	ϕ	Log
Binomial	Logit	$0, \ldots, N$	$\mu(1-\mu)/N$	Probit, cloglog
Poisson	Log	$0, 1, 2, \ldots$	μ	Identidad, raíz
Gamma	Inversa	$(0, \infty)$	$\phi\mu^2$	Log, identidad

Tabla 2.2: Familias, enlaces y varianzas condicionales típicas.

$$L(\boldsymbol{\beta}, \phi, \mathbf{D} \mid \mathbf{y}) \;=\; \prod_{i=1}^{m} \int_{\mathbb{R}^q} \Big[\prod_{j=1}^{n_i} f_{Y_{ij}\mid \mathbf{b}_i}\big(y_{ij} \mid \mathbf{b}_i; \boldsymbol{\beta}, \phi\big) \Big] \, \varphi_q(\mathbf{b}_i; \mathbf{0}, \mathbf{D}) \, d\mathbf{b}_i, \tag{2.6}$$

donde m es el número de sujetos, q la dimensión de $\mathbf{b}_i$, $f_{Y_{ij}\mid \mathbf{b}_i}$ la densidad condicional de la familia exponencial elegida y $\varphi_q(\cdot; \mathbf{0}, \mathbf{D})$ la densidad normal q-variante con media nula y matriz de covarianza $\mathbf{D}$.

La verosimilitud marginal no admite una forma cerrada salvo para distribuciones normales. En la práctica se recurre a *aproximaciones numéricas* que equilibran precisión y costo computacional:

- **Laplace.** Expande la integral de (2.6) en torno al modo de la densidad condicional de los efectos aleatorios, logrando una precisión $\mathcal{O}(p^{-1})$ con una sola evaluación Hessiana.
- **Cuadratura adaptativa de Gauss–Hermite (AGQ).** Refina Laplace mediante K nodos por dimensión a costa $O(K^q)$ evaluaciones de densidad.
- **Linearización de primer orden (PQL).** Linealiza el predictor con un desarrollo de Taylor, convierte el problema en un *LMM de trabajo* y actualiza por iteración `IRLS`+BLUP. Es rápida pero sesgada en conteos muy pequeños.
- **Métodos Monte Carlo.** *MCEM* y *MCNR* estiman la integral mediante muestreo (MCMC importance sampling) y son ventajosos cuando q es grande o la verosimilitud es multimodal.

En `R`, `lme4::glmer` implementa Laplace y AGQ; `glmmTMB` añade varianzas de dispersión y familia binomial negativa. Todos los métodos retornan estimadores de máxima verosimilitud (ML) —o ML restringida cuando se re-optimiza la dispersión en etapas internas— y calculan la información observada para las varianzas estándar.

2.5.4. Diagnóstico y selección de modelo

Una vez ajustado un modelo GLMM, es fundamental evaluar su adecuación y compararlo con modelos alternativos. Existen diversas herramientas de diagnóstico que permiten detectar posibles problemas de ajuste, identificar estructuras inapropiadas y guiar la selección del modelo más adecuado.

Una de las técnicas más utilizadas actualmente es el análisis de *residuos condicionales simulados*, implementado en el paquete `DHARMa`. Esta metodología permite diagnosticar problemas como sobredispersión, infra-dispersión o la presencia de ceros estructurales, simulando nuevos conjuntos de datos a partir del modelo ajustado y comparando la distribución de los residuos simulados con los observados. Su ventaja radica en que no requiere supuestos fuertes sobre la distribución exacta de los residuos y proporciona evaluaciones gráficas e inferenciales claras.

En cuanto a la selección de modelo, los criterios de información como el AIC (Akaike Information Criterion) y BIC (Bayesian Information Criterion) ofrecen medidas comparativas útiles. Estos criterios permiten evaluar distintas combinaciones de familias de distribución, funciones de enlace o estructuras de efectos aleatorios.

Por último, los gráficos de valores predichos frente a observados, es decir, $\hat{y}$ versus y, constituyen una herramienta sencilla pero poderosa para detectar patrones de falta de ajuste sistemático. Estos gráficos pueden revelar discrepancias importantes en las colas de la distribución o en subgrupos específicos de la muestra, indicando la necesidad de refinar la especificación del modelo o considerar alternativas más flexibles.

En conjunto, estas técnicas permiten asegurar la validez del modelo ajustado, mejorar su capacidad predictiva y fundamentar la elección entre estructuras rivales bajo una base empírica sólida.

2.5.5. Elección de variantes de GLMM según la estructura de los datos

Dentro del marco de los modelos mixtos lineales generalizados (GLMM), la elección de una distribución específica para la variable de respuesta —y posibles extensiones como la incorporación de inflación de ceros— depende de las propiedades estadísticas particulares del conjunto de datos.

Cuando se modelan variables de conteo con una relación media-varianza cercana, un GLMM con distribución Poisson e interceptos aleatorios puede ser suficiente para capturar la variabilidad intra-sujeto. Sin embargo, en presencia de *sobredispersión* —una situación común en estudios microbiológicos— esta suposición se vuelve inadecuada. La sobredispersión ocurre cuando la variabilidad de los datos es mayor a la que predice el modelo de Poisson, es decir, cuando la desviación estándar observada excede de manera sistemática a la media ($\mathrm{Var}(Y) > \mathbb{E}[Y]$). En tales casos, el modelo de Poisson subestima la varianza real de los conteos, lo que puede llevar a errores en la inferencia estadística y a una sobreestimación de la significancia de los efectos fijos. Para abordar este problema, los modelos binomiales negativos mixtos (NBMM) resultan más apropiados, al incluir un parámetro adicional que permite modelar explícitamente dicha sobredispersión (Zhang et al., 2018).

Cuando además se observa una cantidad considerable de ceros —ya sea por razones biológicas (ausencia real de un taxón) o técnicas (errores de detección)— los modelos inflados en ceros, como los ZINBMM, ofrecen una estructura más realista. Estos modelos permiten separar el proceso generador de ceros del proceso generador de conteos positivos, evitando sesgos en la estimación de los efectos fijos y mejorando la interpretación de los parámetros (Zhang et al., 2020).

En resumen, todos estos modelos —Poisson, NBMM, ZINBMM, entre otros— forman parte de la familia de GLMM, diferenciándose principalmente en la especificación de la distribución de la respuesta y en las estructuras adicionales incorporadas para modelar adecuadamente fenómenos como la sobredispersión o la inflación de ceros. La elección entre ellos debe guiarse por el comportamiento empírico de los datos y por consideraciones teóricas sobre el proceso generador subyacente.

2.6. Modelos binomiales negativos mixtos y su extensión cero-inflada

Los *conteos de abundancia* derivados de la secuenciación 16S-rRNA o shot-gun metagenómica presentan dos rasgos empíricos ineludibles:

(a) **Sobredispersión** —la varianza excede con creces a la media incluso tras ajustar la profundidad de lectura, en donde el concepto *profundidad de lectura* hace noción al número total de secuencias obtenidas por muestra, que determina la sensibilidad de detección de los taxones presentes—,

(b) **Inflación de ceros** —una fracción considerable de las celdas contabiliza ausencia de lectura para un taxón dado, aun cuando la biblioteca total sea elevada.

La combinación de estos fenómenos compromete los supuestos de los modelos lineales generalizados convencionales (p. ej. Poisson, quasi-Poisson, log-normal-Gaussian) y exige estrategias específicas que:

- modelen la *sobredispersión* mediante la distribución binomial negativa (NB), cuya varianza es $\mathrm{Var}(Y) = \mu + \mu^2/\theta$, donde $\mu > 0$ representa la media condicional de la respuesta y $\theta > 0$ es el parámetro de dispersión inversa (cuanto mayor es θ, menor es la sobredispersión);
- capturen la *correlación intra-sujeto* introducida por el diseño longitudinal a través de **efectos aleatorios**;
- distingan entre *ceros estructurales* (ausencia biológica real) y *ceros de muestreo* (limitaciones técnicas), incorporando una capa de mezcla cero-inflada cuando sea necesario.

2.6.1. Definición de la distribución binomial negativa

La distribución binomial negativa $\mathrm{NB}(\mu, \theta)$ es una generalización de la distribución de Poisson que permite sobredispersión. Su función de masa de probabilidad para un valor entero $y \in \mathbb{N}_0$ está dada por:

$$\mathbb{P}(Y = y) = \frac{\Gamma(y + \theta)}{\Gamma(\theta)\, y!} \left(\frac{\mu}{\mu + \theta}\right)^{y} \left(\frac{\theta}{\mu + \theta}\right)^{\theta},$$

donde:

- $\mu > 0$ es el parámetro de media,
- $\theta > 0$ es el parámetro de dispersión inversa,

- $\Gamma(\cdot)$ es la función gamma.

La esperanza y varianza de esta distribución son:

$$\mathbb{E}(Y) = \mu, \qquad \mathrm{Var}(Y) = \mu + \frac{\mu^2}{\theta},$$

lo que permite modelar una varianza mayor a la media (sobredispersión), característica frecuente en datos metagenómicos.

2.6.2. Modelos jerárquicos mixtos para conteos sobredispersos

Estas consideraciones conducen naturalmente a dos clases de modelos jerárquicos:

Modelo Mixto Binomial Negativo (NBMM)

El **NBMM** supone que *todos* los ceros provienen de la misma distribución NB que genera el resto de los conteos, resolviendo la sobredispersión y la heterogeneidad entre sujetos mediante efectos aleatorios.

Modelo Mixto Binomial Negativo Inflado con Ceros (ZINBMM)

El **ZINBMM** añade una componente de mezcla Bernoulli/Logit que explica la presencia de ceros adicionales, manteniendo la parte NB para los conteos positivos y preservando la estructura aleatoria jerárquica.

En términos unificados, sea Y_{ij} el conteo observado del taxón h en la muestra j del sujeto i, y sea T_{ij} el **tamaño de biblioteca** correspondiente, es decir, el número total de lecturas (reads) asignadas a la muestra j del individuo i. Este valor refleja la *profundidad de secuenciación* alcanzada para cada muestra, que puede variar considerablemente debido a factores técnicos y biológicos, como la carga microbiana presente o la eficiencia de extracción del ADN. La inclusión del término $\log(T_{ij})$ como offset en el modelo permite estandarizar los conteos, ajustando por estas diferencias de profundidad y facilitando la interpretación de μ_{ij} como una medida de abundancia relativa.

Entonces, el modelo se expresa como:

$$Y_{ij} \mid \mathbf{b}_i, \mathbf{a}_i \sim \begin{cases} 0, & \text{con probabilidad } p_{ij} = \mathrm{logit}^{-1}\big(\mathbf{Z}_{ij}^\top \boldsymbol{\alpha} + \mathbf{G}_{ij}^\top \mathbf{a}_i\big), \\ \mathrm{NB}\big(\mu_{ij}, \theta\big), & \text{con probabilidad } 1 - p_{ij}, \end{cases}$$

$$\log(\mu_{ij}) = \log(T_{ij}) + \mathbf{X}_{ij}^\top \boldsymbol{\beta} + \mathbf{G}_{ij}^\top \mathbf{b}_i, \quad \mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_b),\ \mathbf{a}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_a),$$

donde:

- p_{ij} representa la probabilidad de que la observación sea un cero estructural,
- $\mathbf{X}_{ij}$, $\mathbf{Z}_{ij}$, $\mathbf{G}_{ij}$ son vectores de covariables para las distintas partes del modelo,
- $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ son vectores de coeficientes fijos,
- $\mathbf{b}_i$, $\mathbf{a}_i$ son efectos aleatorios específicos del sujeto i.

En el caso NBMM (sin inflación de ceros), se asume que $p_{ij} \equiv 0$ para todo i, j, eliminando así la capa de mezcla.

La estimación se basa en log-verosimilitud penalizada o en algoritmos EM-IWLS, delegándose los detalles computacionales al paquete `NBZIMM` de **R**.

2.6.3. Motivación para datos de microbiota

El diseño de esta investigación requiere modelos que reproduzcan con fidelidad las *cinco* características empíricas más citadas de los conteos microbianos longitudinales obtenidos por secuenciación 16S rRNA o metagenómica:

(a) **Sobredispersión intrínseca** La varianza de los conteos supera sistemáticamente a la media, lo que invalida la suposición $\mathrm{Var}(Y) = \mu$ de la familia de Poisson . La distribución **binomial negativa** introduce un parámetro de dispersión θ capaz de acomodar esta variabilidad extra.

(b) **Estructura longitudinal jerárquica** Cada sujeto aporta varias mediciones en el tiempo (t_{ij}), generando correlación intra-individuo. Los *efectos aleatorios* $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$ de los NBMM/ZINBMM modelan interceptos y pendientes sujeto-específicos, evitando subestimar los errores estándar y permitiendo inferencias tanto poblacionales como individuales.

(c) **Profundidad de secuenciación variable** El tamaño de biblioteca T_{ij} varía entre muestras; incluirlo como *offset* $\log(T_{ij})$ en el predictor lineal conecta naturalmente los conteos brutos con las abundancias relativas sin necesidad de escalar los datos previamente.

(d) **Inflación de ceros** Una parte de los ceros corresponde a verdadera ausencia biológica (*ceros estructurales*) y otra a límites de detección (*ceros de muestreo*) . El componente logístico de la capa cero-inflada de los **ZINBMM** discrimina entre ambos mecanismos, reduciendo falsos positivos cuando los ceros dominan la matriz OTU/ASV.

(e) **Composicionalidad y covariables múltiples** El uso conjunto de offsets, efectos fijos $\boldsymbol{\beta}$ (p. ej. grupo, tiempo, interacciones) y efectos aleatorios permite analizar abundancias relativas sin descartar lecturas y cuantificar la variabilidad debida a factores clínicos o ambientales, tal como se requiere en el estudio de la microbiota vaginal gestacional.

En conjunto, los NBMM y su extensión ZINBMM satisfacen de forma integrada todos los requisitos anteriores, justificando su selección como marco inferencial principal para los análisis de abundancia bacteriana desarrollados en este trabajo.

2.6.4. Resumen de estudios previos

La literatura sobre análisis de datos de microbiota ha evolucionado en paralelo con las limitaciones empíricas descritas en la Sección 2.6.3. A continuación se sintetizan los hitos metodológicos más relevantes y sus aplicaciones biológicas[1]:

[1] La revisión se centra en trabajos específicos sobre microbiota; no incluye la extensa bibliografía general sobre GLMM.

(i) De Poisson/Quasi-Poisson a modelos con dispersión adaptativa. Los primeros estudios longitudinales utilizaron GLMM Poisson con interceptos aleatorios para modelar conteos bacterianos ; la infravaloración sistemática de la varianza motivó la adopción de la familia binomial negativa. La formulación *two-part mixed model* propuesta por Chen and Li (2016) separó el proceso de generación de ceros y el de conteos positivos, abriendo la vía para las extensiones cero-infladas.

(ii) Introducción de los NBMMs. Zhang et al. (2017) formalizaron el **Negative Binomial Mixed Model** (NBMM) para datos de microbioma, demostrando en simulaciones que mantiene el control de la tasa de falsos descubrimientos (FDR) frente a modelos Poisson sobredispersos. Posteriormente, Zhang X (2018) extendieron el enfoque a datos temporales con múltiples visitas por sujeto, mientras que Zhang and Yi (2020) desarrollaron un algoritmo IWLS eficiente e implementado en el paquete `NBZIMM`.

(iii) Modelos cero-inflados mixtos (ZINBMMs). La evidencia de ceros estructurales condujo a incorporar una capa de mezcla Bernoulli—Zhang et al. (2020) propusieron el **Zero-Inflated Negative Binomial Mixed Model** (ZINBMM) y mostraron ganancias sustanciales de potencia cuando más del 50 % de las celdas son ceros. Para el ajuste, se popularizó el algoritmo EM-IWLS descrito por Zhang and Yi (2020), también disponible en `NBZIMM`.

(iv) Evaluaciones comparativas y revisiones. Simulaciones extensas bajo distintos escenarios de dispersión, inflación de ceros y tamaños de muestra —p. ej. Kodikara et al. (2022)— confirman que NBMM y ZINBMM superan a modelos transformados (CLR/ILR) en control de error tipo I y potencia. Las revisiones de Kodikara et al. (2022) y Correa Morales and Salazar Uribe (2016) sintetizan estos hallazgos y recomiendan explícitamente NBMM/ZINBMM como métodos de referencia para estudios longitudinales de microbiota.

(v) Herramientas de software. El paquete `NBZIMM` agrupa las rutinas `glmm.nb` (NBMM) y `glmm.zinb` (ZINBMM) sobre la plataforma `nlme`, lo que facilita especificar estructuras de correlación (p. ej. AR(1)) y efectos aleatorios complejos. Su adopción en investigaciones clínicas y ambientales respalda la viabilidad computacional de los modelos aquí empleados.

En conjunto, la evidencia metodológica y aplicada respalda el uso de NBMM y ZINBMM como marcos preferentes para analizar la dinámica temporal de la microbiota, razón por la cual se integran explícitamente en el flujo de trabajo de la presente investigación.

