UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE CIENCIA
DEPARTAMENTO DE MATEMÁTICAS Y CIENCIA DE LA COMPUTACIÓN

Clasificación de datos astronómicos bajo regresión logística funcional

Guillermo Alberto Grandón Rodríguez

Profesor Guía: Felipe Elorrieta López
Comisión: Andrés Iturriaga Jofre
Wilfredo Palma Manriquez

Trabajo de Titulación presentado en conformidad a los requisitos para obtener el grado de Ingeniero Estadístico

Santiago - Chile
2023

Resumen

El presente estudio se enfoca en el análisis de modelos de clasificación funcional aplicados a curvas adaptadas de estrellas variables del tipo RR Lyrae, utilizando datos del Survey Vista Variables in the Vía Láctea (VVV) en el infrarrojo cercano. El objetivo principal es cartografiar tres sectores del Bulbo Galáctico, con especial atención en la selección de estrellas del tipo ab.

El estudio abarca el pretratamiento de los datos, evaluando tres enfoques de imputación: el clásico, el promedio de la magnitud diaria en cada curva y la transformación fold de las series. Se presentan criterios de imputación mediante regresión armónica de datos no observados, incluyendo la elección de armónicos mediante validación cruzada generalizada y el uso de cuatro armónicos por recomendación. Se comparan los diferentes enfoques, realizando transformaciones de base tanto con bspline como con Fourier para el análisis funcional. Por último, se evalúan cuatro funciones de enlace distintas para identificar los modelos más adecuados para cada escenario, considerando su capacidad predictiva.

Palabras claves: RR Lyrae - datos funcionales - r-studio - series de tiempo - análisis funcional - regresión logística funcional - clasificación supervisada.

Abstract

The current study focuses on the analysis of functional classification models applied to adapted curves of variable stars of the RR Lyrae type, using data from the Vista Variables in the Vía Láctea (VVV) Survey in the near-infrared. The main objective is to map three sectors of the Galactic Bulge, with special attention to the selection of type ab stars.

The study covers data pre-processing, evaluating three imputation approaches: classical, daily magnitude averaging in each curve, and fold transformation of the series. Imputation criteria are presented using harmonic regression of unobserved data, including the choice of harmonics through generalized cross-validation and the use of four harmonics by recommendation. The different approaches are compared, performing base transformations with both bspline and Fourier for functional analysis. Finally, four different link functions are evaluated to identify the most suitable models for each scenario, considering their predictive capacity.

Keywords: RR Lyrae - functional data - R-Studio - time series - functional analysis - functional logistic regression - supervised classification.

A toda mi familia y amigos que nunca dejaron de creer en mi.

Agradecimientos

Quiero expresar mi más sincero agradecimiento a todas las personas que han sido parte de mi vida y han contribuido de manera significativa a mi crecimiento. Sin su apoyo y palabras de aliento, no habría llegado hasta aquí.

En primer lugar, quiero agradecer a mi familia, en especial a mis padres y abuelos, quienes nunca me dejaron solo en este proceso, las noches de desvelo y las conversaciones inspiradoras fueron el impulso que necesitaba para no rendirme. A mi madre Judith, cuyas frases como ”se va a tener que poder” y ”hay que tirar pa arriba nomas po” me alentaron a seguir adelante. A mi padre Raúl, que ha sido un terapeuta personal y un ejemplo a seguir en mi vida, a mis abuelos que me hicieron entender lo difícil que es la vida y mas aun la vida honrada. Los quiero mucho, son los pilares de mi vida, gracias a ustedes soy lo que soy.

A Fabian y su hermosa familia, gracias por ser un fiel compañero y hacer mi tiempo en la universidad mucho más entretenida e interesante, espero que sigamos creciendo juntos. A Fernando Carvajal, Nicol Gonzales y Tomas Taucare, tres personas maravillosas a quienes amo por su forma de ser y pensar. A Patricia, con su innato positivismo, que siempre tenía una sonrisa a pesar de las adversidades. Gracias por la confianza y las conversaciones que me ayudaron muchísimo. A Andy Gomez por ser un hermano mas, un resolvedor del cual hay que aprender.

Alexis y Alan, dos filósofos incomprendidos, gracias por sus conversaciones llenas de sabiduría que me animaron en momentos difíciles. Los valoro mucho, sus consejos y ánimo me ayudaron mucho en esta etapa académica.

Estoy profundamente agradecido por todas las personas que han compartido estos años conmigo. Cada uno de ustedes ha dejado una huella en mi vida, y este logro no habría sido posible sin su apoyo.

Es difícil encontrar las palabras adecuadas para expresar cuánto significó para mí el apoyo de cada uno de ustedes. Solo resta decir gracias ser parte de mi camino.

Tabla de contenidos

Índice de Ilustraciones

Capítulo 1

Formulación del Proyecto

1.1. Introducción

En el vasto escenario del cosmos, la astronomía utiliza su poderosa lente para desvelar los secretos de los objetos celestes. Cada estrella, con propiedades que van desde la temperatura hasta la composición química, se convierte en una página meticulosamente etiquetada en el gran libro cósmico. Estas clasificaciones actúan como claves maestras que desbloquean la complejidad y la evolución estelar, permitiendo a la humanidad comprender la narrativa que se despliega en el inmenso teatro del universo.

En este relato estelar, la fotometría se presenta como el artista que captura instantáneas de la esencia luminosa de las estrellas en el lienzo cósmico. La fotometría de apertura, como un pintor celeste, utiliza diferentes pinceles para resaltar detalles específicos en las luces parpadeantes de las estrellas. Más que una simple medición de la luz, esta técnica da vida a la personalidad única de cada estrella, permitiendo a los observadores apreciar su brillo cambiante y comprender los misterios que esta esconde.

En este lienzo, las estrellas RR Lyrae emergen como pulsantes guardianas. Su brillo, con una regularidad tan asombrosa que permite predecir sus pulsaciones y no solamente despierta fascinación, sino que también se utiliza para definir distancias galácticas, siendo la fotometría la que se convierte en la clave para descifrar lo que esconden estas estrellas variables.

A medida que avanzamos en este viaje, surge la importancia del análisis de datos funcionales. Este enfoque no solo estudia puntos aislados en las curvas de luz estelares, sino que también comprende la dinámica continua y su evolución a lo largo del tiempo. El análisis de datos funcionales se convierte así en la clave para desvelar patrones más

complejos en el comportamiento estelar, ofreciendo la flexibilidad necesaria para capturar cambios sutiles y modelar evoluciones temporales, permitiendo ahondar aún más en los misterios que se desarrollan en este escenario. Su aplicabilidad se extiende más allá de la astronomía.

En este contexto, la regresión logística funcional se presenta como una herramienta innovadora y poderosa que al aplicar este método a las curvas de luz, no solo se analizan puntos individuales (Kokoszka et al, 2017 [3]), sino que se captura la esencia dinámica de la variabilidad estelar, para desentrañar las complejidades de los patrones, permitiendo discernir entre clases con mayor precisión.

1.2. Formulación del Problema

¿Se puede mejorar la clasificación en observaciones telescópicas mediante la aplicación de modelos de regresión logística funcional?

En la astronomía, las estrellas variables, especialmente las RR Lyrae, han sido esenciales para comprender la estructura del universo. Estas estrellas, con períodos específicos, se utilizan como ”velas estándar” para estimar distancias (Minniti et. al, 2010 [2]). A pesar de su importancia, mapear la estructura del bulbo galáctico con RR Lyrae es desafiante debido a la alta extinción y la extensa área a cubrir, la aplicación de modelos estadísticos, como la regresión logística funcional, a estos datos astronómicos complejos aún presentan desafíos no resueltos (Elorrieta et. al. [4]). Este estudio se propone mejorar la predicción y detección de estrellas variables RR Lyrae para contribuir a la cartografía tridimensional del bulbo galáctico y avanzar en el análisis estadístico de datos astronómicos.