Capítulo 3

Metodología general

3.1. Flujo de trabajo

En ambos estudios (microbiota y crecimiento vegetal) se siguió un flujo de trabajo reproducible que abarca desde la limpieza de los datos brutos hasta la generación de gráficas finales. La Figura 3.1 resume gráficamente las fases

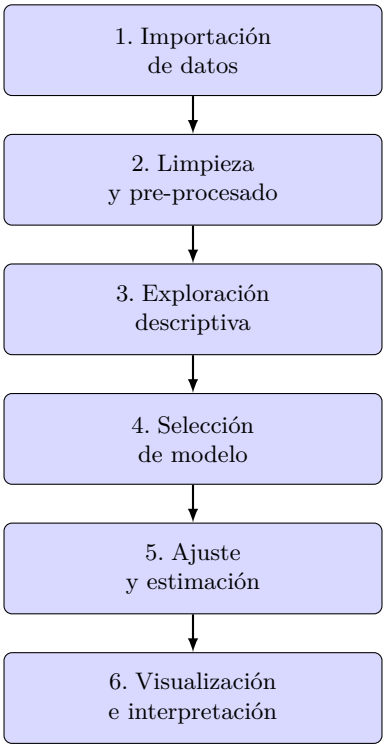


Figura 3.1: Flujo de trabajo general para el análisis longitudinal desde la importación de datos hasta la comunicación de resultados.

A continuación se detallan los pasos comunes y las particularidades de cada estudio.

1. **Importación de datos**

   - **Estudio I (microbiota).** Se importó el objeto `Romero` del paquete `NBZIMM`, separando la matriz OTU/ASV (`otu`) y la tabla de metadatos (`sam`).
   - **Estudio II (plantines).** Se leyeron siete hojas de un archivo `Excel` con mediciones de altura y DAC utilizando `readxl`; se añadió la fecha correspondiente a cada hoja.

2. **Limpieza y pre-procesado**

   (a) Normalización de nombres de columnas (`stringr`) y transformación de tipos (`dplyr::mutate`).

   (b) Filtrado de registros sin variabilidad (por ejemplo, bacterias de varianza nula) o con mediciones aisladas.

   (c) En los plantines se eliminaron plantas con *NAs* intercalados y aquellas sin medición basal (fecha 1).

3. **Exploración descriptiva**

   - Conteo de observaciones por sujeto y distribución de covariables (`dplyr`).
   - Listado de taxones y su frecuencia; identificación de sobredispersión preliminar mediante `var(y) / mean(y)`.
   - Gráficos de perfil longitudinal para altura y DAC diferenciados por tratamiento (`ggplot2`).

4. **Selección de modelos**

   - **Microbiota:** NBMM y ZINBMM con offset $\log(T_{ij})$, covariables `preg`, `Days` e interacción `preg:Days`; intercepto aleatorio por sujeto.
   - **Plantines:** LMMs con interceptos y pendientes aleatorias (`nlme`), incorporando día, tratamiento y covariables ambientales acumuladas.

5. **Ajuste, comparación y selección**

   (a) `glmmTMB` (`family = nbinom2`) para NBMM/ZINBMM; `lme` (`nlme`) para LMMs.

   (b) *Estudio I:* contraste de parámetros mediante Tests de Wald con corrección FDR sobre los p-valores de la interacción.

   (c) *Estudio II:* comparación jerárquica de modelos con ΔAIC, ΔBIC y pruebas de razón de verosimilitud (LRT) entre modelos anidados.

6. **Visualización e interpretación**

   - Barras de abundancias promedio por grupo gestacional.
   - Mapas de calor de abundancia a lo largo del tiempo para taxones significativos.
   - Curvas observadas *vs.* predichas para plantas representativas bajo cada tratamiento, comparando varios modelos candidatos.

3.2. Software y paquetes estadísticos

Todos los análisis se realizaron en **R** (`v4.3.x`) dentro de la IDE **RStudio**. El flujo de trabajo se articuló en cuadernos `R Markdown` (`.Rmd`).

A continuación se resumen los paquetes clave y su papel dentro del proyecto.

(I) **Modelización mixta**

- nlme: ajuste de Modelos Mixtos Lineales (`lme()`) y No Lineales (`nlme()`); indispensable para los LMM y NLMM.
- lme4: alternativa moderna para modelos mixtos gaussianos y GLMM (`lmer()`, `glmer()`); usada para comprobaciones de robustez.
- glmmTMB: motor flexible que permite efectos aleatorios + familias NB y ZINB; fue la "primera pasada" para el barrido masivo de taxones.
- NBZIMM: implementación del algoritmo IWLS/EM específico para *NBMM* y *ZINBMM*. Funciones principales: `glmm.nb()` y `glmm.zinb()`.

(II) **Manipulación y limpieza de datos**

- dplyr, tidyr, stringr, purrr (*tidyverse*): flujo de operaciones para filtrar, pivotar y mapear listas de modelos.
- readxl: importación de las hojas de cálculo con mediciones de plántulas.

(III) **Visualización**

- ggplot2: generación de perfiles longitudinales, mapas de calor, Box-Plots y comparaciones modelo–datos.

3.2.1. Disponibilidad de código

El código completo utilizado en este proyecto, incluidos los scripts y cuadernos `R Markdown`, está disponible en el repositorio GitHub:

https://github.com/GabrielaGutierrez2025/tesis-microbiota-plantines

3.3. Criterios de comparación de modelos

La estrategia de selección se sustenta en una combinación complementaria de *criterios numéricos* (información, contraste y predicción) y *diagnósticos gráficos*. Cada criterio se calcula con las funciones nativas de los paquetes `nlme`, `lme4`, `glmmTMB` y `NBZIMM`, o con rutinas auxiliares incluidas en los `scripts` propios del repositorio del proyecto.

3.3.1. Criterios de información

- **Akaike Information Criterion (AIC)**

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}}) + 2\,k,$$

donde $\ell(\hat{\boldsymbol{\theta}})$ es la log-verosimilitud maximizada y k el número de parámetros libres. Penaliza la complejidad con $2k$ y es asintóticamente equivalente a la validación cruzada leave-one-out (LOO) para $n \to \infty$ (Correa Morales and Salazar Uribe, 2016).

- **Bayesian Information Criterion (BIC)**

$$\text{BIC} = -2\ell(\hat{\boldsymbol{\theta}}) + k\log(n),$$

con una penalización más severa ($\log n$) que favorece modelos parsimoniosos cuando el tamaño muestral es grande. Se reporta sistemáticamente en la comparación en las sucesivas anidaciones de NLMM para los plantines (Correa Morales and Salazar Uribe, 2016).

3.3.2. Contrastes de verosimilitud

- **Likelihood Ratio Test (LRT)** Se utiliza para comparar modelos *anidados*: $\mathcal{M}_0 \subset \mathcal{M}_1$. El estadístico $\chi^2 = -2\big[\ell_0 - \ell_1\big]$ se evalúa con `anova(m0,m1)` en `nlme` y `glmmTMB`. Fue el instrumento principal en el estudio de crecimiento vegetal para jerarquizar **M1**–**M8**. La significancia se juzgó con $\alpha = 0{,}05$.
- **Wald test** Para los modelos NBMM y ZINBMM se contrasta $H_0 : \beta_k = 0$ mediante

$$Z = \frac{\hat{\beta}_k}{\text{SE}(\hat{\beta}_k)} \quad \sim \mathcal{N}(0,1).$$

Los p-valores se obtienen con `summary()` y se ajustan por FDR (`p.adjust(method="fdr")`) al escanear todos los taxones.

3.3.3. Síntesis de aplicación de criterios

Estudio	**Info.**	**Contraste**
Microbiota (NBMM)	AIC/BIC	Wald Z
Microbiota (ZINBMM)	AIC/BIC	Wald Z
Plantines (NLMM)	AIC/BIC	LRT χ^2

Estos criterios, aplicados de forma coherente con la estructura de cada conjunto de datos, garantizan una selección de modelos equilibrada entre ajuste, parsimonia y capacidad predictiva.

Capítulo 4

Caso de Estudio I – Microbiota vaginal en mujeres gestantes y no gestantes

La microbiota vaginal constituye una comunidad compleja y dinámica de microorganismos, cuya composición y estabilidad han sido asociadas a la salud ginecológica y obstétrica de las mujeres. Durante la gestación, se ha observado una tendencia hacia configuraciones más dominadas por especies del género *Lactobacillus*, así como una reducción en la diversidad y variabilidad microbiana, lo que sugiere la existencia de mecanismos de estabilización microbiana relacionados con la gestación (Romero et al., 2014).

Este estudio se centra en el análisis longitudinal de datos de abundancia bacteriana obtenidos a través de secuenciación 16S en mujeres gestantes y no gestantes, con el propósito de explorar las diferencias en su microbiota vaginal a lo largo del tiempo. A diferencia de análisis transversales clásicos, la naturaleza longitudinal de estos datos permite capturar trayectorias individuales y evaluar cambios dinámicos asociados al estado gestacional y a otras variables clínicas.

Desde el punto de vista estadístico, este tipo de datos impone múltiples desafíos: la presencia de conteos con alta proporción de ceros, la sobredispersión, y la estructura jerárquica debida a observaciones repetidas dentro de cada sujeto. Para abordar estos desafíos, en este capítulo se emplean modelos mixtos binomiales negativos (NBMM) y modelos mixtos binomiales negativos inflados con ceros (ZINBMM), los cuales permiten modelar la variabilidad interindividual, la estructura longitudinal y la naturaleza discreta e inflada en ceros de los datos microbiológicos.

El análisis se realiza a nivel de taxón individual, ajustando modelos separados para cada unidad bacteriana observada. Este enfoque permite evaluar el efecto del grupo gestacional, del tiempo gestacional y de su interacción en la abundancia relativa de cada microorganismo. Los resultados obtenidos ofrecen una caracterización detallada de la evolución microbiana en mujeres gestantes versus no gestantes.

4.1. Objetivo

El propósito central de este estudio es **caracterizar la dinámica longitudinal de la microbiota vaginal** y su asociación con la gestación, empleando modelos estadísticos apropiados para datos de conteo sobredispersos e inflados con ceros (NBMM y ZINBMM), de esta manera, el objetivo está dado por:

Describir y modelar los cambios temporales en la abundancia de cada taxón bacteriano, comparando mujeres gestantes y no gestantes, e identificar los factores que explican dichas variaciones.

De este objetivo general se desprenden las siguientes preguntas de investigación, orientadas a comprender las diferencias en la composición y estabilidad de la microbiota vaginal entre mujeres gestantes y no gestantes, así como la influencia de variables clínicas relevantes en su dinámica temporal.

En primer lugar, se busca determinar si existen diferencias sistemáticas en la *composición bacteriana promedio* entre mujeres gestantes y no gestantes a lo largo del periodo de seguimiento. Esta comparación permitirá evaluar si la gestación induce patrones microbianos diferenciables de manera global.

En segundo lugar, se evalúa si la microbiota vaginal de mujeres gestantes presenta una *mayor estabilidad temporal* que la de mujeres no gestantes, en línea con hallazgos previos que sugieren una menor variabilidad durante la gestación (Romero et al., 2014).

Finalmente, se plantea identificar *taxones individuales* cuya abundancia se vea significativamente influida por factores clínicos específicos. En particular, se analizarán los efectos del estado de gestación (`preg`), del tiempo de gestación (`GA_Days`), y de la interacción entre ambos (`preg` $\times$ `GA_Days`) sobre la dinámica de cada taxón bacteriano.

En las secciones 4.4.1 y 4.4.2 se detallan los modelos estadísticos utilizados para responder estas preguntas, mientras que la Sección 4.5 presenta los resultados obtenidos y su interpretación biológica.

4.2. Diseño del estudio y recolección de muestras

4.2.1. Aspectos éticos y disponibilidad de datos

El protocolo fue aprobado por el Human Investigation Committee de la Wayne State University y la Institutional Review Board del *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. Los datos de mujeres no gestantes están disponibles en el Sequence Read Archive (acceso `SRA026073`), y los metadatos en `dbGaP` (estudio `phs000261`); los datos de gestantes se publicaron íntegramente en Romero et al. (2014).

4.2.2. Tipo de estudio y población

El trabajo se planteó como una **cohorte longitudinal prospectiva** con dos grupos paralelos:

a) **Gestantes normales**: mujeres sin comorbilidades obstétricas, quirúrgicas ni médicas, con embarazo único y parto a término (38–42 semanas).

b) **No gestantes**: mujeres sanas en edad reproductiva, libres de enfermedad clínica.

Todas las participantes otorgaron consentimiento informado escrito antes de cualquier procedimiento del estudio.

4.2.3. Calendario de visitas y muestreo

Gestantes. Cada participante fue citada a controles vaginales con espéculo:

- cada 4 semanas hasta la semana 24 de gestación;
- cada 2 semanas desde la semana 24 hasta la última visita prenatal.

En cada visita, un obstetra o matrona recolectó fluido del fórnix posterior con hisopo de *Dacron*®.

No gestantes. Las voluntarias tomaron auto-muestras vaginales dos veces por semana durante 16 semanas, siguiendo un protocolo previamente validado.

Todos los hisopos se mantuvieron a $-70°C$ hasta su análisis. Se prepararon extendidos para tinción de Gram y se evaluó el índice de Nugent como referencia morfológica de la microbiota.

4.2.4. Procesamiento de laboratorio

Extracción de ADN. Los hisopos congelados se incubaron en tampón de lisis enzimática (pre-calentado a $55°C$) seguido de digestión con proteinasa K, SDS y ARNasa A. La disrupción mecánica adicional se realizó en *bead-beater*, y el ADN se purificó con el kit ZR Fecal DNA para eliminar inhibidores de PCR. Se obtuvieron 2.5–5 µg de ADN genómico de alta calidad por muestra.

Amplificación 16S y pirosecuenciación. Se amplificó la región V1–V2 del gen 16S rRNA mediante PCR con los cebadores:

```
27F    GCCTTGCCAGCCCGCTCAGTCAGAGTTTGATCCTGGCTCAG
338R   GCCTCCCTCGCGCCATCAGNNNNNNNNN CATGCTGCCTCCCGTAGGAGT
```

Los amplicones se verificaron por gel, cuantificaron con *PicoGreen*® y se secuenciaron en plataforma 454 FLX Titanium del Genomics Resource Center (University of Maryland School of Medicine).

4.3. Estructura de los datos y variables

El estudio generó dos tablas fundamentales:

a) Matriz de abundancias $\mathbf{Y} \in \mathbb{N}^{n \times p}$:

- $n = 900$ muestras longitudinales.
- $p = 143$ taxones (columnas).
- La entrada y_{ij} representa el número bruto de lecturas asignadas al taxón j en la muestra i.

b) Tabla de metadatos $\mathbf{S} \in \mathbb{R}^{n \times 9}$:

$$\mathbf{S} = \big[\text{Subject\_ID, Sample\_ID, GA\_Days, Age, Race, Nugent.Score, CST, Total.Read.Counts, pregnant}\big].$$

Unidades de análisis. Las $n = 900$ observaciones provienen de $N_{\text{subj}} = 54$ mujeres (22 gestantes, 32 no gestantes). Cada sujeto i aporta n_i muestras secuenciales ($\sum_i n_i = 900$); el vector `Subject_ID` permite modelar esta jerarquía con efectos aleatorios.

Covariables principales.

- `pregnant` (0/1): indicador de embarazo (factor de dos niveles).
- `GA_Days`: días de gestación en la fecha de muestreo (t_{ij}). Se estandarizó $t^{\star}_{ij} = \text{scale}(\texttt{GA\_Days})$ para mejorar la convergencia numérica.
- `Age`: edad materna (años), también centrada y escalada.
- `Total.Read.Counts` $= T_{ij}$: profundidad de secuenciación empleada como *offset* $\log(T_{ij})$ en los modelos de conteo.
- `Race`, `Nugent.Score`, `CST`: variables categóricas u ordinales disponibles para análisis exploratorios; no se incluyeron en los modelos principales de esta memoria.

Propiedades de los conteos. Los vectores $\mathbf{y}_{i\cdot}$ exhiben:

- **Sobredispersión**: la varianza supera a la media, circunstancia que justifica el empleo de la distribución binomial negativa.
- **Inflación de ceros**: múltiples taxones presentan abundancia nula en una fracción apreciable de las muestras, motivando la extensión ZINBMM.

Este diseño jerárquico, junto con la alta dimensionalidad de taxones y la presencia de sobredispersión e inflación de ceros, motiva la estrategia de modelación descrita.

4.4. Modelación

4.4.1. Modelo Mixto Binomial Negativo (NBMM)

El modelo NBMM es un enfoque estadístico utilizado para estudiar la relación entre covariables y la abundancia de un taxón bacteriano específico, teniendo en cuenta la estructura longitudinal de los datos. Este modelo es especialmente útil cuando los conteos observados presentan *sobredispersíon*, es decir, cuando la varianza excede a la media esperada bajo un modelo de Poisson.

Modelo

Sea y_{ij} la abundancia observada de un taxón bacteriano para el sujeto i en el tiempo j. Se asume que:

$$y_{ij} \sim \text{NB}(\mu_{ij}, \theta), \quad i = 1, \ldots, n, \; j = 1, \ldots, n_i,$$

donde:

- μ_{ij} es la media esperada de la distribución binomial negativa.
- $\theta > 0$ es el parámetro de dispersión, que controla el grado de sobredispersíon.

La media esperada μ_{ij} se modela mediante un enlace logarítmico como:

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \cdot \text{preg}_i + \beta_2 \cdot \text{Days}_{ij} + \beta_3 \cdot (\text{preg}_i \times \text{Days}_{ij}) + b_i,$$

donde:

- β_0 es el intercepto.
- β_1 es el coeficiente asociado al estado gestacional (`preg`), que captura el efecto principal de gestación sobre la abundancia bacteriana.
- β_2 es el coeficiente asociado al tiempo (`Days`), que modela el cambio de la abundancia bacteriana en función de los días de seguimiento.
- β_3 es el coeficiente de la interacción `preg:Days`, que evalúa si el efecto de gestación varía a lo largo del tiempo.
- b_i es un efecto aleatorio específico del sujeto i, que captura la heterogeneidad interindividual y la correlación entre mediciones repetidas de un mismo sujeto, con $b_i \sim \mathcal{N}(0, \sigma_b^2)$.

Este modelo permite evaluar no solo los efectos principales de las covariables, sino también si la relación entre el estado gestacional y la abundancia bacteriana cambia con el tiempo. La inclusión de un término de interacción (`preg:Days`) es fundamental en estudios longitudinales de microbiota, ya que la composición bacteriana puede variar de forma diferente en mujeres gestantes y no gestantes a medida que avanza el tiempo de seguimiento.

4.4.2. Modelo Mixto Binomial Negativo Inflado en Ceros (ZINBMM)

El **ZINBMM** extiende el NBMM incorporando un mecanismo de *inflación de ceros*, que permite distinguir entre:

- *Ceros estructurales*: corresponden a la ausencia verdadera del taxón, ya sea por razones biológicas (el organismo no está presente) o técnicas (limitaciones en la detección).
- *Ceros de muestreo*: son ceros que provienen de la cola izquierda de la distribución de conteo (binomial negativa) y que podrían ser distintos de cero en otras réplicas o mediciones.

Este enfoque es especialmente relevante en datos de microbiota, donde muchos taxones presentan abundancia nula en una fracción considerable de las muestras longitudinales.

Modelo

Sea y_{ij} el conteo de lecturas para el taxón en la muestra j del sujeto i. El modelo asume que:

$$y_{ij} \mid b_i \sim \begin{cases} 0, & \text{con probabilidad } p_{ij}, \\ \text{NB}(\mu_{ij}, \theta), & \text{con probabilidad } 1 - p_{ij}, \end{cases} \qquad i = 1, \ldots, n,\ j = 1, \ldots, n_i,$$

donde:

- p_{ij} es la probabilidad de que la observación sea un cero estructural.
- μ_{ij} es la media del submodelo de conteo.
- $\theta > 0$ es el parámetro de dispersión de la binomial negativa.
- $b_i \sim \mathcal{N}(0, \sigma_b^2)$ es el efecto aleatorio que captura la heterogeneidad entre sujetos y la dependencia temporal de las observaciones repetidas.

Parte de conteos (submodelo NB). La media μ_{ij} se especifica mediante un enlace logarítmico:

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \cdot \text{preg}_i + \beta_2 \cdot \text{Days}_{ij} + \beta_3 \cdot (\text{preg}_i \times \text{Days}_{ij}) + b_i,$$

donde:

- β_0 es el intercepto.
- β_1 es el efecto principal del estado gestacional (`preg`) sobre la abundancia.
- β_2 es el efecto del tiempo (`Days`) sobre la abundancia.
- β_3 es el efecto de la interacción `preg:Days`, que evalúa si la relación entre gestación y abundancia varía con el tiempo.
- b_i es el intercepto aleatorio por sujeto.

Parte de ceros (submodelo logístico). La probabilidad de cero estructural p_{ij} se modela con un enlace logit:

$$\text{logit}(p_{ij}) = \gamma_0 + \gamma_1 \cdot \text{preg}_i + \gamma_2 \cdot \text{Days}_{ij},$$

donde:

- γ_0 es el intercepto del submodelo de ceros.
- γ_1 y γ_2 representan cómo el estado gestacional y el tiempo afectan la propensión a observar ceros estructurales.

En este modelo, no se incluyen efectos aleatorios en la parte de ceros.

Resumen. El ZINBMM conserva las ventajas del NBMM para manejar sobredispersión y correlación intra-sujeto, añadiendo un mecanismo explícito para los ceros inflados. Esto permite:

(a) Estimar por separado el efecto de las covariables sobre la **probabilidad de ausencia** del taxón (p_{ij}) y sobre la **abundancia promedio** μ_{ij}, condicionada a que el taxón esté presente.

(b) Diferenciar ceros biológicos reales de ceros técnicos, mejorando la inferencia en taxones raros.

(c) Incorporar interceptos y pendientes aleatorias (b_i), lo que permite capturar la variabilidad individual y las trayectorias particulares de cada sujeto a lo largo del tiempo.

En el análisis aplicado se ajustó un ZINBMM por cada uno de los 143 taxones, permitiendo identificar diferencias significativas ligadas al estado gestacional y su interacción con el tiempo de gestación.

4.5. Resultados y visualizaciones

4.5.1. Criterio de significancia: Test de Wald

El **Test de Wald** se empleó para evaluar la significancia estadística de los coeficientes estimados en los modelos NBMM y ZINBMM. En particular, este test permite verificar si existe una relación significativa entre una covariable de interés (por ejemplo, `preg` o su interacción `preg:Days`) y la abundancia bacteriana.
El test se formula para contrastar la siguiente hipótesis nula sobre un parámetro β_k (coeficiente de la covariable de interés):

- $\mathbf{H_0}$**:** $\beta_k = 0$ (el parámetro no tiene un efecto significativo sobre la abundancia bacteriana).
- $\mathbf{H_1}$**:** $\beta_k \neq 0$ (el parámetro tiene un efecto significativo).

En este contexto, si la hipótesis nula es rechazada, se concluye que existe una asociación significativa entre la covariable de interés (por ejemplo, `preg` o `preg:Days`) y la variable dependiente (abundancia bacteriana).
El estadístico de prueba del Test de Wald se calcula de la siguiente manera:

$$Z = \frac{\hat{\beta}_k}{\mathrm{SE}(\hat{\beta}_k)},$$

donde:

- $\hat{\beta}_k$ es el coeficiente estimado de la covariable.
- $\mathrm{SE}(\hat{\beta}_k)$ es el error estándar asociado al coeficiente estimado.

El estadístico Z sigue una distribución normal estándar bajo la hipótesis nula. El valor p asociado al estadístico se compara con un nivel de significancia predefinido ($\alpha = 0{,}05$):

- Si $p < \alpha$, se rechaza H_0 y se concluye que la covariable tiene un efecto significativo sobre la abundancia.
- Si $p \geq \alpha$, no se rechaza H_0 y se concluye que no hay evidencia suficiente para afirmar una relación significativa.

En el análisis de datos longitudinales de microbiota, el Test de Wald desempeña un papel clave al validar la significancia de los parámetros en los modelos NBMM y ZINBMM. Este test permite identificar covariables relevantes y evaluar su relación con la abundancia bacteriana, garantizando la robustez de los resultados obtenidos.
En estudios longitudinales de la microbiota vaginal, la aplicación de modelos NBMM y ZINBMM con interacciones permite:

- Evaluar la relación entre la abundancia de bacterias específicas y covariables como el embarazo (`preg`).
- Analizar cómo esta relación varía a lo largo del tiempo mediante la interacción `preg:Days`.
- Identificar bacterias cuya abundancia está significativamente asociada a los factores considerados, utilizando el Test de Wald para evaluar la significancia estadística de los parámetros.

4.5.2. Bacterias asociadas a la gestación

Resultados del Modelo NBMM

El modelo NBMM detectó 18 bacterias con coeficientes significativos para la covariable `preg` o su interacción con el tiempo. Estas bacterias se encuentran detalladas en la Modelo NBMM (ver table A.1).

Resultados del Modelo ZINBMM

El modelo ZINBMM detectó 20 bacterias significativas, lo que indica un mayor poder para identificar efectos en datos con ceros inflados, en concordancia con lo reportado en la literatura para estudios de microbiota (Romero et al., 2014). Los nombres taxonómicos correspondientes se presentan en la Modelo ZINBMM (ver table A.2).

Los resultados obtenidos mediante los modelos NBMM y ZINBMM indican que las bacterias listadas presentan coeficientes significativos para la covariable `preg`, o para su interacción con el tiempo (`Days`). Esto significa que la abundancia de estas bacterias se ve afectada de manera significativa por el estado gestacional de la mujer y, en algunos casos, por la evolución temporal.
Los resultados obtenidos revelan diferencias importantes entre los dos modelos aplicados:

- El modelo **NBMM** identificó 18 bacterias significativas. Esto sugiere que este modelo es capaz de captar relaciones entre la covariable `preg` y la abundancia bacteriana, incluso en presencia de sobredispersíon.

- El modelo **ZINBMM** identificó 20 bacterias significativas, lo cual es superior al NBMM. Esto se debe a que el ZINBMM maneja de manera más eficiente los ceros excesivos en los datos, diferenciando entre ceros estructurales y ceros aleatorios.

La mayor cantidad de bacterias significativas detectadas por el ZINBMM indica que este modelo es más adecuado cuando los datos presentan conteos con ceros inflados. En estudios de microbiota, los ceros inflados pueden deberse a la ausencia biológica de ciertas bacterias o a limitaciones técnicas en la secuenciación.

Es importante destacar que entre las bacterias identificadas se encuentran especies clave como *Lactobacillus* y *Lactobacillus jensenii*, que son conocidas por su rol protector en la microbiota vaginal. La identificación de estas bacterias refuerza la relevancia de evaluar su relación con el estado gestacional y su evolución temporal.

4.5.3. Análisis del Modelo NBMM

El modelo NBMM fue utilizado para identificar bacterias significativas cuya abundancia se ve afectada por la variable `preg` (estado de gestación).

Abundancia Promedio de Bacterias Significativas

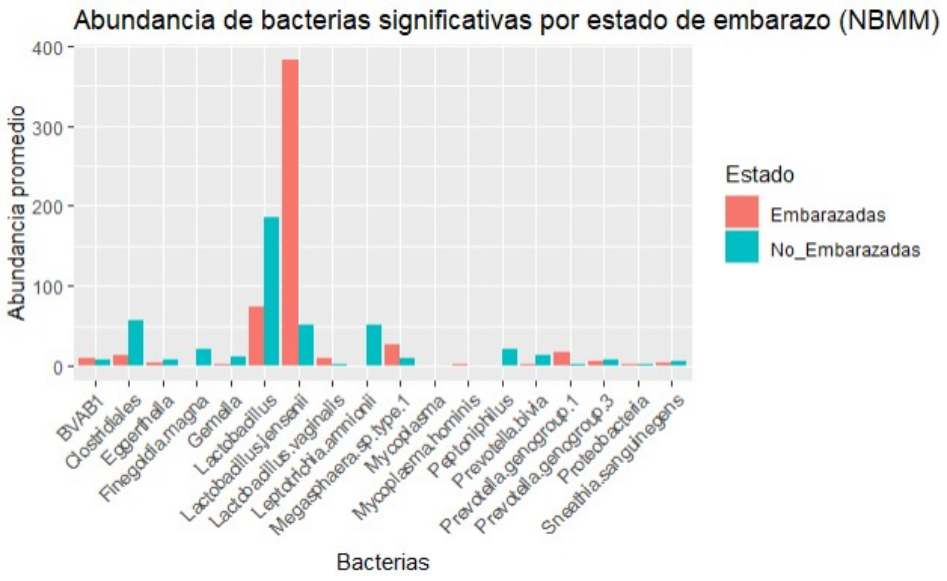


Figura 4.1

El gráfico 4.1 muestra el promedio de abundancia de las bacterias significativas identificadas en mujeres gestantes y no gestantes. Las observaciones más relevantes son:

- **Dominancia de *Lactobacillus*:** Las bacterias del género *Lactobacillus*, en particular *Lactobacillus jensenii*, presentan una abundancia considerablemente mayor en mujeres gestantes

en comparación con las no gestantes. Esta observación respalda la hipótesis de que el embarazo promueve un ambiente vaginal dominado por *Lactobacillus*, lo cual contribuye a la protección contra patógenos.

- **Variabilidad en mujeres no gestantes:** En mujeres no gestantes, se observa una distribución más heterogénea de las bacterias significativas, con abundancias promedio menores y distribuidas entre varios taxones.
- **Implicaciones:** La alta abundancia de *Lactobacillus* en mujeres gestantes sugiere una función estabilizadora clave de estas bacterias, asociada a cambios hormonales y fisiológicos propios del embarazo, respaldada por estudios longitudinales que muestran una mayor estabilidad de la microbiota en gestantes, dominada por *Lactobacillus spp.* (DiGiulio et al., 2015).

4.5.4. Análisis del Modelo ZINBMM

El modelo ZINBMM ajusta por los ceros inflados presentes en los datos, permitiendo una identificación más robusta de bacterias significativas.

Abundancia Promedio de Bacterias Significativas

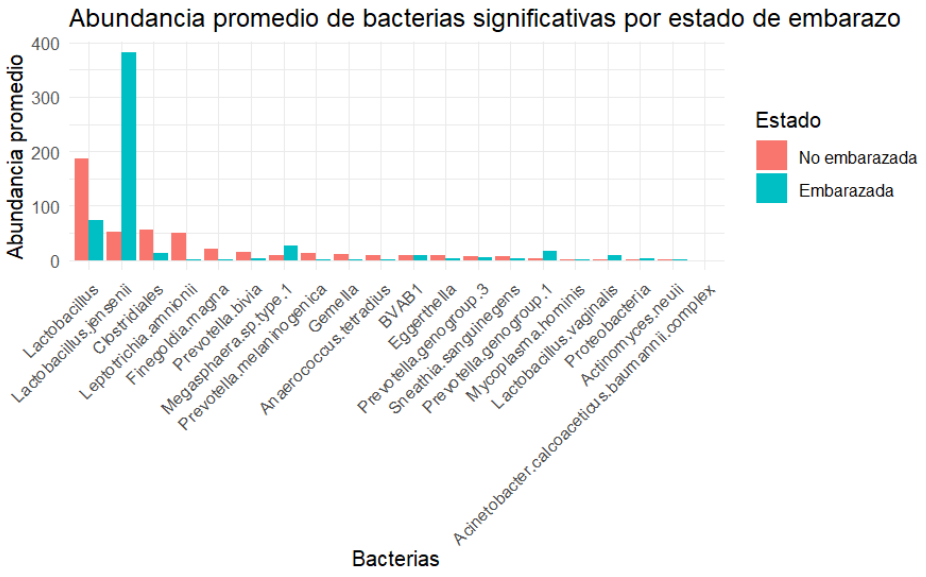


Figura 4.2

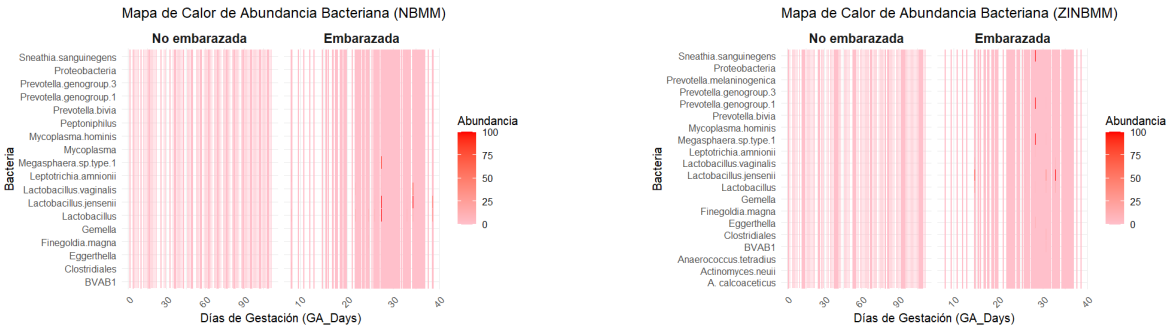
En el gráfico 4.2 que muestra la abundancia promedio de bacterias significativas para el modelo ZINBMM, se observa lo siguiente:

- **Ajuste por ceros:** El modelo ZINBMM identifica más bacterias significativas que el NBMM debido a su capacidad para ajustar los ceros inflados. Esto permite detectar patrones adicionales en la composición bacteriana.

- **Dominancia de bacterias protectoras:** Al igual que en el modelo NBMM, *Lactobacillus* sigue siendo dominante en mujeres gestantes, reforzando su rol protector durante la gestación.
- **Implicaciones:** La capacidad del modelo ZINBMM para manejar ceros inflados permite una comprensión más precisa de la composición bacteriana, destacando diferencias clave entre gestantes y no gestantes.

4.5.5. Mapa de Calor de Abundancia Bacteriana

El mapa de calor (heatmap) es una herramienta que permite visualizar de forma integrada la abundancia relativa de las bacterias significativas a lo largo del tiempo y su relación con el estado de gestación. En la Figura 4.3 se presentan los resultados para mujeres gestantes (`1`) y no gestantes (`0`), utilizando tanto el modelo NBMM como el modelo ZINBMM.



(a) Modelo NBMM

(b) Modelo ZINBMM

Figura 4.3: Comparación de heatmaps para los modelos NBMM y ZINBMM.

Comparación entre NBMM y ZINBMM

Ambos modelos capturan patrones biológicamente relevantes, pero con matices:

- **NBMM:** Identifica de manera clara los taxones dominantes y las diferencias globales de abundancia entre grupos, resaltando la fuerte estabilidad de *Lactobacillus* en mujeres gestantes. Sin embargo, suponer que todos los ceros provienen del mismo mecanismo puede limitar la detección de bacterias poco abundantes.
- **ZINBMM:** Ajusta explícitamente por ceros inflados, separando la probabilidad de ausencia de la abundancia condicionada a presencia. Esto permite detectar taxones adicionales cuya señal podría quedar enmascarada bajo un modelo NBMM, especialmente en mujeres no gestantes.

Patrones observados en mujeres gestantes

- **Dominancia de *Lactobacillus*:** Tanto en NBMM como en ZINBMM, las especies de *Lactobacillus* (*L. jensenii*, *L. crispatus*) mantienen niveles altos y constantes a lo largo de

la gestación.

- **Estabilidad temporal:** La homogeneidad de colores en el heatmap refleja un equilibrio microbiano sostenido, probablemente influenciado por cambios hormonales y condiciones fisiológicas que favorecen un ambiente protector.