1.3. Objetivos del proyecto

1.3.1. Objetivo General

Analizar y comparar el impacto de la clasificación bajo el marco del análisis de datos secuenciales en astronomía mediante el uso de enfoques basados en datos funcionales.

1.3.2. Objetivos Específicos

- Diseñar e implementar un modelo de regresión logística funcional que permita clasificar correctamente las curvas de estrellas.
- Evaluar rendimiento del modelo y compararlo con otros enfoques de clasificación para datos funcionales.
- Evaluar y analizar la posibilidad de mejoras en la clasificación usando transformación ”folder”.

1.4. Metodología

La metodología se detalla a continuación:

- Revisión Bibliográfica de Desafíos en Datos Funcionales: Se llevará a cabo una revisión de la literatura existente sobre los desafíos asociados con el análisis de datos funcionales. Se explorarán conceptos claves relacionados con la naturaleza de este análisis y las metodologías aplicadas en investigaciones previas.
- Manipulación de los Datos: Se utilizarán los datos reales recopilados durante la proyecto astronómico ”VISTA Variables in the Via Lactea (VVV)” (consultar Minniti et al., 2010 [2]). Estos datos proporcionarán información sobre la variabilidad estelar en la Vía Láctea. Se realizará una cuidadosa manipulación, así como imputación de datos no observados con la información de expertos en el tema según Elorrieta et al (2016) [4] y procesamiento de estos datos para su posterior análisis.
- Modelación con un Enfoque Logístico Funcional: Se implementarán modelos logístico funcional (Oviedo, 2020 [5]) para comprender y caracterizar la variabilidad en los datos funcionales obtenidos. Este enfoque permitirá la configuración adecuada, facilitando la clasificación efectiva en los datos recopilados durante la proyecto astronómico VVV (ESO, 2009 [1]).

Capítulo 2

Marco Teórico

2.1. Introducción a datos funcionales

El análisis de datos funcionales es una técnica estadística que se utiliza para estudiar datos que cambian de manera continua, como por ejemplo el tiempo, el espacio u otro dominio. A diferencia del análisis de datos tradicional, que trata los datos como observaciones individuales, el análisis de datos funcionales los considera como funciones o curvas continuas (Kokoszka, 2017 [3]).

El análisis de datos funcionales se centra en estadísticas funcionales en lugar de estadísticas tradicionales y técnicas importantes es el Análisis de Componentes Principales Funcionales (FPCA), que se utiliza para descomponer la variación en los datos funcionales en componentes principales, que permite identificar patrones significativos en los datos y reducir la dimensionalidad de manera efectiva, este análisis se aplica en una amplia variedad de campos, como astronomía, economía, medicina, meteorología y más.

2.1.1. Espacio de Hilbert

Un espacio de Hilbert es crucial para abordar datos funcionales, proporcionando una estructura matemática coherente. Se inicia con la definición de un espacio vectorial, que establece operaciones como la suma y el producto por un escalar, agregando una estructura adicional al espacio vectorial.

Dentro de este contexto, las funciones cuadradas integrables (L^2) forman un espacio vectorial con propiedades clave, como la capacidad de definir un producto interno, el cual

mide la similitud y ortogonalidad entre funciones, siendo esencial para tareas como la identificación de patrones y la proyección de datos.

El espacio de Hilbert también se relaciona con espacios normados y la noción de ortonormalidad, que se utiliza para construir bases 2.1.2. La elección de una base, como las B-splines o la serie de Fourier, permite representar funciones de manera más simple y comprensible mediante una combinación lineal, esta representación base es una técnica clave para descomponer funciones en términos de funciones base, facilitando así el análisis y la interpretación de datos funcionales, para mas detalle revisar Kokoszka (2017) [3].

2.1.2. Representación Base

Es una representación de una función como una combinación lineal de funciones o elementos de una base, básicamente, se trata de expresar una función en términos de otras funciones que forman una base en un espacio vectorial. Esta técnica se utiliza para descomponer una función en componentes más simples y comprensibles. El proceso implica encontrar los coeficientes adecuados para combinar las funciones de la base de manera que se reproduzca la función original.

Suponiendo que las curvas $X_n(t)\ ;\ n = 1, \ldots, N$ son iid en L^2 se tiene la siguiente aproximación:

$$X_n(t) \approx \sum_{m=1}^{M} c_{nm} B_{nm}(t) \quad ; \quad 1 \leq n \leq N \tag{2.1}$$

donde M corresponde a la cantidad de funciones base.

- B-Spline: Es una combinación lineal de funciones base polinómicas. La flexibilidad de las B-splines proviene de la capacidad de controlar localmente la forma de la curva o superficie mediante la manipulación de los coeficientes de control (c_{nm}), permitiendo una representación precisa de la forma deseada.

- Serie de Fourier: B_{nm} estaría dada por las funciones trigonométricas seno y cosenos, los coeficientes a_{nm} y b_{nm} de amplitud que determinan la contribución de cada componente a la función $X_n(t)$. La suma se extiende a través de un número infinito de términos para representar la función de manera precisa:

$$X_n(t) \approx \sum_{m=0}^{M} [a_{nm} \cos(2\pi m t) + b_{nm} \sin(2\pi m t)]$$

(Kokoszka et al.,2017 [3] Oviedo, 2020[5])

2.1.3. Análisis Exploratorio de datos funcionales

2.1.3.1. Medidas de Tendencia y dispersión

Las medidas de tendencia central para datos funcionales son herramientas que se utilizan para resumir características de las funciones.

$$\bar{X}_n(t) = \frac{1}{N}\sum_{n=1}^{N} X_n(t) \qquad SD(X_n(t)) = \left(\frac{1}{N-1}\dot{\sum}_{n=1}^{N}(X_n(t) - \bar{X}_n(t))^2\right)^{\frac{1}{2}} \tag{2.2}$$

La media $\bar{X}_n(t)$ y la desviación estándar funcional $SD(X_n(t))$ tienen un significado análogo que en la estadística tradicional. La desviación estándar da una idea sobre la variabilidad típica de curvas en cualquier punto (Kokoszka, 2017 [3] y Oviedo, 2020[5]).
También se define la covarianza:

$$\hat{c}(t,s) = \frac{1}{N-1}\sum_{n=1}^{N}(X_n(t) - \hat{X}_n(t))(X_n(s) - \hat{X}_n(s))$$

Su interpretación es la misma que para la matriz de varianza-covarianza habitual.

2.1.3.2. Derivadas

Una vez que la observación $X_n(t)$ se expresa como una suma de bases $B_m(t)$, la función de derivada de orden k se define como:

$$X_n^{(k)}(t) \approx \sum_{m=1}^{M} c_{nm} B^{(k)}(t)$$

donde:

- se requiere que $X_n(t)$ sea k veces continuamente diferenciable.
- Las funciones bases $B_m(t)$ tengan al menos k derivadas.

En las transformadas de fourier, no existe ningún problema (Kokoszka, 2017 [3]).