Patrones observados en mujeres no gestantes

- **Mayor heterogeneidad:** El heatmap revela variaciones abruptas en abundancia y una mayor diversidad de taxones, con presencia intermitente de bacterias anaeróbicas.
- **Reducción de *Lactobacillus*:** La menor presencia y variabilidad en su abundancia sugiere un microbioma más susceptible a alteraciones.

Conclusiones integradas

El análisis visual mediante heatmaps confirma que:

1. La gestación se asocia con un microbioma vaginal más estable y dominado por bacterias protectoras.
2. Las mujeres no gestantes presentan un patrón más variable, con fluctuaciones marcadas y menor presencia de *Lactobacillus*.
3. El modelo ZINBMM ofrece una caracterización más fina al identificar taxones que el NBMM podría subestimar, resaltando la importancia de modelar ceros estructurales en datos de microbiota.

4.6. Conclusión general del estudio

Este estudio examinó la dinámica longitudinal de la microbiota vaginal en mujeres gestantes y no gestantes mediante dos herramientas de modelación especialmente adecuadas para datos de conteo: el *Modelo Mixto Binomial Negativo* (NBMM) y el *Modelo Mixto Binomial Negativo Inflado en Ceros* (ZINBMM). El objetivo principal fue cuantificar la relación entre la abundancia de taxones específicos y covariables clave—los días gestacionales (`GA_days`) y el estado gestacional (`preg`). A partir de los resultados se obtienen las conclusiones que se detallan a continuación.

Importancia de los modelos NBMM y ZINBMM

- El **NBMM** permitió modelar adecuadamente la sobredispersión inherente a los conteos bacterianos y evaluar el efecto de las covariables sobre la microbiota vaginal.
- El **ZINBMM** fue esencial para tratar la gran proporción de ceros, distinguiendo entre ceros estructurales (ausencia biológica) y ceros aleatorios (limitaciones técnicas de secuenciación).
- Ambos modelos incorporaron efectos aleatorios a nivel de sujeto, capturando la variabilidad intraindividual en datos temporales y proporcionando estimaciones robustas para `GA_days` y `preg`.

Relación entre la gestación y la microbiota vaginal

- **Mayor abundancia de bacterias protectoras.** Las especies del género *Lactobacillus*—en particular *L. jensenii* y *L. crispatus*—presentaron una abundancia significativamente mayor en mujeres gestantes, respaldando su papel en la protección contra infecciones.
- **Estabilidad durante la gestación.** La abundancia de *Lactobacillus* se mantuvo alta y estable a lo largo de los días de gestación, lo que sugiere un equilibrio microbiano favorecido por cambios hormonales y fisiológicos propios de la gestación.
- **Variabilidad en mujeres no gestantes.** Las participantes no gestantes mostraron una microbiota más diversa y variable, con menor presencia de *Lactobacillus* y mayor representación de taxones anaeróbicos.

Comparación entre NBMM y ZINBMM

- El **NBMM** identificó 18 bacterias cuya abundancia se asoció significativamente con `preg` y/o su interacción con el tiempo (`Days`).
- El **ZINBMM** detectó 20 bacterias significativas, mostrando mayor sensibilidad gracias al ajuste explícito de los ceros inflados y revelando patrones más sutiles en taxones menos abundantes.

Estos hallazgos subrayan la relevancia de seleccionar metodologías que capten la complejidad de los datos longitudinales de microbiota.

Interpretación de los resultados visuales

- La **abundancia promedio** de *Lactobacillus* fue sustancialmente superior en gestantes, lo que se interpreta como un efecto causal plausible: estas bacterias, por su reconocida función protectora frente a infecciones y su capacidad para mantener un pH vaginal estable, tienden a establecerse en mayor proporción durante la gestación. De este modo, no solo se observa una asociación estadística, sino también un respaldo biológico que explica por qué su presencia es más marcada en mujeres gestantes.
- La **distribución de abundancia** mostró menor dispersión en gestantes, indicando una colonización más estable por *Lactobacillus*.
- Los **mapas de calor** evidenciaron que la microbiota de gestantes mantiene un perfil homogéneo a lo largo del tiempo, mientras que en no gestantes predominan la variabilidad y la diversidad taxonómica.

Relevancia biológica y clínica

- Durante la gestación se constata una mayor estabilidad del ecosistema vaginal, dominado por *Lactobacillus*, lo que ha sido ampliamente reportado en la literatura médica como un

factor protector frente a infecciones y complicaciones (véase, por ejemplo, (Parakriti Gupta, 2020)), obteniendo así resultados consistentes con esta evidencia.

- La mayor diversidad observada en mujeres no gestantes concuerda con reportes previos que asocian este patrón a un riesgo incrementado de disbiosis, incluyendo vaginosis bacteriana (Parakriti Gupta, 2020).
- El uso del **ZINBMM** resulta particularmente valioso en estudios de microbiota con ceros inflados, ya que permite identificar patrones relevantes vinculados a la ausencia de determinados taxones, aportando robustez al análisis estadístico.

Conclusión final

En síntesis, el estudio demuestra que el estado gestacional influye profundamente en la composición y estabilidad de la microbiota vaginal, promoviendo un entorno dominado por *Lactobacillus*. La aplicación conjunta de NBMM y ZINBMM permitió identificar bacterias clave cuya abundancia varía según el estado gestacional y su evolución temporal, siendo el ZINBMM especialmente eficaz frente a los ceros inflados. Estos resultados refuerzan la importancia de emplear modelos adecuados en estudios longitudinales de microbiota y contribuyen al entendimiento de los factores que modulan la salud vaginal materna y, potencialmente, la salud fetal.

Disponibilidad de código

El código completo de este Estudio I está disponible en el repositorio GitHub:

https://github.com/GabrielaGutierrez2025/tesis-microbiota-plantines/blob/main/microbiota/microbiota.Rmd

Capítulo 5

Caso de Estudio II – Crecimiento longitudinal de plantines

Este capítulo presenta el segundo estudio desarrollado en esta memoria, centrado en el análisis del crecimiento longitudinal de *Quillaja saponaria* (quillay) en vivero, en el marco de un experimento aplicado en el Jardín Botánico Nacional de Viña del Mar. A diferencia de estudios simulados o académicamente controlados, aquí se trabaja con datos reales obtenidos por un equipo de investigación de la Escuela de Ingeniería Ambiental, en colaboración con profesionales del área ecológica y forestal. Este aspecto, aunque enriquecedor, representa un desafío no menor en el contexto estadístico: la naturaleza misma de los datos, las condiciones de campo, y las decisiones prácticas durante la recolección imponen restricciones y consideraciones que deben ser tenidas en cuenta en cada etapa del análisis.

Adicionalmente, este estudio se desarrolló en estrecha colaboración con investigadores de distintas formaciones, cuyas perspectivas complementarias aportaron enfoques experimentales y biológicos al fenómeno estudiado. Esta experiencia potenció la comunicación interdisciplinaria y la capacidad de traducir preguntas biológicas y ambientales en formulaciones cuantitativas adecuadas para modelación estadística, favoreciendo un trabajo conjunto enriquecedor.

Previo al ajuste de modelos, se realizó una revisión exhaustiva de la literatura con el fin de identificar las estrategias estadísticas más apropiadas para el análisis de datos longitudinales en contextos ecológicos y de producción vegetal. Entre las alternativas evaluadas, los modelos mixtos lineales (LMM) emergen como una herramienta versátil y potente para capturar tanto la estructura jerárquica de los datos como la variabilidad interindividual en las trayectorias de crecimiento.

Las secciones que siguen describen el contexto experimental del estudio, el objetivo estadístico, la estructura de los datos recopilados, y los modelos propuestos para representar y comparar el crecimiento de los plantines bajo distintos tratamientos bioestimulantes.

5.1. Objetivo y contexto del estudio

Marco general del proyecto

Los datos analizados provienen de un proyecto de dos años desarrollado por la Escuela de Medio Ambiente cuyo objetivo central es *formular una estrategia de producción y aplicación de cianobacterias fijadoras de nitrógeno para la recuperación de bosques nativos afectados por incendios forestales*. Dentro de esta meta global, la **etapa de plantines** persigue un objetivo específico: evaluar el efecto bioestimulante de *Anabaena sp.* sobre el crecimiento temprano de *Quillaja saponaria* (quillay) bajo condiciones controladas de vivero, y generar información base para la posterior etapa de reforestación en terreno.

Objetivo estadístico de este estudio

Mientras que el proyecto biológico evalúa la eficacia bioestimulante de los tratamientos, **el propósito del presente estudio es puramente estadístico**:

> *identificar el modelo mixto lineal (LMM) que mejor describa y pronostique la trayectoria longitudinal de crecimiento—altura o DAC—de los plantines bajo los cuatro tratamientos mencionados.*

Con este fin se ajustarán y compararán diferentes especificaciones de LMM (intercepto aleatorio, pendiente aleatoria, interacción tiempo × tratamiento, incorporación de temperatura acumulada, etc.), evaluando su capacidad predictiva mediante AIC, BIC y pruebas de razón de verosimilitud.

Ubicación y descripción del sitio experimental

El vivero se instaló en la zona de educación ambiental del Jardín Botánico Nacional de Viña del Mar (Región de Valparaíso, Chile), cuyas coordenadas UTM son 19H $E = 266\,315{,}21$ m, $N = 6\,340\,422{,}52$ m. Para proporcionar un micro-ambiente estable se construyó un *sombreadero* de $3{,}2\,\text{m} \times 9{,}6\,\text{m}$ cubierto con malla raschel de 60 % de sombreo y malla gallinero perimetral. En su interior se dispusieron mesones de madera que soportan las bandejas de germinación y los contenedores individuales utilizados tras el repique.

Recursos e infraestructura

- **Materiales de construcción**: madera, malla raschel 60 %, malla antimaleza, malla gallinero.
- **Materiales de viverización**: bandejas forestales de 50 alveolos, contenedores de polietileno (25 cm × 15 cm) y etiquetas plásticas.
- **Sustratos**: mezcla 2:1:1:1 de corteza de pino compostada, compost vegetal, fibra de coco y perlita.
- **Herramientas**: palas, rastrillos, regaderas, pie de metro, higrómetros, entre otros.

Diseño experimental y tratamientos

Para asegurar la homogeneidad inicial se adquirieron **2 000** plantines de quillay con edad y tamaño similares, repicados (es decir, trasplantados) entre el 10 y 14 de marzo de 2025 a los contenedores individuales descritos. Se establecieron **4** tratamientos ($n = 500$ plantines cada uno), aplicando las soluciones por imbibición al momento del repique:

Trat.	**Descripción**	**Dosis / Volumen**
A	Agua potable (control negativo)	50 mL plantín^{-1}
B	Bioestimulante comercial (*Algafert®– Spirulina*)	50 mL plantín^{-1} a 3 mL L^{-1}
C	*Anabaena sp.* (dosis 1)	8 g m^{-2} de inóculo
D	*Anabaena sp.* (dosis 2)	16 g m^{-2} de inóculo

Variables registradas

Símbolo / nombre	**Descripción y unidades**
$Y_{ij}^{(\mathrm{Alt})}$, `altura`	Altura total (cm) de la planta i en la fecha j.
$Y_{ij}^{(\mathrm{DAC})}$, `dac`	Diámetro a nivel del cuello (mm) en la planta i, fecha j.
`tratamiento`	Factor categórico con niveles *TA*, *TB*, *TC*, *TD*.
`fecha`	Fecha calendario (tipo `Date`); se transforma a `dia` = días transcurridos desde la primera medición.
`temp_mean_cum`	Temperatura media acumulada (°C · día) desde el trasplante hasta `fecha`.
`humedad_mean_cum`	Humedad relativa media acumulada (%).
`rad_mean_cum`	Radiación solar acumulada (MJ m^{-2}).
`lluvia_sum_cum`	Precipitación acumulada (mm).

Tabla 5.1: Variables incluidas en la base depurada.

Las cuatro covariables ambientales provienen de la hoja `env_acumulado_plantas.xlsx` y se asociaron a cada observación por la columna *fecha*. Todas las variables continuas se escalaron antes de la modelación.

Las mediciones están anidadas jerárquicamente:

$$\text{fecha } j \;\subset\; \text{planta } i \;\subset\; \text{tratamiento } k.$$

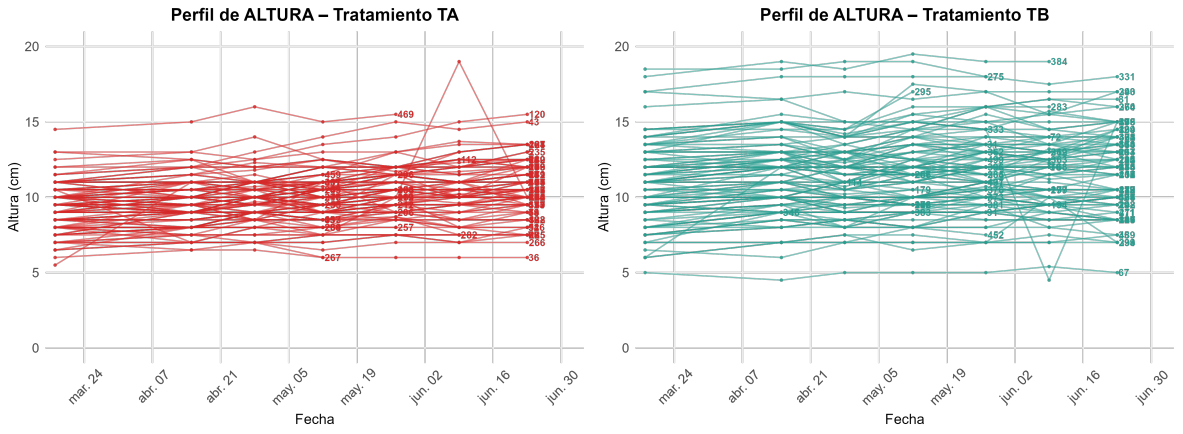
Este esquema motivó el uso de Modelos Mixtos Lineales (LMM), en los cuales *planta* se modela como factor de efectos aleatorios (intercepto y, en algunos casos, pendiente de tiempo), mientras que *tratamiento*, *tiempo* e interacciones se tratan como efectos fijos. El detalle de los modelos comparados se presenta en la Sección 5.4.

5.2. Exploración inicial de trayectorias y supervivencia por tratamiento

5.2.1. Exploración inicial de trayectorias

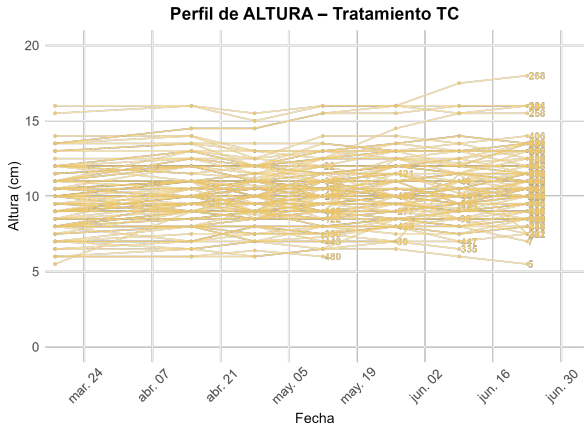
Antes de proceder al ajuste de los Modelos Mixtos Lineales (LMM), se realizó una inspección gráfica detallada de las trayectorias longitudinales de crecimiento de los plantines, considerando dos variables fundamentales: la *altura total* y el *diámetro a la altura del cuello* (DAC). Para ello se emplearon herramientas de visualización en `ggplot2`, generando perfiles individuales para cada una de las plantas en las siete fechas de medición, correspondientes al periodo marzo–junio de 2025. Las trayectorias se agruparon por tratamiento experimental (*TA*, *TB*, *TC* y *TD*), lo que permitió explorar visualmente las tendencias de crecimiento diferenciadas entre condiciones.

En cada gráfico, las plantas fueron representadas mediante líneas conectadas por puntos en cada fecha de medición, con el identificador rotulado en su última observación. Esta estrategia permitió detectar de forma visual valores atípicos, trayectorias inesperadas o registros con comportamiento errático.

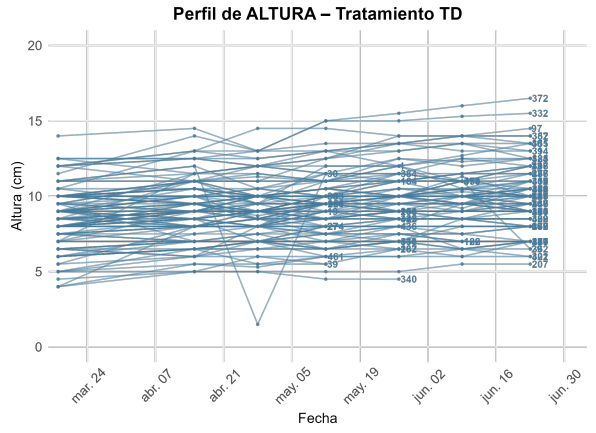


(a) Tratamiento TA

(b) Tratamiento TB



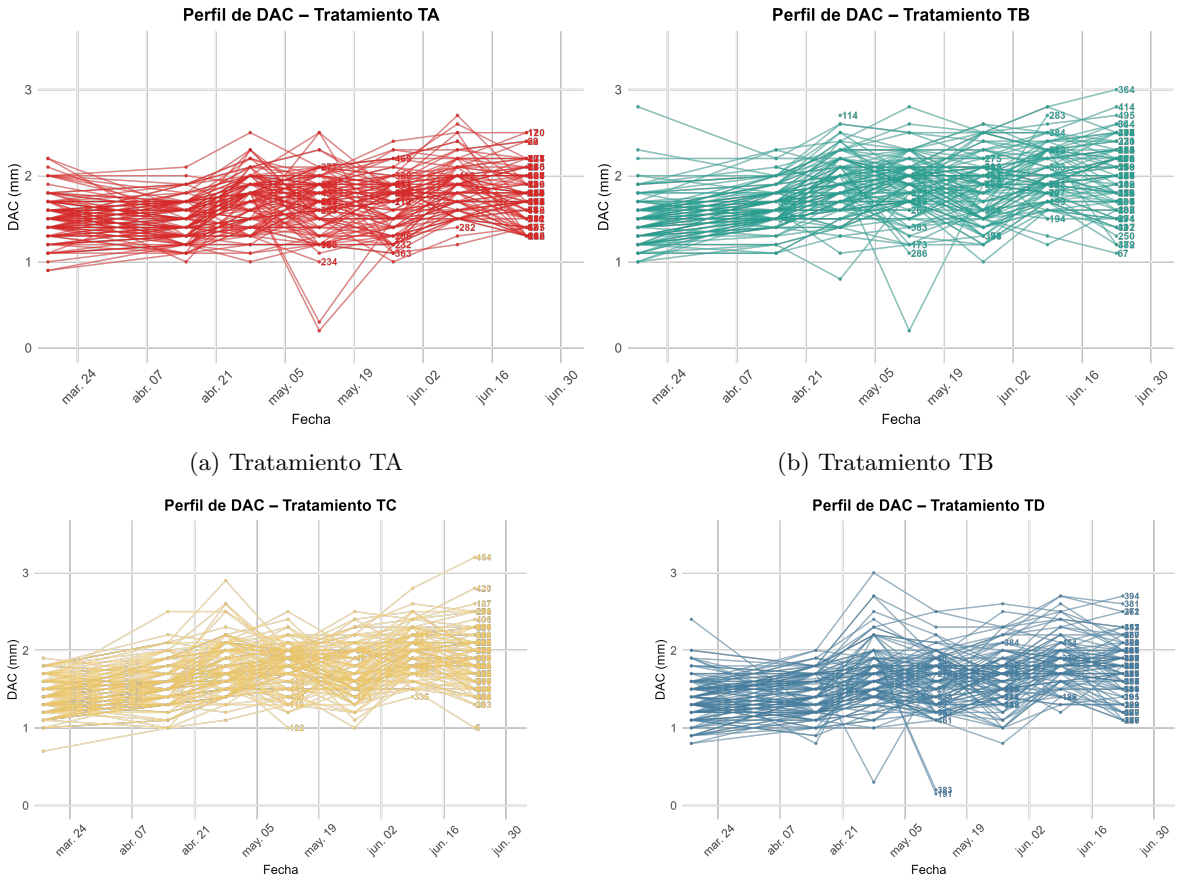
(c) Tratamiento TC



(d) Tratamiento TD

Figura 5.1: Perfiles longitudinales de **altura** para los cuatro tratamientos experimentales. Cada línea representa una planta individual; los identificadores están rotulados en la última fecha para facilitar la detección de valores extremos o trayectorias atípicas.

Como se puede observar en la Figura 5.1, las trayectorias de **altura** muestran patrones diferenciados por tratamiento. En los grupos *TA* y *TC*, la mayoría de las plantas presenta una evolución moderadamente ascendente, con rangos intercuartílicos situados aproximadamente entre 8 y 13 cm. No obstante, también se detectaron algunas trayectorias con pendiente negativa, lo cual podría deberse a variabilidad biológica natural o a errores de medición. El tratamiento *TB* exhibe alturas globalmente mayores, con valores que se extienden hasta los 18 cm y una dispersión más amplia entre plantas. Esto podría reflejar un efecto estimulante del tratamiento aplicado, aunque también podría deberse a una mayor heterogeneidad inicial en el lote de plantines asignado a este grupo. Por su parte, el tratamiento *TD* muestra un patrón más compacto, con alturas que oscilan entre 5 y 13 cm y menor variabilidad relativa.



(a) Tratamiento TA

(b) Tratamiento TB

(c) Tratamiento TC

(d) Tratamiento TD

Figura 5.2: Perfiles longitudinales de **DAC** (diámetro del cuello). La dispersión acotada y la tendencia ascendente generalizada refuerzan su elección como variable de respuesta principal en los modelos mixtos.

En cuanto al **DAC**, representado en la Figura 5.2, se observa un comportamiento más homogéneo entre tratamientos. Los cuatro grupos presentan rangos intercuartílicos estrechos, aproximadamente entre 1 y 2,3 mm, lo que es coherente con la escala reducida de esta variable. A pesar de que la dispersión general es menor en comparación con la altura, se mantiene una ligera tendencia ascendente en casi todos los tratamientos, reflejando un patrón de engrosamiento progresivo del tallo a lo largo del tiempo.

A partir de estas observaciones, se tomó la decisión analítica de adoptar el *diámetro a la altura del cuello (DAC)* como variable de respuesta principal para los modelos mixtos lineales. Esta elección se fundamenta en su menor variabilidad relativa, su escala adecuada para modelación lineal y su relevancia biológica como indicador del desarrollo estructural de la planta. Cabe destacar que la metodología implementada es totalmente replicable para la variable altura, con mínimas modificaciones en los filtros y escalas de los gráficos.

5.2.2. Supervivencia de los plantines por tratamiento

Aunque el objetivo central del Estudio II es identificar el modelo con mejor capacidad predictiva para las trayectorias de crecimiento de los plantines (altura o DAC), también resulta relevante evaluar la *supervivencia hasta la última fecha de medición* como un indicador temprano del desempeño agronómico de los tratamientos aplicados. Esta dimensión es especialmente importante en estudios de aplicación real, donde el éxito de un bioestimulante no solo se mide por su efecto en el crecimiento, sino también por su capacidad de no comprometer la viabilidad de las plantas.

Con este objetivo, se planteó el siguiente contraste de hipótesis:

> $\mathbf{H}_0$: la proporción de plantas que completan el experimento es igual en los cuatro tratamientos.

Una diferencia significativa respecto a esta hipótesis nula implicaría que alguno de los tratamientos genera estrés fisiológico, toxicidad o alguna deficiencia que afecta la permanencia de los individuos a lo largo del ensayo.

El análisis se realizó a partir del conjunto depurado de mediciones, el cual se obtuvo filtrando aquellas plantas que estuvieron presentes en la **primera fecha de medición** ($t_1 = 18$ de marzo de 2025). A partir de este conjunto, se identificó para cada planta si también estaba presente en la última fecha de medición ($t_7 = 23$ de junio de 2025). Con esta información, se definió una variable dicotómica de supervivencia: igual a 1 si la planta estaba presente en ambas fechas, y 0 en caso contrario. Esta variable permitió calcular las tasas de supervivencia por tratamiento.

La Tabla 5.2 resume el número de individuos al inicio y al final del experimento en cada grupo, así como el porcentaje de supervivencia observado. La Tabla 5.2 presenta una visualización clara de estas proporciones, facilitando la comparación entre tratamientos.

Tabla 5.2: Número de plantas iniciales y finales por tratamiento

Tratamiento	n_{inicial}	n_{final}	Supervivencia (%)
TA	125	96	77
TB	125	85	65
TC	124	99	79
TD	125	95	74

Para evaluar si las diferencias observadas son estadísticamente significativas, se aplicó una prueba de χ^2 de homogeneidad con $df = 3$, obteniéndose:

$$\chi^2 = 5{,}11, \qquad p = 0{,}1638.$$

Dado que este valor es mayor al umbral de significancia común ($\alpha = 0{,}05$), no se rechaza la hipótesis nula de homogeneidad, por lo tanto, no se encontraron diferencias estadísticamente significativas en las proporciones de plantas que llegaron al final del experimento entre los distintos tratamientos.

Sin embargo, por lo observado en la Tabla 5.2 en términos prácticos, el tratamiento **TB** fue el que presentó la menor tasa de supervivencia (65 %), con 40 pérdidas documentadas. Este resultado sugiere que el uso del bioestimulante comercial aplicado en este tratamiento podría estar asociado a un mayor nivel de estrés o a una menor tolerancia por parte de los plantines. Por el contrario, el tratamiento **TC**, correspondiente a la dosis 1 de *Anabaena sp.*, alcanzó el mayor porcentaje de supervivencia (79 %), lo que refuerza su perfil favorable como posible alternativa bioestimulante.

Desde una perspectiva metodológica, si bien según el test no se puede asegurar que se tiene diferencias significativas en las proporciones de plantas que llegaron al final del tratamiento, es importante considerar que la exclusión de las plantas fallecidas podría conducir a una ligera subestimación de la varianza residual.

En síntesis, las diferencias detectadas en supervivencia no solo justifican la modelación diferenciada por tratamiento, sino que también aportan evidencia preliminar sobre la tolerancia de los plantines a las distintas soluciones aplicadas.

5.3. Preparación de datos para modelos mixtos

Antes de ajustar los modelos mixtos lineales (LMM) para describir y predecir la evolución del diámetro a la altura del cuello (DAC) en los plantines, fue necesario preparar cuidadosamente la base de datos a utilizar. Este proceso implicó una serie de transformaciones orientadas a organizar la información en un formato compatible con la estructura jerárquica del análisis, integrando variables experimentales y ambientales.

En primer lugar, se codificó el tiempo como una variable numérica que indicara los días transcurridos desde la primera medición. Esta transformación fue crucial para modelar tendencias de crecimiento de forma continua y permitir la estimación de pendientes temporales tanto fijas como aleatorias.

Luego, se filtraron los datos para centrarse exclusivamente en la variable DAC, definida previamente como respuesta principal del estudio. Cada planta se identificó de forma única, y se conservaron únicamente aquellas observaciones completas y consistentes a lo largo del período de estudio.

El tratamiento asignado a cada planta se representó mediante variables categóricas, lo que permitió incluir directamente cada nivel de tratamiento como un efecto fijo en los modelos. Al trabajar con efectos fijos categóricos sin establecer un grupo de referencia, cada coeficiente asociado a un tratamiento refleja explícitamente la diferencia de ese tratamiento respecto a la media general del conjunto de datos, en lugar de compararse únicamente con un grupo base. Este enfoque facilita interpretaciones más directas: por ejemplo, es posible evaluar de manera individual el efecto de cada tratamiento sobre la variable de respuesta sin introducir sesgos derivados de la elección arbitraria de un tratamiento de referencia. Además, permite realizar comparaciones simultáneas entre tratamientos, lo que resulta particularmente útil para estudios experimentales con múltiples condiciones, ya que se puede cuantificar de forma clara cómo varía la respuesta de los plantines bajo distintos bioestimulantes.

Posteriormente, se integraron las covariables ambientales acumuladas (temperatura, humedad, radiación y precipitación), obtenidas a partir de registros externos y asociadas a cada fecha de medición. Estas variables resumen la exposición ambiental reciente de los plantines y reflejan condiciones que pueden influir en su desarrollo. Al tratarse de valores acumulados, se mitigan las fluctuaciones diarias y se capta mejor la "historia climática" relevante para cada observación.

Con la información organizada y enriquecida, se generó un conjunto de datos estructurado para el ajuste de modelos, donde cada fila representa una medición específica de una planta en una fecha determinada, e incluye tanto las variables explicativas como la respuesta. Esta preparación permitió establecer una base sólida para comparar distintas especificaciones de modelos mixtos, explorando la contribución de los tratamientos, el tiempo y las condiciones ambientales al crecimiento del DAC.

El diseño de modelos considerado se resume en tres dimensiones clave: (i) inclusión de covariables ambientales y del tiempo como predictores fijos, (ii) incorporación de efectos aleatorios por planta (ya sea solo intercepto o intercepto más pendiente temporal), y (iii) posibilidad de modelar interacciones entre tratamiento y tiempo para capturar patrones diferenciados de crecimiento. Los detalles de estos modelos y su evaluación se presentan en las secciones siguientes.

5.4. Modelos mixtos lineales evaluados

5.4.1. Formulación general del modelo mixto lineal y estrategia de construcción

En este capítulo se exploran distintas formulaciones de modelos mixtos lineales (LMM) para describir la evolución temporal del diámetro a la altura del cuello (DAC) en los plantines sometidos a cuatro tratamientos experimentales. Antes de detallar cada una de las especificaciones contrastadas, presentamos la estructura general del modelo y la lógica progresiva seguida en su construcción.

Modelo general y componentes estructurales

El modelo mixto lineal responde a la necesidad de capturar tanto efectos poblacionales comunes (como el tratamiento o las condiciones ambientales) como variaciones individuales entre plantas. En términos generales, el modelo puede expresarse como

$$Y_{ij} = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\top}\mathbf{b}_i + \varepsilon_{ij}, \tag{5.1}$$

con $i = 1, \ldots, n$ y $j = 1, \ldots, n_i$ ($n_i \in \{2, 3, 4, 5, 6, 7\}$), donde:

- Y_{ij} representa la medición de DAC de la planta i en la fecha j,
- $\mathbf{x}_{ij}$ es el vector de covariables fijas asociadas a la observación ij,
- $\boldsymbol{\beta}$ contiene los coeficientes poblacionales (efectos fijos),

- $\mathbf{z}_{ij}$ es el diseño de efectos aleatorios para la planta i,
- $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$ representa los efectos aleatorios específicos de la planta,
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ es el término de error residual.

En este contexto, cada observación individual Y_{ij} para la planta i en la fecha j puede representarse como:

$$\begin{aligned} Y_{ij} \;=\; & \underbrace{\beta_0 + \beta_1\,\mathrm{dia}_{ij} + \boldsymbol{\beta}_2^\top \mathbf{T}_i + \boldsymbol{\beta}_3^\top (\mathrm{dia}_{ij} \cdot \mathbf{T}_i) + \boldsymbol{\beta}_4^\top \mathbf{E}_j}_{\text{efectos fijos}} \\ & + \underbrace{b_{0i} + b_{1i}\,\mathrm{dia}_{ij}}_{\text{efectos aleatorios}} + \varepsilon_{ij}. \end{aligned} \tag{5.2}$$

Los términos involucrados son los siguientes:

- dia_{ij} es el número de días transcurridos desde la primera medición (18-mar-2025), codificado como un entero entre 1 y 97.
- $\mathbf{T}_i$ es el vector de indicadores (dummies) de tratamiento para la planta i, que permite distinguir entre TA (referencia), TB, TC y TD.
- $\mathrm{dia}_{ij} \cdot \mathbf{T}_i$ representa el conjunto de variables de interacción entre el número de días y las dummies de tratamiento (TB, TC y TD), es decir, términos del tipo $\mathrm{dia}_{ij} \cdot \mathbb{1}_{\mathrm{TB}_i}$, $\mathrm{dia}_{ij} \cdot \mathbb{1}_{\mathrm{TC}_i}$, etc., que permiten estimar *pendientes específicas* para cada tratamiento, modelando así diferencias en la tasa de crecimiento entre grupos.
- $\mathbf{E}_j$ corresponde a las covariables ambientales acumuladas hasta la fecha j, incluyendo temperatura, humedad, radiación y precipitación.
- b_{0i} y b_{1i} son el intercepto y la pendiente aleatorios de la planta i, con distribución conjunta normal.
- ε_{ij} es el error residual asociado a cada observación.

Estrategia de construcción y comparación

La formulación del modelo no fue única, sino que siguió una lógica escalonada. Se ajustaron ocho especificaciones (denotadas M1 a M8), cada una añadiendo complejidad sobre la anterior en dos dimensiones principales:

- **Estructura aleatoria**: comenzando con modelos que consideran solo interceptos aleatorios y luego incorporando también pendientes aleatorias por planta, permitiendo así trayectorias individuales más flexibles.
- **Efectos fijos**: partiendo desde modelos con tratamiento como único predictor, y añadiendo progresivamente el tiempo, las covariables ambientales acumuladas, y finalmente interacciones del día con el tratamiento.

El criterio para seleccionar los modelos más apropiados se basó en una combinación de indicadores estadísticos: AIC, BIC y pruebas de razón de verosimilitud (*Likelihood Ratio Test*) entre modelos

anidados. Todos los modelos fueron ajustados mediante el paquete `nlme` de R utilizando el método REML para estimar los parámetros.

Los detalles específicos de cada modelo, junto con sus coeficientes estimados, significancia estadística y métricas comparativas, se resumen en las Tablas 5.3 y 5.4.

5.4.2. Modelos mixtos lineales evaluados

Con el objetivo de modelar el crecimiento longitudinal del diámetro a la altura del cuello (DAC) en plantines bajo distintos tratamientos y condiciones ambientales, se ajustaron ocho modelos mixtos lineales de complejidad creciente, en donde estos modelos se enmarcan en la formulación general 5.1.

En las siguientes dos tablas se resumen, respectivamente, (i) los coeficientes estimados para cada modelo junto con sus significancias y criterios de ajuste global, y (ii) la estructura de diseño de cada modelo en términos de los predictores incluidos en los componentes fijos (X) y aleatorios (Z).

Tabla 5.3: Resumen comparativo de los ocho modelos mixtos lineales ajustados. Se presentan los coeficientes estimados ($\hat{\beta}_i$) y sus respectivos p-valores entre paréntesis. El coeficiente $\hat{\beta}_0$ corresponde al intercepto, mientras que $\hat{\beta}_1$ al día. Los coeficientes $\hat{\beta}_2$ a $\hat{\beta}_4$ corresponden a los tratamientos TB, TC y TD respectivamente, $\hat{\beta}_5$ a $\hat{\beta}_8$ corresponden a las variables ambientales acumuladas: temperatura (`temp_cum`), humedad (`hum_cum`), radiación (`rad_cum`) y lluvia (`rain_cum`). Además, $\hat{\beta}_9$, $\hat{\beta}_{10}$ y $\hat{\beta}_{11}$ representan las interacciones entre el tiempo (`día`) y los tratamientos TB, TC y TD, respectivamente.