2.1.3.3. Suavizado penalizado

Su objetivo principal es encontrar un equilibrio entre ajustar una función suave a los datos y evitar un ajuste excesivo al ruido (Kokoszka, 2017 [3]), lo que preserva la estructura subyacente de los datos. Esto se logra mediante la introducción de un término de penalización en la función de optimización, lo que limita la complejidad del modelo y promueve funciones suaves en lugar de adaptarse excesivamente a las fluctuaciones en los datos. Cuando las curvas muestran un ruido significativo, este enfoque permite mitigar ese ruido y puede interpretar las fluctuaciones como el resultado de fuentes externas, dependiendo del conocimiento del contexto de los datos.
Suponer que y_j en los $t_j \in [T_1, T_2]$ son denotados como $x_n(t_j, n)$, de esta forma nos centramos en una curva ruidosa cualquiera:

$$y_j = x(t_j) + \epsilon_j \quad ; \quad x(t_j) \in [T_1, T_2]$$

donde ϵ es el error aleatorio y $\mathbb{E}(\epsilon_j) = 0$, la idea principal del suavizado es eliminar la contribución del error.
Para obtener una aproximación de $x(t_j)$:

$$x(t) \approx x_k(t) = \sum_{k=1}^{K} c_k B_k(t) \quad ; \quad M < K \text{ Mucho mayor}$$

2.1.4. Operador diferencial

Se define el operador diferencial L() como una combinación lineal de las m derivadas de $x(t)$ (Kokoszka, 2017 [3]):

$$L(x(t)) = \alpha_0(t)x(t) + \alpha_1(t)x^{(1)}(t) + \cdots + \alpha_m(t)x^{(m)}(t)$$

Se quiere encontrar la curva x, equivalente a encontrar los coeficientes que minimizan la suma de cuadrados penalizada:

$$PSS\lambda(c_1, c_2, \ldots, c_k) = \sum (y_j - x_k(t_j))^2 + \lambda \int_{T_1}^{T_2} [L(x_k(t))]^2 dt$$

El término de penalización puede variar según las necesidades y contextos, siendo la penalización de Tikhonov una opción común. La elección del parámetro de penalización es fundamental, ya que regula la cantidad de suavizado aplicado.

2.1.5. Criterio de Validación

Los parámetros de penalización pueden determinarse mediante técnicas como la validación cruzada generalizada (GCV) acompañado de inspecciones, esta proporciona una cercana aproximación a la validación cruzada (CV) y es computacionalmente eficiente.

$$\text{Cross-validation: } CV(\nu) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{r}^{\nu}(x_i))^2 w(x_i)}{1 - S_{ii}}$$

Donde:

- $\hat{r}_n^{\nu}(x_i)$: hace referencia a las predicciones en el punto t_i obtenido, omitiendo el par (x_i, y_i).
- S_{ii}: el elemento i diagonal de la matriz de suavizado.
- $w(x_i)$: Peso del dato x en el tiempo t_i.
- **Nota**: Función CS.S()

El objetivo es seleccionar el valor de hiperparámetro que minimice el GCV, lo que indica el ajuste óptimo del modelo.

$$\text{Generalized Cross-validation: } GCV(\nu) = \frac{\sum_{i=1}^{n}(y_i - \hat{r}^{\nu}(x_i))^2 w(x_i)}{n(1 - tr(S)m^{-1})^2}$$

(Oviedo, 2020[5])

2.2. Regresión funcional con respuesta escalar

Estos modelos de regresión funcional permite modelar la relación entre una variable de respuesta escalar (dependiente) y una o más variables predictoras (independientes) que son funciones en lugar de valores discretos. El primer tipo de regresión funcional es una extensión de la regresión lineal tradicional que permite modelar la relación entre variables. A diferencia de la regresión lineal simple o múltiple, que trabaja con variables numéricas, la regresión funcional se enfoca en variables que son funciones. Este enfoque es especialmente útil cuando los datos pueden describirse mejor mediante curvas o tendencias funcionales en lugar de relaciones lineales simples. Algunos conceptos clave:

- **Variable Respuesta**: Es un valor escalar que puede representar, por ejemplo, una medida continua como la temperatura, la concentración de un compuesto químico, o cualquier otra variable numérica.

- **Variable Funcional**: En lugar de tener datos puntuales, se considera una variable funcional, que es una función que asigna un valor a cada punto en un dominio específico. Estas funciones pueden representar, por ejemplo, curvas temporales o perfiles de comportamiento.

- **Modelo Funcional**: El objetivo es encontrar una función que mejor se ajuste a los datos observados. Se busca un modelo funcional que describa la relación entre la variable dependiente y las variables independientes funcionales.

- **Bases Funcionales**: En la práctica, se utilizan bases funcionales para representar las funciones en términos de combinaciones lineales de funciones base.

- **Coeficientes Funcionales**: Al igual que en la regresión lineal, la regresión funcional implica encontrar los coeficientes que ponderan las funciones base para obtener la mejor aproximación a la función objetivo.

- **Problemas de Regularización**: Dado que hay muchas maneras de representar una función en términos de bases funcionales, se pueden emplear técnicas de regularización para evitar el sobre-ajuste y seleccionar las funciones base más importantes.

Es importante destacar que la regresión funcional es un campo avanzado y puede requerir conocimientos sólidos en estadísticas, análisis funcional y matemáticas aplicadas. Además, la elección de las bases funcionales y la gestión de la complejidad del modelo son aspectos cruciales para obtener resultados significativos.

Representación del modelo:

La ecuación de regresión funcional para una respuesta escalar toma la siguiente forma:

$$y_i = \langle X, \beta \rangle + \epsilon_i = \int_T X_i(t)\beta_i(t) + \epsilon(t)$$

Donde y_i es la i-esima variable respuesta, $X_i(t)$ es una función que toma valores a lo largo de algún dominio en este caso evaluado en un tiempo t y $\beta_i(t)$ esta asociados con las bases funcionales. Estos coeficientes ponderan la contribución de cada función base en la predicción del valor de la variable dependiente.

2.2.1. Regresión Lineal Funcional Generalizado

El Modelo Lineal Funcional Generalizado (GFLM) es un marco estadístico que extiende el modelo lineal tradicional para manejar datos funcionales (Oviedo, 2020[5]).
En modelo lineal tradicional generalmente se supone que $y_i|x_i$ puede ser elegido dentro del conjunto de distribuciones pertenecientes a la familia exponencial.
La función de densidad de probabilidad:

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

Donde:

- ϕ: representa un parámetro de escala o dispersión.
- θ: es el parámetro canónico de distribución.
- $a(), b()yc()$: son funciones conocidas diferentes para las distintas distribuciones de Y. (Normal, Binomial o Poisson)

Los parámetros deben ser estimados maximizando la función de verosimilitud:

$$l(\theta, \phi, y) = \log f(\theta, \phi, y) = \frac{y\theta - b(\theta)}{a(\theta)} + c(y, \theta)$$

y se especifica de la siguiente manera:

$$\mathbb{E}(y|X) = b'(\theta) = \mu \qquad \mathbb{V}ar[y|X] = b''(\theta) = V(\mu)\phi$$

μ es el valor esperado de la respuesta y $V(\mu)$ es la varianza condicional. La función de enlace es la siguiente:

$$g(\mu) = \left(\int_T X\beta + dt \right) + Z\beta$$

En R, puedes definir una clase muy flexible de funciones de enlace al especificar la distribución principal junto con la función de enlace. El enlace conecta el valor esperado de la variable de respuesta con el predictor lineal en el modelo de regresión.