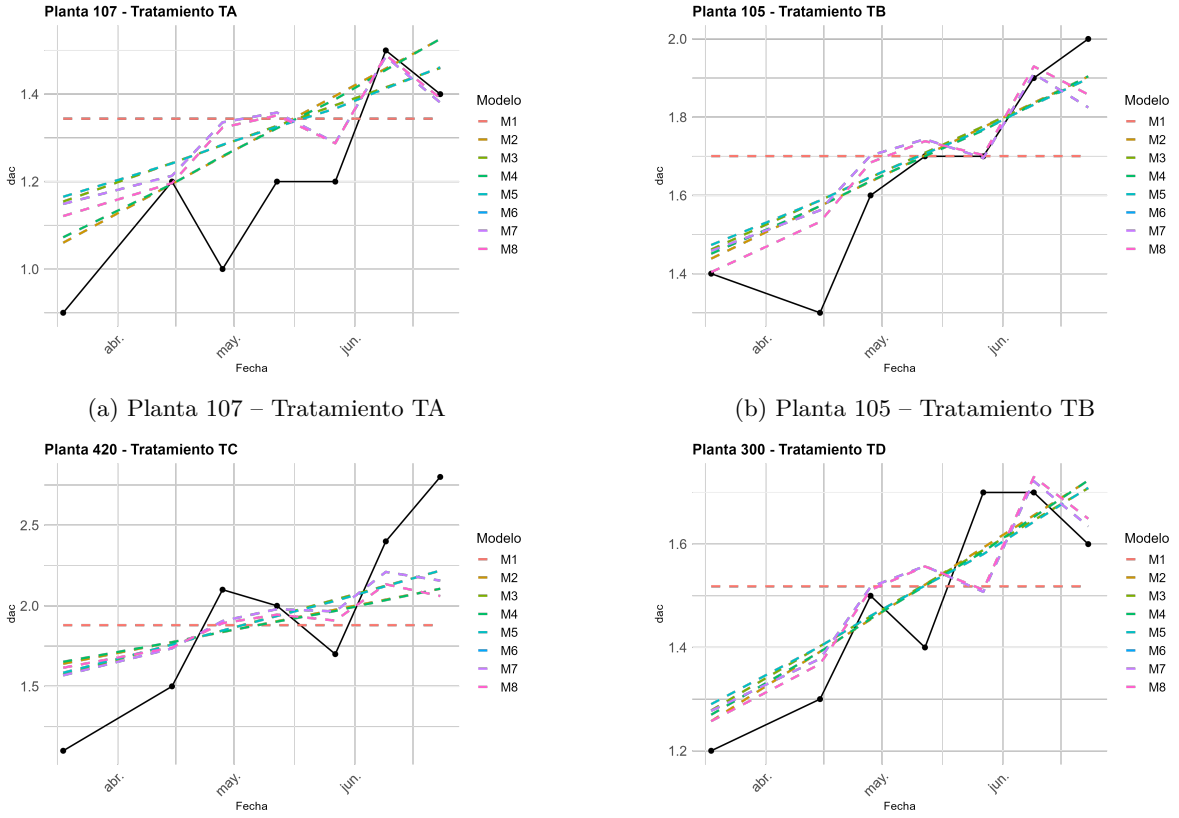
Modelo	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	AIC	BIC	LogLik
M1	1.67 (0.00)	–	0.12 (0.00)	0.06 (0.0015)	-0.07 (0.0001)	–	–	–	–	–	–	–	2010.42	2046.85	-999.21
M2	1.43 (0.00)	0.04 (0.00)	0.12 (0.00)	0.05 (0.02)	-0.08 (0.00)	–	–	–	–	–	–	–	1211.62	1254.11	-598.81
M3	1.44 (0.00)	0.004 (0.00)	0.11 (0.00)	0.04 (0.01)	-0.09 (0.00)	–	–	–	–	–	–	–	1190.21	1244.85	-586.10
M4	1.53 (0.00)	0.003 (0.16)	0.12 (0.00)	0.05 (0.002)	-0.08 (0.00)	-0.06 (0.35)	–	–	–	–	–	–	1216.39	1264.95	-600.19
M5	1.54 (0.00)	0.002 (0.16)	0.11 (0.00)	0.04 (0.01)	-0.09 (0.00)	-0.06 (0.35)	–	–	–	–	–	–	1195.31	1256.01	-587.65
M6	1.66 (0.04)	0.0005 (0.96)	0.11 (0.00)	0.04 (0.01)	-0.09 (0.00)	-0.88 (0.00)	-0.13 (0.00)	0.53 (0.07)	-0.14 (0.00)	–	–	–	1051.67	1130.57	-512.83
M7	1.66 (0.04)	0.0005 (0.96)	0.11 (0.00)	0.04 (0.01)	-0.09 (0.00)	-0.88 (0.00)	-0.13 (0.00)	0.53 (0.07)	-0.14 (0.00)	–	–	–	**1051.67**	**1130.57**	-512.83
M8	1.74 (0.03)	-0.001 (0.94)	0.02 (0.30)	-0.03 (0.19)	-0.15 (0.00)	-0.89 (0.00)	-0.13 (0.00)	0.52 (0.08)	-0.14 (0.00)	0.001 (0.00)	0.001 (0.00)	0.001 (0.003)	1099.72	1184.67	-535.86

Nota: Aunque algunos coeficientes presentan p-value mayores a 0,05, esto no implica que el modelo deba rechazarse. En modelos mixtos, la dependencia entre efectos fijos y aleatorios, la inclusión de covariables de control y criterios de ajuste global como AIC, BIC y log-verosimilitud son considerados para la selección del modelo. Asimismo, la interpretación se complementa con el conocimiento biológico y la plausibilidad de los efectos, evitando decisiones basadas únicamente en la significancia estadística individual.

Tabla 5.4: Especificación de las matrices de diseño $\mathbf{X}$ (efectos fijos) y $\mathbf{Z}$ (efectos aleatorios) para cada modelo M_1–M_8. Se marca con ✓ si el término está presente.

Bloque	**Término**	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
Matriz X: efectos fijos									
	Intercepto fijo	✓	✓	✓	✓	✓	✓	✓	✓
	Día		✓	✓	✓	✓	✓	✓	✓
	Tratamiento (TB–TD)	✓	✓	✓	✓	✓	✓	✓	✓
	Temp. acumulada				✓	✓	✓	✓	✓
	Humedad acumulada						✓	✓	✓
	Radiación acumulada						✓	✓	✓
	Precipitación acumulada						✓	✓	✓
	Día × Tratamiento								✓
Matriz Z: efectos aleatorios (por planta)									
	Intercepto aleatorio	✓	✓	✓	✓	✓	✓		✓
	Pendiente aleatoria (día)			✓		✓	✓	✓	

Tras ajustar y comparar los ocho modelos mixtos lineales descritos, se puede observar una evolución progresiva tanto en el poder explicativo como en la adecuación visual de las trayectorias estimadas. A medida que se incorporan nuevas covariables y se complejiza la estructura aleatoria, los modelos logran capturar con mayor precisión la dinámica de crecimiento del DAC en las distintas condiciones experimentales.



(a) Planta 107 – Tratamiento TA

(b) Planta 105 – Tratamiento TB

(c) Planta 420 – Tratamiento TC

(d) Planta 300 – Tratamiento TD

Figura 5.3: Curvas de crecimiento observadas (curva negra) y predichas (líneas discontinuasde colores) por los ocho modelos mixtos lineales (M1–M8) para cuatro plantas seleccionadas, una por tratamiento.

Para apoyar esta comparación, se graficaron las trayectorias predichas por cada modelo sobre los datos reales de cuatro plantas seleccionadas (una por tratamiento)(Figura 5.3). Estas curvas permiten evaluar visualmente cómo la incorporación sucesiva de efectos mejora el ajuste a lo largo del tiempo. Las figuras muestran que los modelos iniciales tienden a subestimar el crecimiento o a capturarlo de forma muy rígida, mientras que las versiones posteriores reflejan de manera más fiel las pendientes diferenciadas entre tratamientos y el efecto modulador de las condiciones ambientales.

Síntesis y conclusiones

La tabla comparativa y la estructura matricial de los modelos evaluados permiten visualizar con claridad la progresión de complejidad en las ocho especificaciones consideradas. En particular, se observa que la inclusión de covariables ambientales acumuladas (a partir de M4), así como la incorporación de pendientes aleatorias por planta (M3, M5–M7), contribuyen a una mejora progresiva en los indicadores de ajuste (log-verosimilitud, AIC y BIC), con un rendimiento destacable de los modelos M6, M7 y M8.

A modo complementario, se elaboraron gráficos de ajuste individual sobre cuatro plantas representativas —una por cada tratamiento— donde se comparan visualmente las trayectorias predichas por los ocho modelos. Estas curvas permiten evidenciar que, si bien los modelos iniciales (M1–M2) capturan parcialmente el crecimiento promedio, tienden a subestimar la heterogeneidad individual, especialmente en tratamientos con mayor variabilidad intra-grupo. En cambio, los modelos con estructura aleatoria más completa (como M6–M8) logran un seguimiento más preciso de la dinámica de cada planta, reflejando tanto su patrón específico como el efecto acumulado del ambiente.

Al observar el comportamiento de los modelos evaluados, se aprecia cómo la incorporación progresiva de términos mejora la capacidad de ajuste. Si bien los modelos más simples permiten una primera aproximación a las tendencias generales del crecimiento, es en las especificaciones más completas donde se logra capturar con mayor precisión la influencia combinada del tiempo, los tratamientos y el entorno ambiental.

Dentro de estos, se destaca el modelo **M7**, que incorpora efectos aleatorios en la pendiente por planta junto con todos los efectos fijos relevantes (tiempo, tratamiento y clima). Este modelo logra un equilibrio notable entre parsimonia y capacidad explicativa. De hecho, cabe mencionar que M6 y M7 entregan resultados idénticos en términos de valores estimados y métricas de ajuste (log-verosimilitud, AIC y BIC), lo cual podría sugerir que ambos modelos son equivalentes. Sin embargo, su estructura aleatoria difiere: mientras M6 considera intercepto y pendiente aleatoria por planta, M7 únicamente incluye pendiente aleatoria. Esto revela que el intercepto aleatorio no aporta información adicional sustantiva en esta etapa del análisis, y que el componente clave en la heterogeneidad individual se encuentra en la tasa de crecimiento, más que en el valor inicial del DAC. Por razones prácticas de simplicidad interpretativa y sin pérdida de ajuste, se adopta **M7** como el modelo óptimo para el resto del estudio.

Según los resultados de este modelo, el intercepto estimado es $\hat{\beta}_0 = 1{,}66$, lo que representa el valor promedio del DAC al inicio del seguimiento para plantas del tratamiento de referencia (TA) bajo condiciones promedio de clima.

El coeficiente asociado al tiempo (`dia`) es $\hat{\beta}_1 = 0{,}0005$, lo cual indica un crecimiento diario promedio de 0,0005 mm, ajustado por tratamiento y ambiente. Respecto a los efectos de tratamiento, se estimaron $\hat{\beta}_2 = 0{,}11$ para TB, $\hat{\beta}_3 = 0{,}04$ para TC y $\hat{\beta}_4 = -0{,}09$ para TD, lo que sugiere que TB genera un crecimiento adicional considerable en comparación con TA, mientras que TD incluso muestra un leve efecto negativo.

En cuanto a las covariables climáticas, se obtuvo $\hat{\beta}_5 = -0{,}88$ para la temperatura acumulada, lo que indica que mayores temperaturas sostenidas tienden a inhibir el crecimiento del DAC, posiblemente por estrés térmico. La humedad acumulada presenta un efecto moderadamente negativo ($\hat{\beta}_6 = -0{,}13$), mientras que la radiación acumulada tiene un efecto positivo importante ($\hat{\beta}_7 = 0{,}53$), coherente con su rol en la fotosíntesis. Finalmente, la precipitación acumulada se asocia negativamente ($\hat{\beta}_8 = -0{,}14$), posiblemente por saturación del suelo o drenaje deficiente.

Por todos estos motivos —ajuste superior, parsimonia, interpretabilidad y coherencia biológica— el modelo **M7** se considera el más adecuado en esta etapa del análisis.

Cabe destacar que el detalle completo de los resultados obtenidos para cada modelo —incluyendo los coeficientes estimados, sus errores estándar, valores p, así como las métricas de ajuste como log-verosimilitud, AIC y BIC— se encuentra disponible en el **Anexo B**. Dicho anexo complementa la síntesis presentada en esta sección, ofreciendo una visión integral y reproducible del proceso de ajuste y comparación de los modelos lineales mixtos evaluados.

5.4.3. Limitaciones de los datos y justificación para no ajustar NLMM

Antes de plantear un *modelo mixto no lineal* (NLMM)—por ejemplo logístico, Gompertz o Richards—se evaluó la pertinencia de tal complejidad a la luz de las características empíricas del ensayo de plantines. Los siguientes puntos resumen las restricciones que motivaron mantener la familia lineal en esta fase del proyecto.

L1. Curvatura poco manifiesta. Las siete mediciones disponibles abarcan apenas ~ 100 días (18 mar–23 jun 2025). Los perfiles exploratorios muestran pendientes casi constantes dentro de este intervalo —incluso después de estratificar por tratamiento—, sin evidencias visuales de desaceleración ni de meseta típicas del crecimiento sigmoideo que justificarían un NLMM.

L2. Horizonte temporal acotado. Modelos como Gompertz requieren observar una fracción sustancial de la vida útil de la planta para estimar de forma identificable la asíntota y el punto de inflexión. Con apenas cuatro meses de seguimiento el crecimiento sigue en fase aproximadamente exponencial, de modo que la convergencia de un NLMM quedaría fuertemente dictada por las condiciones iniciales y no por la información contenida en los datos.

L3. Orientación del estudio. El objetivo declarado es hallar un *modelo predictivo* parsimonioso para el diámetro (DAC). Con los criterios de información (AIC/BIC) el LMM M7 ofrece la mejor combinación de ajuste y complejidad; añadir no linealidad no mejoraría la precisión predictiva sin prolongar primero el periodo de observación.

L4. Recomendación futura. Si el seguimiento continúa y se documenta la fase de saturación del crecimiento, entonces será aconsejable reconsiderar un NLMM y evaluar su desempeño frente al LMM lineal.

En resumen, con siete mediciones distribuidas en un intervalo todavía lineal de crecimiento, un LMM con pendiente aleatoria (Modelo M7) describe adecuadamente la dinámica del DAC. Adoptar un NLMM en esta etapa habría implicado una complejidad paramétrica injustificada y riesgos de inestabilidad numérica.

5.5. Conclusiones parciales

El segundo estudio abordó el reto de modelar el crecimiento longitudinal (DAC) de *Quillaja saponaria* bajo cuatro tratamientos de bioestimulantes. Tras una preparación de los datos, la comparación sistemática de ocho Modelos Mixtos Lineales (M1–M8) permite extraer las siguientes conclusiones intermedias:

1. Elección de la respuesta y estructura temporal

- La variable *diámetro del cuello* (DAC) presentó menor heterogeneidad relativa y curvas más suaves que la altura, justificando su selección como respuesta principal.
- Siete mediciones equidistantes entre marzo y junio de 2025—iniciadas a los 7 días post-repique—proporcionaron el soporte mínimo necesario para estimar tendencias lineales e incorporar pendientes aleatorias.

2. Modelo con mejor desempeño

- El **Modelo 7** -que incluye una pendiente aleatoria y todas las covariables ambientales acumuladas— presentó el AIC más bajo ($AIC = 1051{,}67$) y una log-verosimilitud REML de $-512{,}83$.
- Añadir interacciones entre el tiempo y el tratamiento (Modelo 8) no mejoró el desempeño, y de hecho incrementó ligeramente el AIC, lo que sugiere que *la tasa media de crecimiento es aproximadamente constante entre tratamientos*, siendo más relevante capturar la heterogeneidad inter-planta mediante pendientes aleatorias específicas.

3. Importancia de los predictores

- *Tratamientos*: Los resultados indicaron que los tratamientos TB (bioestimulante comercial) y TC (Anabaena *sp.*, dosis 1) promovieron un incremento en el diámetro del cuello (DAC) en comparación con el control TA, con efectos estimados de +0,11 mm y +0,04 mm, respectivamente. En contraste, el tratamiento TD (Anabaena *sp.*, dosis 2) mostró una reducción del DAC de −0,09 mm respecto al control. Estas diferencias sugieren respuestas fisiológicas tempranas dependientes del tipo y concentración del inoculante aplicado. De forma complementaria, es relevante destacar que el equipo de investigadores ambientales con quienes se desarrolló esta colaboración ha reportado observaciones concordantes en terreno. Tras compartir los resultados del modelo, dichos expertos confirmaron que los patrones estimados estadísticamente coinciden con lo observado directamente en la dinámica de crecimiento de los plantines, otorgando respaldo empírico y biológico a los hallazgos del presente análisis.

- *Tiempo*: el coeficiente asociado al tiempo transcurrido ($\hat{\beta}_1 \approx 0{,}0005$) resultó el más pequeño entre todos los modelos evaluados y carece de significancia estadística en el modelo M7 ($p = 0{,}96$). Esto indica que, al incorporar explícitamente las covariables ambientales acumuladas (temperatura, humedad, radiación y precipitación), gran parte del efecto anteriormente atribuido al tiempo lineal queda absorbido por estas variables, las cuales capturan de forma más precisa la dinámica de crecimiento.

- *Covariables ambientales*: temperatura media acumulada mostró el efecto más pronunciado ($\beta_{temp} = -0{,}88$ mm), seguida de la radiación incidente ($\beta_{rad} = +0{,}53$ mm) y la humedad relativa ($\beta_{hum} = -0{,}13$ mm); la precipitación acumulada tuvo un efecto negativo también ($\beta_{prec} = -0{,}14$ mm).
  Estos resultados se alinean con el conocimiento ecofisiológico actual sobre los factores que regulan el crecimiento vegetal. En primer lugar, se ha documentado que temperaturas elevadas y sostenidas pueden alterar profundamente el metabolismo de las plantas, inhibiendo la fotosíntesis neta, aumentando la fotorrespiración y comprometiendo procesos clave como la síntesis de proteínas y la estabilidad de las membranas celulares (Chaves-Barrantes and Gutiérrez-Soto, 2020). Esto se traduce, en términos productivos, en un menor desarrollo estructural, lo cual es coherente con el efecto negativo estimado para la temperatura acumulada en este estudio.
  Por otro lado, la radiación solar —en particular la fracción fotosintéticamente activa (PAR)— representa la fuente energética esencial para la fotosíntesis. Diversos estudios han mostrado una relación casi lineal entre la radiación interceptada por la planta y su tasa de crecimiento, especialmente en condiciones controladas como invernaderos mediterráneos (Castilla et al., 2002). Este rol positivo de la radiación queda reflejado en el coeficiente positivo asociado a dicha variable, sugiriendo que una mayor disponibilidad lumínica ha favorecido el incremento en el diámetro del cuello.
  En contraste, niveles elevados de humedad relativa pueden reducir la transpiración y, por tanto, limitar el flujo de agua y nutrientes desde la raíz hacia la parte aérea. Además, una humedad excesiva en el entorno radicular puede conducir a condiciones hipóxicas, que afectan negativamente la respiración celular y el desarrollo radicular, disminuyendo la eficiencia de absorción de nutrientes (de Jesús Moreno Roblero et al., 2021). Esto sustenta el efecto negativo observado para esta variable.
  Finalmente, la precipitación acumulada, en exceso, puede inducir anegamientos temporales o saturación del suelo, afectando tanto la oxigenación de la rizósfera como la estructura del sustrato. Este efecto ha sido ampliamente asociado con disminuciones en el rendimiento y en el crecimiento de biomasa aérea, especialmente en especies sensibles a la hipoxia (de Jesús Moreno Roblero et al., 2021).
  En conjunto, estos hallazgos no solo son estadísticamente robustos, sino que también encuentran sustento en principios fisiológicos ampliamente documentados. El modelo ajustado logra reflejar adecuadamente la interacción entre factores ambientales clave y la dinámica de crecimiento, resaltando la necesidad de considerar variables ecofisiológicas al analizar datos longitudinales de desarrollo vegetal.

4. Componentes de varianza

- La desviación estándar del intercepto aleatorio fue $\widehat{\sigma}_{b_0} = 0{,}15$ mm y la de la pendiente $\widehat{\sigma}_{b_1} = 0{,}0014$ mm/día, reflejando diferencias iniciales y ritmos de crecimiento específicos de cada planta.
- El término residual se redujo progresivamente de 0,30 mm (M1) a 0,25 mm (M7), evidenciando la capacidad de los predictores incorporados para explicar la variabilidad intra-planta.

5. Proyecciones y trabajo futuro

- A medida que se incorporen nuevas mediciones longitudinales, se podrá evaluar con mayor solidez la validez de la suposición de linealidad en el crecimiento. En este contexto, se contempla la aplicación de *modelos no lineales mixtos* (NLMM), que permitirían capturar trayectorias sigmoideas o asintóticas más coherentes con los procesos fisiológicos del desarrollo vegetal.
- Asimismo, se considera explorar modelos *semiparamétricos* que flexibilicen la forma funcional del crecimiento sin requerir una especificación estricta, como modelos aditivos generalizados mixtos (GAMM), permitiendo capturar no linealidades suaves en el tiempo u otras covariables.
- También se proyecta el uso de *modelos con colas pesadas* (por ejemplo, con errores t-student) para mejorar la robustez ante outliers o valores extremos en la respuesta.
- Finalmente, se abre la posibilidad de incorporar *modelos conjuntos* que integren simultáneamente el proceso de crecimiento y otros desenlaces relevantes (como biomasa total o estado fitosanitario), permitiendo modelar la dependencia entre procesos longitudinales correlacionados.

En síntesis, la modelación jerárquica confirmó que *la combinación de un término de crecimiento lineal, efectos de tratamiento y covariables ambientales acumuladas, junto pendiente aleatorios por planta, describe de forma parsimoniosa y predictivamente sólida el crecimiento radial temprano de los plantines.*

Disponibilidad de código

El código completo de este Estudio II está disponible en el repositorio GitHub:

https://github.com/GabrielaGutierrez2025/tesis-microbiota-plantines/blob/main/plantines/plantines.Rmd

Capítulo 6

Conclusiones generales

Desde un punto de vista personal, esta investigación implicó un valioso desafío en la interacción con especialistas de otras áreas. Fue necesario familiarizarme con su lenguaje técnico y comprender sus enfoques experimentales, al mismo tiempo que desarrollaba la capacidad de comunicar de manera clara y accesible los conceptos y resultados propios de la estadística y la modelación matemática. Esta experiencia reforzó la importancia de la colaboración multidisciplinaria y de la comunicación efectiva entre distintos campos del conocimiento.

El presente capítulo cierra la memoria integrando los hallazgos obtenidos a lo largo de los dos estudios aplicados —microbiota vaginal (Cap. 4) y crecimiento longitudinal de plantines (Cap. 5)— con las cuestiones metodológicas desarrolladas en los capítulos previos. Se ofrece, en primer lugar, un balance de los aportes prácticos derivados del trabajo (Sección 6.1); y, posteriormente, se discuten las limitaciones residuales y las líneas de investigación que se abren a partir de los resultados alcanzados (Sección 6.2).

Este esquema persigue dos propósitos complementarios:

1. **Conectar lo metodológico con lo aplicado**. Los capítulos intermedios demostraron que los Modelos Mixtos —lineales y binomiales negativos (con o sin inflación de ceros)— constituyen un marco flexible para datos longitudinales con sobredispersión, correlación intra-sujeto y estructuras de covariables complejas. Aquí se sintetiza cómo dichos modelos resolvieron problemas reales en microbiología y silvicultura.

2. **Extraer lecciones transferibles**. Más allá de los resultados numéricos, se destaca qué decisiones analíticas (elección de la variable respuesta, esquema de aleatoriedad, criterios de selección, etc.) resultaron críticas y en qué medida pueden guiar futuros estudios con características similares.

6.1. Resumen de aportes prácticos

El trabajo realizado aporta de forma conjunta al *avance metodológico* en modelos estadísticos para datos longitudinales y a la *comprensión aplicada* de dos problemas biológicos de interés —la dinámica de la microbiota vaginal y el crecimiento de plantines de *Quillaja saponaria*. A continuación se detallan los principales logros alcanzados.

1. **Microbiota vaginal durante la gestación**. El modelo ZINBMM identificó 20 taxones cuya abundancia se asocia significativamente al estado de gestación y a su evolución temporal, confirmando el predominio y la estabilidad de *Lactobacillus spp.* en mujeres gestantes. Esta coherencia biológica refuerza la validez del modelo, ya que *Lactobacillus* es ampliamente reconocido como un género protector clave en la salud vaginal, debido a su capacidad de mantener un pH ácido, inhibir patógenos y modular la respuesta inmune local. En cambio, las mujeres no gestantes presentaron mayor variabilidad composicional, lo que coincide con el conocimiento previo sobre los cambios hormonales y la menor estabilidad microbiana fuera del contexto gestacional. Además, el enfoque inflacionado en ceros resultó crucial para capturar adecuadamente la distribución sesgada y la presencia de taxones poco frecuentes, mejorando la sensibilidad del modelo ante especies de baja abundancia pero potencial relevancia clínica.
2. **Predicción de crecimiento de plantines**. Entre ocho especificaciones LMM evaluadas, el Modelo 7 presentó el mejor ajuste según el criterio AIC y reprodujo con alta verosimilitud las trayectorias de crecimiento del diámetro del cuello (DAC). El modelo reveló que el tratamiento TB fue el que más estimuló el crecimiento en DAC respecto al control, con un efecto estimado de +0,11 mm, lo que indica una respuesta fisiológica significativa al inoculante aplicado. Sin embargo, este tratamiento también presentó el menor porcentaje de supervivencia entre los plantines, lo que sugiere posibles efectos secundarios adversos o sensibilidad fisiológica ante la formulación empleada. Esta dualidad fue corroborada por el equipo de especialistas ambientales, quienes confirmaron en terreno que, si bien los individuos del tratamiento TB tendían a crecer más, también exhibían mayores tasas de mortalidad.

   Por otro lado, el tratamiento TC (Anabaena *sp.*, dosis 1) mostró un efecto positivo más moderado sobre el crecimiento (+0,04 mm), pero se destacó por presentar el mayor porcentaje de supervivencia entre todos los grupos tratados. Este equilibrio entre estimulación del crecimiento y viabilidad sugiere que TC podría representar una alternativa bioestimulante más segura y sostenible en contextos de viverización forestal.

   A nivel ecofisiológico, los efectos estimados para las covariables ambientales resultaron coherentes con el comportamiento esperado: temperaturas acumuladas elevadas y humedad excesiva mostraron efectos negativos sobre el crecimiento, debido al estrés térmico, la fotorrespiración y la hipoxia radicular; mientras que la radiación solar incidente presentó un efecto positivo, al actuar como fuente directa de energía para la fotosíntesis. La precipitación acumulada evidenció también un efecto negativo, probablemente vinculado a procesos de anegamiento temporal o compactación del sustrato, que limitan la oxigenación de la rizósfera.

En conjunto, estos hallazgos no solo son estadísticamente consistentes, sino que han sido validados cualitativamente por el equipo de investigadores ambientales del proyecto. Tras compartir con ellos los resultados del modelo, confirmaron que los patrones estimados reflejan fielmente lo observado en terreno durante el seguimiento del ensayo, reforzando así la credibilidad ecológica y aplicabilidad práctica del enfoque adoptado.

3. **Visualización informativa**. La integración de perfiles longitudinales, mapas de calor y curvas observadas vs. predichas permitió una lectura visual inmediata de tendencias, niveles de variabilidad y efectos de tratamiento. Estas herramientas facilitaron la interpretación por parte de audiencias tanto técnicas como no especializadas, aportando valor adicional en la comunicación de los resultados y en la toma de decisiones basada en datos.

En conjunto, estos aportes no solo demuestran la utilidad de los modelos mixtos para analizar datos longitudinales con exceso de ceros y sobredispersión en contextos concretos, sino que también ofrecen criterios claros para la selección y evaluación de modelos mixtos lineales y generalizados. Además, proporcionan evidencia empírica relevante para el estudio de la salud reproductiva femenina y para la planificación de estrategias silvícolas adaptativas en contextos de restauración ecológica. La concordancia entre los hallazgos cuantitativos y los principios biológicos conocidos refuerza la aplicabilidad de los modelos propuestos en escenarios reales.

6.2. Desafíos pendientes y pasos futuros

Aumentar la densidad temporal de las mediciones en plantines. Aunque para los plantines ya se dispone de siete observaciones a lo largo del periodo de estudio, esta frecuencia aún resulta insuficiente para ajustar modelos de crecimiento no lineales—como las curvas logística, Gompertz o Richards—con efectos aleatorios estables. Para capturar adecuadamente la fase inicial de establecimiento, la inflexión de la curva y la meseta asintótica, se requieren más puntos por individuo distribuidos de manera homogénea en el tiempo.

- **Recomendación operativa.** Mantener las repeticiones cada dos semanas actual.
- **Ventaja estadística.** Una mayor granularidad temporal permitirá identificar con menor sesgo los parámetros no lineales (K, r, t_0) y estimar con precisión la variabilidad entre plantas, facilitando la eventual transición de los LMM a NLMM.
- **Impacto en la toma de decisiones.** Con modelos no lineales bien calibrados se podrá predecir la biomasa final y el momento óptimo de trasplante o cosecha, optimizando los recursos del vivero y el éxito de reforestación.

Apéndice A

Anexo: Bacterias significativamente asociadas a mujeres gestantes

Este anexo presenta las bacterias que mostraron asociación significativa con la interacción entre gestación y tiempo de gestación, según los modelos NBMM y ZINBMM aplicados en el Estudio 1 sobre microbiota vaginal. Se indican los nombres taxonómicos identificados en cada caso.

Bacterias significativas según modelos NBMM y ZINBMM

Tabla A.1: Modelo NBMM

Índice	Nombre de la bacteria
1	*Lactobacillus*
2	*Lactobacillus jensenii*
3	*Clostridiales*
4	*Leptotrichia amnionii*
5	*Finegoldia magna*
6	*Peptoniphilus*
7	*Prevotella bivia*
8	*Megasphaera sp. type 1*
9	*Gemella*
10	*BVAB1*
11	*Eggerthella*
12	*Prevotella genogroup 3*
13	*Sneathia sanguinegens*
14	*Prevotella genogroup 1*
15	*Mycoplasma hominis*
16	*Lactobacillus vaginalis*
17	*Proteobacteria*
18	*Mycoplasma*

Tabla A.2: Modelo ZINBMM

Índice	Nombre de la bacteria
1	*Lactobacillus*
2	*Lactobacillus jensenii*
3	*Clostridiales*
4	*Leptotrichia amnionii*
5	*Finegoldia magna*
6	*Prevotella bivia*
7	*Megasphaera sp. type 1*
8	*Prevotella melaninogenica*
9	*Gemella*
10	*Anaerococcus tetradius*
11	*BVAB1*
12	*Eggerthella*
13	*Prevotella genogroup 3*
14	*Sneathia sanguinegens*
15	*Prevotella genogroup 1*
16	*Mycoplasma hominis*
17	*Lactobacillus vaginalis*
18	*Proteobacteria*
19	*Actinomyces neuii*
20	*Acinetobacter calcoaceticus baumannii complex*

Apéndice B

Resultados detallados de los modelos lineales mixtos

Modelo 1: Intercepto aleatorio con efecto fijo de tratamiento

La formulación en forma mixta es

$$Y_{ij} \;=\; \beta_{\mathrm{TA}} + \beta_{\mathrm{TB}}\,\mathbf{1}_{\{\mathrm{TB}\}}(i) + \beta_{\mathrm{TC}}\,\mathbf{1}_{\{\mathrm{TC}\}}(i) + \beta_{\mathrm{TD}}\,\mathbf{1}_{\{\mathrm{TD}\}}(i) + b_i + \varepsilon_{ij},$$

$$b_i \sim \mathcal{N}(0, \sigma_b^2), \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

donde β_{TA} representa el DAC medio inicial bajo el tratamiento TA y los demás β cuantifican la diferencia respecto a TA.

Resultados de la estimación (REML).

Parámetro	Estimación	EE	t	p
β_{TA} (Intercepto)	1.6721	0.0168	99.82	$<$ 0,0001
β_{TB}	0.1274	0.0204	6.24	$<$ 0,0001
β_{TC}	0.0641	0.0202	3.18	0.0015
β_{TD}	–0.0792	0.0206	–3.84	0.0001
σ_b (SD intercepto)	0.1750			
σ (SD residual)	0.3067			

Bondad de ajuste

Modelo	AIC	BIC	ℓ_{REML}
M1: tratamiento	**2010.42**	**2046.85**	–999.21

Modelo 2: intercepto aleatorio + efecto fijo de tiempo

$$y_{ij} = \underbrace{\beta_0 + \beta_1\, t_{ij} + \beta_2\, \mathbb{1}_{\mathrm{TB},i} + \beta_3\, \mathbb{1}_{\mathrm{TC},i} + \beta_4\, \mathbb{1}_{\mathrm{TD},i}}_{\text{efectos fijos}} + \underbrace{b_{0i}}_{\substack{\text{intercepto}\\ \text{aleatorio}}} + \varepsilon_{ij}, \qquad \begin{aligned} b_{0i} &\sim \mathcal{N}(0, \sigma_b^2),\\ \varepsilon_{ij} &\sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{B.1}$$

donde t_{ij} es `día`, $\mathbb{1}_{\mathrm{TB},i}$, $\mathbb{1}_{\mathrm{TC},i}$ y $\mathbb{1}_{\mathrm{TD},i}$ son las dummies de tratamiento, y σ_b^2 y σ^2 las varianzas del intercepto aleatorio y del término de error, respectivamente.