Distribución	Ψ	$\mathbb{E}(\mu)$	$V(\mu)$	Enlace
Binomial	$\frac{1}{n}$	μ	$\mu(1-\mu)$	$\log \mu(1-\mu)$; logit
Poisson	1	μ	μ	$\log \mu$; log
Negative Binomial	1	$\log(\frac{\mu}{1+\frac{1}{\phi}})$	$\mu + \frac{\mu^2}{\phi}$	$\log(\mu(\phi+\mu))$; log
Normal	σ^2	μ	1	μ; identidad
Gamma	$\frac{1}{\nu}$	$\frac{-1}{\nu}$	μ^2	$\mu{-}1$; inversa

Tabla 2.1: Principales distribuciones utilizadas en GLM

2.2.2. Estimación de parámetros para GLMF

La dependencia de la respuesta escalar Y es estimada a través de covariables funcionales $X^j(t)$ y no funcionales $Z^j(t)$:

$$y_i = g^{-1}\left(\underbrace{\alpha + \beta_1 Z_{i,1} + \cdots + \beta_p Z_{i,p}}_{\text{Covariables no funcionales}} + \underbrace{\int_{T_1} \beta_1(t) X_{i,1}(t)dt + \cdots + \int_{T_1} \beta_p(t) X_{i,p}(t)dt}_{\text{Covariables funcionales}}\right) + \epsilon_i.$$

Donde la función de enlace esta dada por $g^{-1}(.)$, ϵ es el error aleatorio con media 0 y varianza σ^2 finita. La estimación tiene la siguiente forma:

$$\hat{y} = g^{-1}\left(\hat{X}\beta\right) = g^{-1}\left(\hat{X}(\hat{X}^T\hat{X})^{-1}\hat{X}\right) y = g^{-1}\left(H\right) y$$

De la forma matricial las primeras columnas de $\hat{X}$ corresponden a las covariables Z no funcionales y las siguientes a las puntuaciones (Oviedo, 2020 [5])

2.2.2.1. Expansión base:

En este caso, la matriz de covariables funcionales se basa en la representación de bases funcionales, como sigue a continuación:

$$\hat{X} = [Z_1, \ldots, Z_p, (C_1)^T\psi(t_1)\phi^T(t_1), \ldots, (C_q)^T\psi(t_q)\phi^T(t_q)]$$

2.2.2.2. Componentes principales funcionales bases:

Es posible también utilizar el siguiente enfoque basado en componentes principales funcionales para realizar la estimación de los parámetros:

$$\hat{X} = [Z_1, \ldots, Z_p, \{f^1_{i,1}, \ldots, f^1_{i,k_1}\}, \ldots, \{f^q_{i,1}, \ldots, f^q_{i,k_q}\}]$$

donde $f^j_{i,j}$ corresponde a la i-ésima puntuación asociado a la j-ésima componente principal funcional. Según [6] McCullagh y Nelder (1989) para realizar la estimación de β se pueden obtener a través del algoritmo de mínimos cuadrados iterativamente (WLS)

2.2.2.3. Devianza

La devianza a menudo se utiliza para la validación de modelos lineales generalizados, su definición general es la siguiente:

$$D = 2(l(y; y) - l(y; u))$$

donde el primer término corresponde a la función de verosimilitud del modelo completo y el segundo término a la del modelo ajustado. Cuando el parámetro es conocido, se define $D^* = \frac{D}{\psi}$ que distribuye aproximadamente χ^2_{n-p}.

2.2.2.4. Sobredispersion

En el contexto del modelado de respuestas discretas, es esencial abordar la posible sobredispersión de los datos, indicada por un parámetro de dispersión estimado que excede las expectativas ($V(y) = \psi\mathbb{E}(y)$).
La sobredispersión, donde la varianza es mayor que la media, puede conducir a subestimar los errores estándar de los parámetros y generar conclusiones erróneas. Para mitigar este problema se recomienda ajustar los errores estándar mediante el parámetro de dispersión estimado, modelar la media y la dispersión conjuntamente mediante un modelo binomial negativo. Estas estrategias abordan la sobredispersión de manera efectiva, garantizando una inferencia más precisa en presencia de datos altamente dispersos. El parámetro de dispersión usualmente es estimado después de estimar los residuos con $\hat{\beta}$ a partir de:

$$\hat{\phi} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{\mu}_i}{V(\hat{\mu}_i)}\right)$$

2.2.2.5. Validación del modelo

Para realizar la validación existen distintos tipos de residuos:

(Residuos de Pearson): $\frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}_i}}$ (Workin): $\frac{y_i - \hat{\mu}_i}{d\mu_i/d\nu_i}$

(Respuesta residual) : $y_i - \hat{\mu}_i$ (Devianza residual) $sign(y_i - \hat{\mu})\sqrt{d_i}$

Donde d_i son las contribuciones especificas de la observación a la estadística de desviación y $sign$ corresponde a la función signo (Oviedo, 2020 [5]).

2.2.3. Regresión Logística Funcional

La regresión logística funcional (RLF) modela la relación entre la respuesta binaria Y y la covariable funcional $X(t)$:

$$y_i = p_i + \epsilon_i \ ; \ i = 1, \ldots, n$$

donde ϵ_i es la componente de error aleatorio, la cual cumple los mismo supuestos que la regresión logística no funcional, es decir, son independientes con una media de 0 y una varianza de $p_i(1-p_i)$. Esta relación permite determinar la probabilidad p_i de que ocurra un evento especifico ($Y=1$) dado $X(t)$:

$$p_i = P[Y=1|x_i(t) : t \in T] = \frac{\exp\{\alpha + \int_T X_i(t)\beta(t)dt\}}{1+\exp\{\alpha + \int_T X_i(t)\beta(t)dt\}}$$

En este caso el intercepto α corresponde a una variable del tipo escalar y β al parámetro funcional asociado a la covariable funcional. A pesar que en el contexto funcional, se hace aun mas difícil la interpretación natural del modelo logístico, se busca un acercamiento a lo tradicional, por ello, es posible definir la función de enlace logit(p) como:

$$logit(p) = l_i = \ln\frac{p_i}{1-p_i} = \alpha + \int_T X_i\beta(t)dt$$

La idea principal es reducir la dimensión de las covariables a pocas bases donde:

$$\hat{l}_i = \alpha + \int_T X_i\beta(t)dt \approx \hat{X}\hat{\beta}$$

para la representación base $\hat{X}\hat{\beta} = (C_i)^T\psi(t)\phi^T(t)$ y para FPCA $\hat{X}\hat{\beta} = \{f_{i,1},\ldots,f_{i,k}\}$. La devianza para este caso corresponde a la siguiente expresión:

$$D = 2\sum_{i=1}^{n}\left[y_i log\left(\frac{y_i}{\hat{p}_{ii}}\right) + (1-y_i)log\left(\frac{1-y_i}{1-\hat{p}_{ii}}\right)\right]$$

(Oviedo, 2020 [5])

2.2.4. Análisis de Clasificación

2.2.4.1. Análisis de Sensibilidad y Especificidad

A partir del modelo ajustado, se pueden utilizar las probabilidades modeladas con la finalidad de clasificar o discriminar una observación a una determinada respuesta. La clasificación se realiza considerando un umbral, de manera que:

$$\hat{Y}_i = \begin{array}{l} 1,\ Si\ p_i > \ Umbral \\ 0,\ Si\ p_i \leq \ Umbral \end{array}$$

Donde p_i corresponde a la probabilidad modelada para la i-ésima observación a través del modelo ajustado.

A partir de esto, es posible construir una tabla de clasificación, con el fín de comparar los valores Y observados con los clasificados por el modelo $\hat{Y}$ dado un determinado umbral:

	Valores Predichos		
Valores Reales	Negativo $\hat{Y} = 0$	Positivo $\hat{Y} = 0$	Totales
Negativo $Y = 0$	Verdadero Negativo (VN)	Falso Positivo (FP)	$VN + FP$
Positivo $Y = 1$	Falso Negativo (FN)	Verdadero Positivo (VP)	$FN + VP$

Donde:

- VP: Cantidad de observaciones positivas clasificadas correctamente.
- VN: Cantidad de observaciones negativas clasificadas correctamente.
- FP: Cantidad de observaciones positivas clasificadas erróneamente.
- FN: Cantidad de observaciones negativas clasificadas erróneamente.

Una vez obtenidos los valores recopilados por la tabla, es posible definir los siguientes conceptos:

- **Sensibilidad**: Es la probabilidad de clasificar correctamente los valores positivos($Y = 1$) predichos por el modelo.

$$\mathbb{P}(\hat{Y} = 1|Y = 1) = \frac{VP}{VP + FN}$$

- **Especificidad**: Es la probabilidad de clasificar correctamente los valores negativos ($Y = 0$) predichos por el modelo.

$$\mathbb{P}(\hat{Y} = 0|Y = 0) = \frac{VN}{VN + FP}$$

- **Precisión**: Mide la proporción de instancias clasificadas como positivas que son verdaderamente positivas.

$$\text{Precisión} = \frac{VP}{VP + FP}$$

- **F1-Score**: El F1-Score es la media armónica entre la precisión y el recall. Proporciona una medida única que tiene en cuenta tanto la precisión como el recall del modelo.

$$\text{F1-Score} = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}$$

- **Tasa de Falsos Positivos** :La tasa de falsos positivos mide la proporción de instancias negativas que son incorrectamente clasificadas como positivas.

$$\text{FPR} = \frac{FP}{FP + TN}$$

- **Tasa de Verdaderos Positivos**: La tasa de verdaderos positivos es equivalente al sensibilidad. Mide la proporción de instancias positivas que son clasificadas correctamente como positivas.

$$\text{TPR} = \frac{TP}{TP + FN}$$

2.2.4.2. Curva ROC - Indicadores de discriminación

La curva ROC representa la relación entre la sensibilidad y la tasa de falsos positivos (1-especificidad). Se genera uniendo todos los posibles umbrales de corte y proporciona una medida de la capacidad del modelo para discriminar una variable dicotómica. Se considera que cuanto mayor sea el área bajo esta curva (AUC), mejor será el **poder de discriminación** del modelo (Zweig et al, 1993[7]). A modo de guía para la interpretación del indicador AUC, se tienen la siguiente categorización:

Valor AUC	Nivel de Discriminación
0.9 - 1.0	Excelente Discriminación
0.8 - 0.9	Buena Discriminación
0.7 - 0.8	Suficiente Discriminación
0.6 - 0.7	Mala Discriminación
0.5 - 0.6	Insuficiente Discriminación

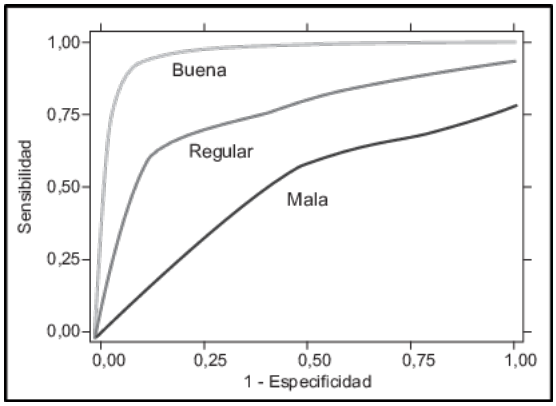


Figura 2.1: Criterios ROC-AUC

2.2.4.3. Punto de Corte Óptimo

El área bajo la curva ROC (AUC) se considera un indicador crucial de la capacidad de discriminación de un modelo predictivo. Para seleccionar un punto óptimo en la curva ROC que maximice esta capacidad de discriminación, se busca aquel que minimice la distancia entre la curva y el valor 1 en el eje de la sensibilidad. En otras palabras, se identifica el punto en la curva que esté más cercano al punto (0,1) en el espacio de sensibilidad y tasa de falsos positivos. Este enfoque se utiliza para determinar el punto de corte que mejor equilibre la sensibilidad y la especificidad del modelo, lo que resulta en una discriminación óptima entre clases en el problema de clasificación

2.2.4.4. Índice de Youden para selección de punto de corte óptimo

El Índice de Youden corresponde a un indicador que permite evaluar la capacidad de discriminación de un modelo o algoritmo de clasificación de una variable del tipo dicotómica en términos de la sensibilidad y especificidad. Este índice fue propuesto por William J. Youden en el año 1950 y es calculado de la siguiente manera:

$$J = \frac{\text{verdaderos negativos}}{\text{verdaderos negativos} + \text{falsos negativos}} + \frac{\text{verdaderos negativos}}{\text{verdaderos negativos} + \text{falsos negativos}} - 1$$

El valor de este indicador fluctúa entre -1 y 1. El valor más cercano a 1 representa el punto de corte óptimo para realizar la discriminación entre ambas clases.

2.3. Survey ”Vista Variables in the Via Lactea” (VVV)

El Proyecto VVV Survey (VISTA Variables en la Vía Láctea) representa una iniciativa colosal, parte del Núcleo Milenio para la Vía Láctea, que involucro a más de 50 astrónomos de Chile y Europa que se centra en las observaciones infrarrojas realizadas por el telescopio VISTA, un instrumento de 4,1 metros ubicado en el observatorio cerro Paranal de la ESO (Observatorio Europeo Austral), Chile 2010-2014.

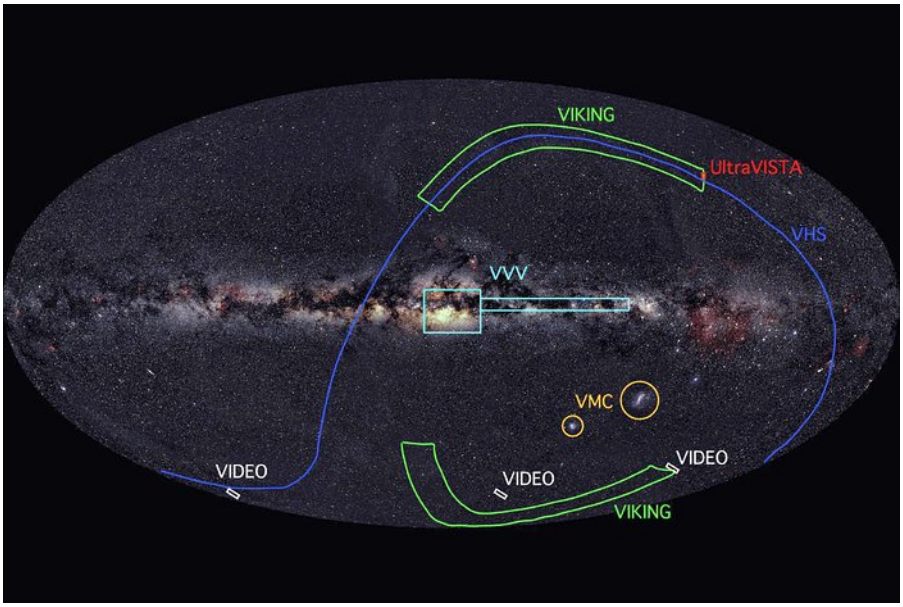


Figura 2.2: Diferentes proyectos astronómicos con el telescopio Vista de ESO [1]

Este proyecto astronómico (VVV) emerge como un faro de conocimiento al explorar aproximadamente 520 grados cuadrados del bulbo y disco galácticos. Esta investigación revela un panorama fascinante del corazón de la Vía Láctea, donde la presencia de estrellas antiguas del tipo RR Lyrae se convierte en protagonista. Este descubrimiento, respaldado por avanzadas observaciones astronómicas, se erige como un hito trascendental, brindando valiosa información sobre la estructura y evolución de nuestra galaxia a lo largo del tiempo cósmico.

En este contexto, el presente trabajo busca replicar y profundizar en los hallazgos del proyecto VVV, centrándose en la utilización de estrellas RR Lyrae como clave para mapear el Bulbo Galáctico, y explorando la aplicación de métodos automatizados para su identificación en datos funcionales.

2.3.1. Bandas ZYJHKs

Las bandas ZYJHKs son un conjunto de filtros utilizados en astronomía para observar diferentes regiones del espectro electromagnético. Cada letra representa una región específica del espectro y los filtros asociados permiten la captura de luz en esas bandas particulares.