Estimación REML

Efecto fijo	$\hat{\beta}$	EE	t	p
Intercepto (β_0)	1.4355	0.0176	81.61	$<$0,0001
`día` (β_1)	0.0048	0.0002	30.69	$<$0,0001
`TB` (β_2)	0.1241	0.0184	6.76	$<$0,0001
`TC` (β_3)	0.0554	0.0181	3.06	0.0023
`TD` (β_4)	–0.0877	0.0186	–4.72	$<$0,0001

Varianzas aleatorias

$$\widehat{\sigma}_b = 0{,}1805 \text{ mm}, \qquad \widehat{\sigma}_\varepsilon = 0{,}2663 \text{ mm}.$$

Bondad de ajuste

Modelo	AIC	BIC	ℓ_{REML}
M2: dia + tratamiento	**1211.622**	**1254.116**	–598.8108

Modelo 3: intercepto y pendiente aleatorios

$$y_{ij} = \underbrace{\beta_0 + \beta_1\, t_{ij} + \beta_2\, \mathbb{1}_{\mathrm{TB},i} + \beta_3\, \mathbb{1}_{\mathrm{TC},i} + \beta_4\, \mathbb{1}_{\mathrm{TD},i}}_{\text{efectos fijos}} + \underbrace{b_{0i} + b_{1i}\, t_{ij}}_{\substack{\text{efectos}\\ \text{aleatorios}}} + \varepsilon_{ij}, \qquad \begin{aligned} \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} &\sim \mathcal{N}\big(\mathbf{0}, \mathbf{D}\big),\\ \varepsilon_{ij} &\sim \mathcal{N}(0, \sigma^2). \end{aligned} \tag{B.2}$$

Estimación REML

Efecto fijo	$\hat{\beta}$	EE	t	p
Intercepto (β_0)	1.4426	0.0164	87.77	$<$0,0001
`día` (β_1)	0.0048	0.0002	27.39	$<$0,0001
`TB` (β_2)	0.1146	0.0181	6.34	$<$0,0001
`TC` (β_3)	0.0452	0.0179	2.53	0.0115
`TD` (β_4)	–0.0965	0.0183	–5.27	$<$0,0001

Varianzas aleatorias

$$\widehat{\sigma}_{b_0} = 0{,}1448 \text{ mm}, \quad \widehat{\sigma}_{b_1} = 0{,}0013 \text{ mm/día}, \quad \widehat{\rho}(b_0, b_1) = 0{,}312, \quad \widehat{\sigma}_{\varepsilon} = 0{,}2631 \text{ mm}.$$

Bondad de ajuste

Modelo	AIC	BIC	ℓ_{REML}
M3: dia + tratamiento; (1 + dia | planta)	**1190.219**	**1244.854**	–586.1095

Modelo 4: intercepto aleatorio + temperatura acumulada

$$y_{ij} = \underbrace{\beta_0 + \beta_1\, t_{ij} + \beta_2\, \mathbb{1}_{\text{TB},i} + \beta_3\, \mathbb{1}_{\text{TC},i} + \beta_4\, \mathbb{1}_{\text{TD},i} + \beta_5\, \mathsf{Temp}_{ij}}_{\text{efectos fijos}} + \underbrace{b_{0i}}_{\text{intercepto aleatorio}} + \varepsilon_{ij}, \tag{B.3}$$

con

$$b_{0i} \sim \mathcal{N}(0, \sigma_b^2),$$
$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

donde Temp_{ij} denota `temp_mean_cum` normalizada (aquellos grados Celsius acumulados a la fecha j de la planta i).

Estimación REML

Efecto fijo	$\hat{\beta}$	EE	t	p
Intercepto (β_0)	1.5388	0.1127	13.66	$<$0,0001
`día` (β_1)	0.0029	0.0021	1.38	0.1683
`TB` (β_2)	0.1243	0.0184	6.77	$<$0,0001
`TC` (β_3)	0.0554	0.0181	3.05	0.0023
`TD` (β_4)	–0.0877	0.0186	–4.72	$<$0,0001
Temp (β_5)	–0.0602	0.0649	–0.93	0.3536

Varianzas aleatorias

$$\widehat{\sigma}_b = 0{,}1805 \text{ mm}, \qquad \widehat{\sigma}_\varepsilon = 0{,}2663 \text{ mm}.$$

Bondad de ajuste

Modelo	AIC	BIC	ℓ_{REML}
M4: + Temp$_{ij}$	**1216.393**	**1264.955**	–600.1965

Modelo 5: intercepto y pendiente aleatorios
(`dia` + `tratamiento` + `temp_mean_cum`)

$$y_{ij} = \underbrace{\beta_0 + \beta_1\, t_{ij} + \beta_2\, \mathbb{1}_{\text{TB},i} + \beta_3\, \mathbb{1}_{\text{TC},i} + \beta_4\, \mathbb{1}_{\text{TD},i} + \beta_5\, \text{temp}_{ij}}_{\text{efectos fijos}} + \underbrace{b_{0i} + b_{1i}\, t_{ij}}_{\substack{\text{intercepto} \\ \text{\& pendiente aleatorios}}} + \varepsilon_{ij},$$

donde

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0},\, \begin{bmatrix} \sigma^2_{b0} & \sigma_{b0b1} \\ \sigma_{b0b1} & \sigma^2_{b1} \end{bmatrix} \right),$$

$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2).$$

Estimación REML

Efecto fijo	$\hat{\beta}$	EE	t	p
Intercepto (β_0)	1.5459	0.1115	13.86	$<$0,0001
`día` (β_1)	0.0029	0.0021	1.39	0.1656
`TB` (β_2)	0.1149	0.0181	6.36	$<$0,0001
`TC` (β_3)	0.0452	0.0179	2.53	0.0113
`TD` (β_4)	–0.0964	0.0183	–5.27	$<$0,0001
`temp_cum` (β_5)	–0.0603	0.0643	–0.94	0.3484

Varianzas aleatorias

$$\widehat{\sigma}_{b0} = 0{,}1457 \text{ mm}, \quad \widehat{\sigma}_{b1} = 0{,}0014 \text{ mm/día}, \quad \widehat{\rho}_{b0b1} = 0{,}287, \quad \widehat{\sigma}_{\varepsilon} = 0{,}2631 \text{ mm}.$$

Bondad de ajuste

Modelo	AIC	BIC	ℓ_{REML}
M5: M4 + (1 + dia | planta)	**1195.317**	**1256.019**	–587.6583

Modelo 6: intercepto + pendiente aleatoria y covariables ambientales acumuladas

$$y_{ij} = \underbrace{\beta_0 + \beta_1\, t_{ij} + \beta_2\, \mathbb{1}_{\text{TB},i} + \beta_3\, \mathbb{1}_{\text{TC},i} + \beta_4\, \mathbb{1}_{\text{TD},i} + \beta_5\, T^*_{ij} + \beta_6\, H^*_{ij} + \beta_7\, R^*_{ij} + \beta_8\, P^*_{ij}}_{\text{efectos fijos}} + \underbrace{b_{0i} + b_{1i}\, t_{ij}}_{\text{efectos aleatorios}} + \varepsilon_{ij}, \tag{B.4}$$

con

$$\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D}),$$
$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

donde t_{ij} = `dia`, T^*_{ij} = `temp_mean_cum`, H^*_{ij} = `humedad_mean_cum`, R^*_{ij} = `rad_mean_cum` y P^*_{ij} = `lluvia_sum_cum`.

Estimación REML

Efecto fijo	$\hat{\beta}$	EE	t	p
Intercepto (β_0)	1.6673	0.8162	2.04	0.0412
`día` (β_1)	0.0006	0.0153	0.04	0.9694
`TB` (β_2)	0.1131	0.0177	6.40	$<$0,0001
`TC` (β_3)	0.0444	0.0175	2.54	0.0111
`TD` (β_4)	–0.0973	0.0179	–5.43	$<$0,0001
`temp_cum` (β_5)	–0.8896	0.1887	–4.72	$<$0,0001
`hum_cum` (β_6)	–0.1351	0.0230	–5.87	$<$0,0001
`rad_cum` (β_7)	0.5304	0.2962	1.79	0.0735
`rain_cum` (β_8)	–0.1431	0.0174	–8.22	$<$0,0001

Varianzas aleatorias

$$\widehat{\sigma}_{b_0} = 0{,}1474 \text{ mm}, \quad \widehat{\sigma}_{b_1} = 0{,}0014 \text{ mm día}^{-1}, \quad \mathrm{Corr}(b_0, b_1) = 0{,}269, \quad \widehat{\sigma}_{\varepsilon} = 0{,}2554 \text{ mm}.$$

Bondad de ajuste

Modelo	AIC	BIC	ℓ_{REML}
M6: $t + \mathrm{trat} + T^* + H^* + R^* + P^*$	**1051.673**	**1130.574**	–512.8363

Modelo 7: pendiente aleatoria de tiempo

$$y_{ij} = \underbrace{\beta_0 + \beta_1\, t_{ij} + \beta_2\, \mathbb{1}_{\mathrm{TB},i} + \beta_3\, \mathbb{1}_{\mathrm{TC},i} + \beta_4\, \mathbb{1}_{\mathrm{TD},i} + \beta_5\, \mathrm{Temp}_{ij} + \beta_6\, \mathrm{Hum}_{ij} + \beta_7\, \mathrm{Rad}_{ij} + \beta_8\, \mathrm{Rain}_{ij}}_{\text{efectos fijos}} + \underbrace{b_{1i}\, t_{ij}}_{\substack{\text{pendiente}\\ \text{aleatoria}}} + \varepsilon_{ij}, \tag{B.5}$$

donde

$$b_{1i} \sim \mathcal{N}(0, \sigma^2_{b_1}),$$
$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2).$$

Estimación REML

Efecto fijo	$\hat{\beta}$	EE	t	p
Intercepto	1.6673	0.8162	2.04	0.0412
`día`	0.0006	0.0153	0.04	0.9694
`TB`	0.1131	0.0177	6.40	$<$0,0001
`TC`	0.0444	0.0175	2.54	0.0111
`TD`	–0.0973	0.0179	–5.43	$<$0,0001
Temp. acumulada	–0.8896	0.1887	–4.72	$<$0,0001
Humedad acumulada	–0.1351	0.0230	–5.87	$<$0,0001
Radiación acumulada	0.5304	0.2962	1.79	0.0735
Lluvia acumulada	–0.1431	0.0174	–8.22	$<$0,0001

Componentes de varianza.

$$\widehat{\sigma}_{b_0} = 0{,}1474 \text{ mm}, \qquad \widehat{\sigma}_{b_1} = 0{,}0014 \text{ mm día}^{-1}, \qquad \widehat{\rho}_{b_0 b_1} = 0{,}269, \qquad \widehat{\sigma} = 0{,}2554 \text{ mm}.$$

Bondad de ajuste

Modelo	AIC	BIC	ℓ_{REML}
M7:M6 + (dia | planta)	1051.673	1130.574	–512.8363

Modelo 8: intercepto aleatoria (`~ 1 | planta`)

$$y_{ij} = \underbrace{\beta_0 + \beta_1\, t_{ij} + \beta_2\, \mathbb{1}_{\text{TB},i} + \beta_3\, \mathbb{1}_{\text{TC},i} + \beta_4\, \mathbb{1}_{\text{TD},i} + \beta_5\, \text{temp}_{ij} + \beta_6\, \text{hum}_{ij} + \beta_7\, \text{rad}_{ij} + \beta_8\, \text{rain}_{ij}}_{\text{efectos fijos}} + \underbrace{b_{0i}}_{\text{efectos aleatorios}} + \varepsilon_{ij}, \tag{B.6}$$

donde $\mathbf{b}_i \sim \mathcal{N}_2(\mathbf{0}, \mathbf{D}), \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Estimación REML

Efecto fijo	$\hat{\beta}$	EE	t	p
Intercepto (β_0)	1.7433	0.8246	2.11	0.0346
`día` (β_1)	–0.0010	0.0154	–0.07	0.9460
`TB` (β_2)	0.0285	0.0280	1.02	0.3086
`TC` (β_3)	–0.0361	0.0278	–1.30	0.1939
`TD` (β_4)	–0.1510	0.0280	–5.38	$<$0,0001
`temp_cum` (β_5)	–0.8932	0.1906	–4.69	$<$0,0001
`hum_cum` (β_6)	–0.1361	0.0233	–5.85	$<$0,0001
`rad_cum` (β_7)	0.5213	0.2991	1.74	0.0815
`rain_cum` (β_8)	–0.1419	0.0176	–8.07	$<$0,0001
`día:TB` (β_9)	0.0019	0.0004	4.39	$<$0,0001
`día:TC` (β_{10})	0.0018	0.0004	4.26	$<$0,0001
`día:TD` (β_{11})	0.0013	0.0004	2.95	0.0033

Varianzas aleatorias

$$\widehat{\sigma}_{b_0} = 0{,}1822 \text{ mm}, \qquad \widehat{\sigma}_{\varepsilon} = 0{,}2581 \text{ mm}.$$

Bondad de ajuste

Modelo	AIC	BIC	ℓ_{REML}
M8: $M6 + \text{dia} \mid \text{planta}$ (pendiente aleatoria)	1099.721	1184.678	–535.8604

Bibliografía

Aagaard, K., Riehle, K., Ma, J., Segata, N., Mistretta, T. A., and Coarfa, C. (2012). A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PLOS ONE*, 7:e36466.

Castilla, N., Hernández, J. J., and Escobar, R. (2002). Efectos de la radiación solar en el rendimiento de los cultivos bajo invernadero. In *Los invernaderos mediterráneos: Tecnología y manejo*, pages 65–96. Ediciones Mundi-Prensa, Madrid.

Chaves-Barrantes, N. F. and Gutiérrez-Soto, M. V. (2020). Respuestas al estrés por calor en los cultivos. i. aspectos moleculares, bioquímicos y fisiológicos. *Agronomía Mesoamericana*, 31(1):261–278.

Chen, E. Z. and Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, 32(17):2611–2617.

Cho, I. and Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13(4):260–270.

Correa Morales, J. C. and Salazar Uribe, J. C. (2016). *Introducción a los Modelos Mixtos*. Universidad Nacional de Colombia, Medellín.

de Jesús Moreno Roblero, M., Godínez, L. M. B., González, G. C., and Ayala, A. M. R. (2021). El oxígeno en la zona radical y su efecto en las plantas. *Ciencia Ergo Sum*, 28(1):e740.

Dempster, A. P., Laird, N., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

DiGiulio, D. B., Callahan, B. J., McMurdie, P. J., Costello, E. K., Lyell, D. J., Robaczewska, A., Sun, C. L., Goltsman, D. S. A., Wong, R. J., Shaw, G., Stevenson, D. K., Holmes, S. P., and Relman, D. A. (2015). Temporal and spatial variation of the human microbiota during pregnancy. *Proceedings of the National Academy of Sciences*, 112(35):11060–11065.

Ghodsi, M., Liu, B., and Pop, M. (2011). DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics*, 12:271.

Gilbert, J. A., Meyer, F., and Bailey, M. J. (2011). The future of microbial metagenomics (or is ignorance bliss). *The ISME Journal*, 5(5):777–779.

Holmes, E., Li, J. V., Athanasiou, T., Ashrafian, H., and Nicholson, J. K. (2011). Understanding the role of gut microbiome–host metabolic signal disruption in health and disease. *Trends in Microbiology*, 19(7):349–359.

Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biology*, 3(2):REVIEWS0003.

Knights, D., Parfrey, L. W., Zaneveld, J., Lozupone, C., and Knight, R. (2011). Human-associated microbial signatures: examining their predictive value. *Cell Host & Microbe*, 10(4):292–296.

Kodikara, S., Ellul, S., and Lê Cao, K.-A. (2022). Statistical challenges in longitudinal microbiome data analysis. *Briefings in Bioinformatics*, 23(4):bbac273.

Lindstrom, M. and Bates, D. (1990). Nonlinear mixed effects models for repeated-measures data. *Biometrics*, 46(03):673–687.

Matsen, F. A., Kodner, R. B., and Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11:538.

Meza, C. (2021). Modelos lineales generalizados mixtos – sesión i. Lecture notes. Disponible en línea, consultado 2025-07-07.

M.J., Bates, D.M., t. . N.-R., and algorithms for linear mixed-effects models for repeated-measures data, E. (1988). *Journal of the American Statistical Association*.

Parakriti Gupta, Mini P Singh, K. G. (2020). Diversity of vaginal microbiome in pregnancy: Deciphering the obscurity. *Proceedings of the National Academy of Sciences*.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

Ritz, C. and Streibig, J. C. (2008). *Nonlinear Regression with R*. Statistics in Practice. Springer, New York.

Romero, R., Hassan, S. S., Gajer, P., and et al. (2014). The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome*, 2:4.

Samuel, B. S. and Gordon, J. I. (2006). A humanized gnotobiotic mouse model of host–archaeal–bacterial mutualism. *Proceedings of the National Academy of Sciences*, 103(26):10011–10016.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., and et al. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–1031.

Turnbull, L. A., Rees, M., and Levine, J. M. (2012). How to fit nonlinear plant growth models and calculate growth rates: an update for ecologists. *Methods in Ecology and Evolution*, 3(2):245–256.

Urbano, J., Villalobos, F. J., and otros (1999). Efecto de la radiación sobre las plantas. Apuntes académicos. Universidad de Córdoba, España.

Wooley, J. C. and Ye, Y. (2009). Metagenomics: facts and artifacts, and computational challenges. *Journal of Computer Science and Technology*, 25(1):71–81.

Zhang, X. (2024). *NBZIMM: Negative Binomial and Zero-Inflated Negative Binomial Mixed Models*. R package version 3.0.

Zhang, X., Guo, B., and Yi, N. (2020). Zero-inflated gaussian mixed models for analyzing longitudinal microbiome data. *PLOS ONE*, 15(11):e0242073.

Zhang, X., Guo, Y., and Yi, N. (2021). Fast zero-inflated negative binomial mixed models for large-scale longitudinal microbiome studies. *Bioinformatics*, 37(8):2345–2351.

Zhang, X., Mallick, H., Tang, Z., and et al. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, 18(1):4.

Zhang, X., Pei, Y.-F., Zhang, L., and et al. (2018). Negative binomial mixed models for analyzing longitudinal microbiome data. *Frontiers in Microbiology*, 9:1683.

Zhang, X. and Yi, N. (2020). Fast zero-inflated negative binomial mixed modeling approach for analyzing longitudinal metagenomics data. *Bioinformatics*, 36(8):2345–2351.

Zhang X, Pei Y-F, Z. L. e. (2018). Negative binomial mixed mod- els for analyzing longitudinal microbiome data. *Front Microbiol*.