- **Z-Banda**: Se encuentra en el extremo del espectro rojo y es utilizada para observaciones en longitudes de onda más largas. Es especialmente útil para estudiar objetos muy distantes o con corrimiento al rojo.
- **Y-Banda**: Similar a la Z-banda, se encuentra en el infrarrojo cercano y se utiliza para estudios que requieren penetración adicional a través del polvo cósmico.
- **J-Banda**: Se encuentra en el infrarrojo cercano y es sensible a longitudes de onda más largas que las bandas ópticas tradicionales. Es útil para estudiar estrellas, galaxias y otros objetos astronómicos que emiten fuertemente en estas longitudes de onda.
- **H-Banda**: Otra banda en el infrarrojo cercano, la H-banda se utiliza para observaciones en longitudes de onda aún más largas que la J-banda. Es especialmente útil para estudiar la formación estelar y otras estructuras en regiones densas de polvo.
- **Ks-Banda**: Se encuentra en el infrarrojo cercano y es sensible a longitudes de onda más largas que las bandas J y H. La Ks-banda es valiosa para estudiar objetos que emiten predominantemente en el infrarrojo.

2.3.2. Bulbo galáctico

El bulbo galáctico, una región central densa de estrellas y gas, constituye un componente esencial de muchas galaxias, incluyendo nuestra propia Vía Láctea.

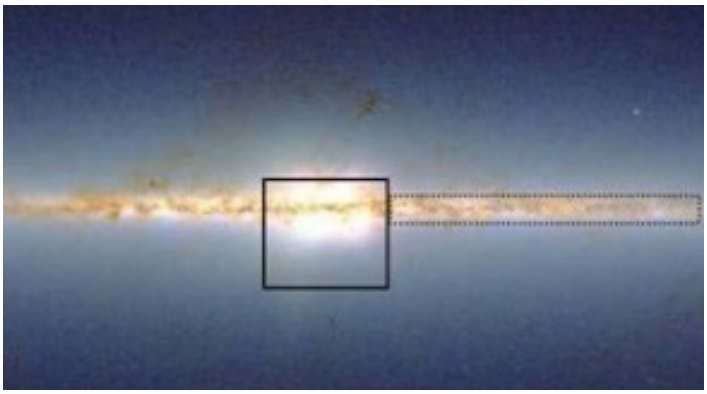


Figura 2.3: Disco y Bulbo galáctico (Imagen proporcionada por Minniti et al, 2010 [2])

En este estudio, nos centraremos en la investigación de tres sectores particulares del bulbo galáctico:

ID	RA(J2000.0)	DEC(2000.0)	Longitude	Latitude	Z	Y	J	H	Ks
b293	18 03 32.280	-28 25 56.280	2.42120	-3.13666	1	1	1	1	10
b294	18 06 45.096	-27 09 32.759	3.88150	-3.13671	1	0	1	1	10
b295	18 09 54.024	-25 52 56.640	5.34179	-3.13672	0	1	1	1	10

Donde se encuentran características como la ID del sector, RA(J2000.0) y DEC(2000.0) que es la ascensión recta y declinación de los objetos en coordenadas J2000.0, es una coordenada angular utilizada para especificar la posición de un objeto en el cielo, la longitud y latitud galáctica y por último sus longitudes de onda según los filtros utilizados.

2.3.3. Estrella variable: RR Lyrae

Utilizando datos proporcionados por el telescopio, se tiene como objetivo principal utilizar las estrellas RR Lyrae para mapear la estructura del Bulbo Galáctico. Dada la gran cantidad de fuentes esperadas, se requiere un mecanismo automatizado para identificar estas estrellas, especialmente aquellas del tipo ab.

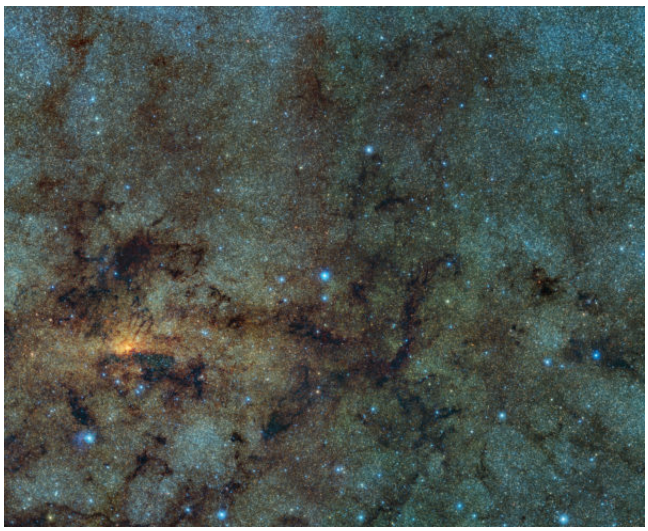


Figura 2.4: Estrellas variables cerca del centro galáctico

Las estrellas RR Lyrae, con más de 10.000 millones de años de antigüedad, destacan por sus variaciones regulares en el brillo, fenómeno asociado con pulsaciones radiales. Estas estrellas actúan como indicadores cruciales de poblaciones estelares antiguas. Su

capacidad para revelar información detallada sobre la evolución estelar y las condiciones en las que se formaron las convierte en herramientas precisas para los astrónomos, contribuyendo de manera significativa a la comprensión de la historia del universo. Las características de sus pulsaciones radiales, que son oscilaciones en el diámetro de la estrella que afectan su brillo, son 2:

- RRab : Pulsan en su modo fundamental, lo que significa que su diámetro cambia principalmente en su modo más básico. Tienen curvas de luz características con ascensos rápidos y descensos más lentos.

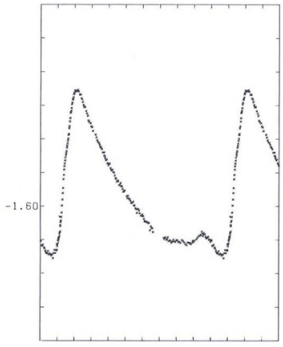


Figura 2.5: Estrella RRb (Imagen obtenida de [1])

- RRc : Pulsan en su primer sobretono, lo que implica que su diámetro cambia en una frecuencia que es aproximadamente el doble de la frecuencia fundamental. Tienen curvas de luz diferentes de las RRab, con ascensos más lentos y descensos más rápidos.

2.3.4. Datos astronómicos como datos funcionales

El fenómeno que se desea modelar es $P(Y_i = 1|X_i(d))$, donde la variable Y_i se define como una respuesta binaria que indica si, el objeto observado i, es o no una RR Lyrae, $X_i(d)$ es la representación de la magnitud del objeto, el cual puede ser representado mediante una estructura funcional $\{X_i(d), d \in D\}$, $i = 1, \ldots, N$, donde N es el número de series de tiempo (Objetos) en el día d representando el tiempo continuo perteneciente a un intervalo D especifico y esta representando los días en el sistema de números de días julianos modificados según la conversión publicada por Mayer (1998)[8].

Es relevante mencionar que el día juliano modificado (MJD) 0 indica la medianoche del 16/17 de noviembre de 1858, marcando el comienzo del sistema MJD en el día 1858-11-17

CE. Al calcular la fecha para el MJD 55690.74857, dividimos el número en dos partes: la parte entera representa el MJD completo, mientras que la parte decimal señala la fracción del día (horas).

$$MJD\ 55690 = Dia\ Juliano + 2400000{,}5 \approx Lunes\ 9\ de\ Mayo\ de\ 2011$$

$$\text{Fracción del día} = 0{,}74857 \approx 17:57:56\ (\text{hora del día})$$

Por lo tanto, el MJD 55690.74857 se traduce aproximadamente a:

$$\text{Fecha} \approx \text{Lunes } 9 \text{ de Mayo de } 2011 \text{ a las } 17:57:56\ (\text{hora del día})$$

Es esencial tener en cuenta que esta es una estimación y no considera ajustes específicos del calendario, como el cambio del calendario juliano al gregoriano. No obstante, la ventaja principal radica en la eficiencia de almacenamiento de estas fechas.

2.3.4.1. Folded light curves (Curvas de luz plegadas)

Es una herramienta importante para analizar los datos de astronómicos que muestran variabilidad periódica, se utiliza para las estrellas variables principalmente, el proceso "folding" o plegado de curvas comienza con la identificación del periodo (El periodo o frecuencia en este estudio fue otorgado por expertos), luego se procede a plegar los datos para representarlos en un periodo de referencia, esto implica que los tiempos de observación al realizarse en diferentes momentos, logrando posicionar todas las mediciones como si se hubieran tomado en el mismo punto del ciclo. Posteriormente, luego de plegar los datos se promedian las mediciones de brillo para cada fase del periodo, osea que se agrupan las mediciones en la misma fase y se calcula su valor promedio, reduciendo el ruido aleatorio de los datos. La fase ϕ de una observación se puede calcular como:

$$\phi = \left(\frac{t - t_0}{p}\right) - \mathbb{E}(t)$$

donde t_0 es el tiempo de referencia, t es el momento en que se tomó la observación, p es el período de la curva de luz y $\mathbb{E}(t)$ es la parte entera de $\frac{t-t_0}{p}$ (Elorrieta 2018 [9]). La fase generalmente se expresa como la fracción del ciclo estelar, tomando valores en el intervalo [0,1]. Para este procedimiento se utiliza la función "foldlc" de la librería iAR (Elorrieta et al, 2022 [10]) junto a la función recursiva "transformacion()" detallada en el anexo 4.4.1

Capítulo 3

Aplicación a datos astronómicos

En el trabajo de Elorrieta et al. (2016) [4], se llevó a cabo la construcción de un clasificador supervisado basado en aprendizaje automático. Este clasificador fue diseñado para asignar una puntuación a las curvas de luz del VVV en la banda Ks, con el propósito de indicar la probabilidad de que una estrella sea del tipo ab de las estrellas RR Lyrae.

En el presente estudio, el objetivo es replicar los resultados obtenidos en su trabajo mediante la implementación de un enfoque similar para datos funcionales. La replicación se centrará en la construcción de un clasificador supervisado adaptado a los datos funcionales, basado no en características de las curvas sino en las series mismas, buscando así confirmar la aplicabilidad de la metodología propuesta por Elorrieta et al. en un contexto más amplio.

3.1. Estructura de la información

Este estudio se basa en la recopilación de datos observacionales astronómicos realizada durante el proyecto astronómico Survey VVV y se organiza en tres conjuntos principales: B293, B294 y B295. Cada uno de estos conjuntos, vinculado a áreas específicas, alberga series temporales que proporcionan información detallada sobre la magnitud y el numero del día Juliano modificado observado de diversos objetos celestes. Además, se incorporan seis bases de datos adicionales para determinar la clasificación (RRLyrae) de cada estrella en sus respectivas zonas, también contienen datos suplementarios, como la frecuencia del objeto observado en las diferentes zonas.

Es importante destacar también que las observaciones astronómicas se han realizado de manera irregular. Queda en evidencia que los días de medición no son consistentes entre las diferentes series. Esta disparidad en la temporalidad de las mediciones se traduce en la

presencia de valores no observados, representados como NA (Not Available), al comparar las observaciones de objetos estelares en distintas zonas. La variabilidad en los días de medición introduce un desafío adicional al realizar análisis comparativos, la identificación y manejo adecuado de los valores no observados se convierte en un aspecto importante para garantizar la integridad de los resultados obtenidos. Este factor temporal irregular agrega complejidad al estudio, subrayando la necesidad de estrategias efectivas para abordar la variabilidad en los intervalos de observación al analizar las series temporales aplicados a datos funcionales.

Ante la irregularidad y la presencia de NA´s, se presentan tres posibles enfoques para abordar esta problemática. El primero consiste en imputar los datos faltantes utilizando regresión armónica directamente para estimar estos valores, la segunda alternativa sugiere promediar las magnitudes de cada día juliano antes de realizar la imputación, es decir, calcular el promedio de los valores observados para cada serie en un día particular, lo que podría reducir la variabilidad temporal y facilitar la imputación, aparte de disminuir considerablemente la cantidad de NA´s sin perder la estructura (ver diagrama 3.1) .

La tercera alternativa (ver el diagrama 3.2) sugiere analizar una transformación fold evaluando la posibilidad de mejora. en este caso al ser los tiempos modificados se presenta una discordancia mayor en los tiempos por lo que se realiza un promedio en las magnitudes según la aproximación de la fase ϕ en 0.01.

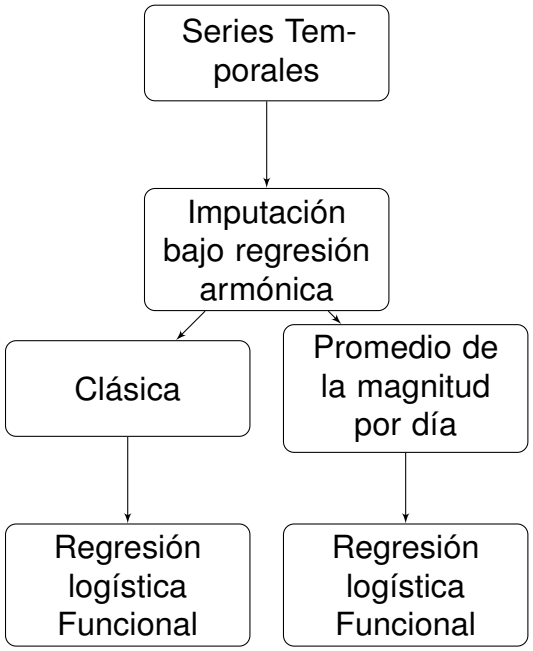


Figura 3.1: Diagrama

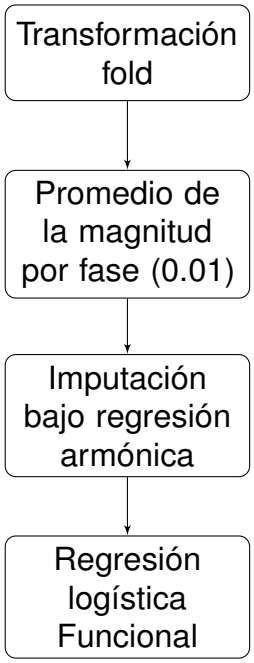


Figura 3.2: Diagrama 2

Cada una de las estrategias tiene sus propias ventajas y desafíos, y la selección del modelo óptimo deberá tener presente estos puntos que serán comentados en el desarrollo del informe.

3.2. Características del Hardware y Software

Respecto al computador personal utilizado para este proyecto, se tienen las siguientes características:

- Sistema Operativo: Windows 10 Pro 64 bits.
- Modelo del sistema: ASUS TUF-Gaming
- Procesador: AMD Ryzen 7 3800X 8-Core Processor 3.90 GHz
- Memoria RAM: 16 GB
- Tarjeta gráfica: Radeon RX 580 series

El proyecto emplea RStudio (2021) [11] , un entorno de desarrollo especializado en R, por su interfaz eficiente en escritura y depuración de código, garantiza la coherencia y reproducibilidad mediante la gestión eficaz de paquetes y dependencias.

- Librerias: fda, fda.usc, data.table, ggplot2, plotly, doParallel, pROC, dplyr, iAR, caTools.

3.3. Visualización de las series

En la fase inicial del análisis, se muestra una gráfica de 20 series temporales escogidas de manera aleatoria de cada uno de los sectores analizados (B293, B294, B295), lo que permite una representación del comportamiento estelar en las áreas específicas bajo estudio, esto permite resaltar similitudes o diferencias intra e intergrupo. El propósito es obtener una comprensión preliminar de la variabilidad, evitando evaluaciones subjetivas.

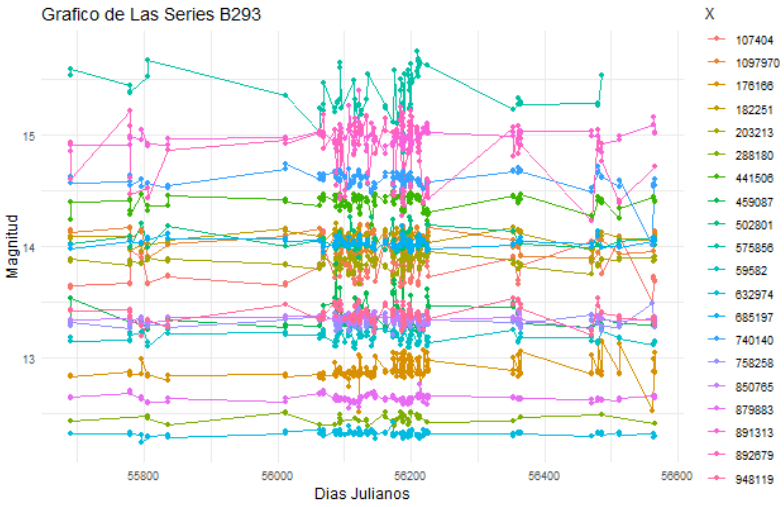


Figura 3.3: Comportamiento de las series correspondiente al grupo B293

En el anexo, se ha incorporado una sección titulada ”Detalle de las series temporales seleccionadas” 4.3.1. En dicha sección, se proporciona información sobre las series temporales elegidas para el análisis. Al examinar las tablas, se observa que los grupos B293 y B294 contienen relativamente pocas series que comparten intervalos temporales coincidentes. Esta discrepancia en los tiempos de observación es la razón detrás de la presencia de un número significativo de valores no observados (NA) en cada serie.

Es plausible que las observaciones hayan sido realizadas en el mismo día, pero en momentos distintos, generando así esta discrepancia temporal o también es posible que simplemente haya mucha diferencia de cantidad de datos en cada serie.

3.4. Visualización de las series bajo transformación fold

La función transformacion() 4.4.1 utiliza la función foldlc() de la libreria iAR, la cual recorre las curvas realizando una transformación de los tiempos en base a la frecuencia de cada curva, como lo muestra las gráficas 3.4, 4.6 y 4.7:

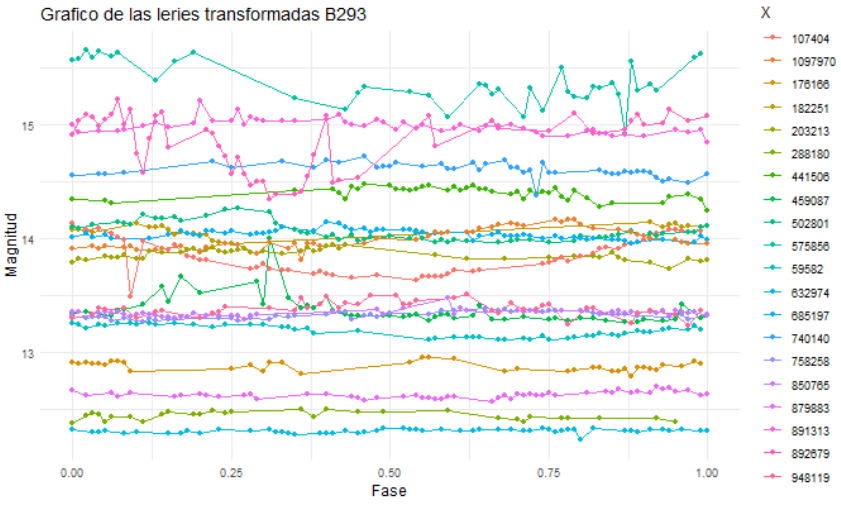


Figura 3.4: Comportamiento de la transformación correspondiente al grupo B293

Existen curvas de luz que no tienen algún patrón definido, se puede esperar que no tenga relación con la categoría buscada, ya que, las RRLyrae tienen un comportamiento periódico. En el anexo 4.4 también es posible encontrar los gráficos para los sectores B294 y B295. Las siguientes tablas muestran información sobre la cantidad de categorías por sector.

Tabla B293	
Clase	Cantidad
0	4999
1	277

Tabla 3.1: B293

Tabla B294	
Clase	Cantidad
0	5586
1	208

Tabla 3.2: B294

Tabla B295	
Clase	Cantidad
0	4153
1	180

Tabla 3.3: B295

Figura 3.5: Tablas de categorías para cada grupo

sin embargo para el análisis posterior se necesita el conteo general 3.4.

Tabla B295	
Clase	Cantidad
0	14679
1	664

Tabla 3.4: Conteo general

3.5. Imputación de datos no observados para las series

La imputación se llevara a cabo mediante la sustitución de los valores no observados en un conjunto de datos con predicciones generadas por un modelo específico. Estos modelo se entrenan exclusivamente con los valores disponibles, incorporando además la información temporal mediante la inclusión de los tiempos en los que se realizaron las observaciones. Además, se consideran las frecuencias correspondientes a cada serie, permitiendo al modelo ajustarse a patrones específicos, la imputación busca proporcionar estimaciones para los valores no observados, manteniendo la esencia de cada serie.

Ejemplo de imputación

- Al observar una serie cualquiera con datos faltantes como por ejemplo:

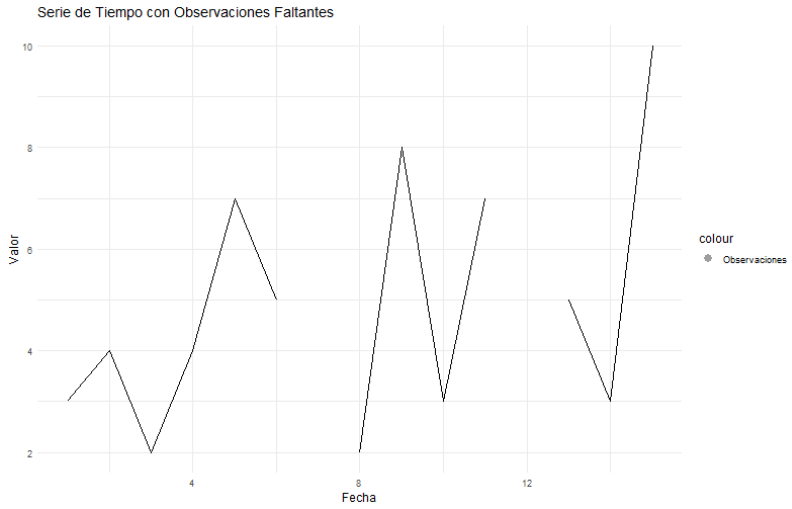


Figura 3.6: Ejemplo de imputación paso 1

- Se procede a ajustar un modelo armónico, solamente utilizando los datos observados:

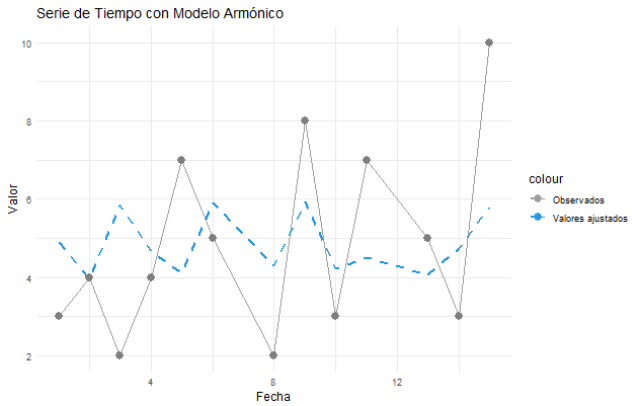


Figura 3.7: Ejemplo de imputación paso 2

- Una vez obtenidos los coeficientes del modelo armónico, se procede a realizar predicciones en los momentos en los que no se han registrado observaciones. Estas predicciones permiten sustituir los valores no observados por estimaciones ajustadas generadas a través del modelo armónico.

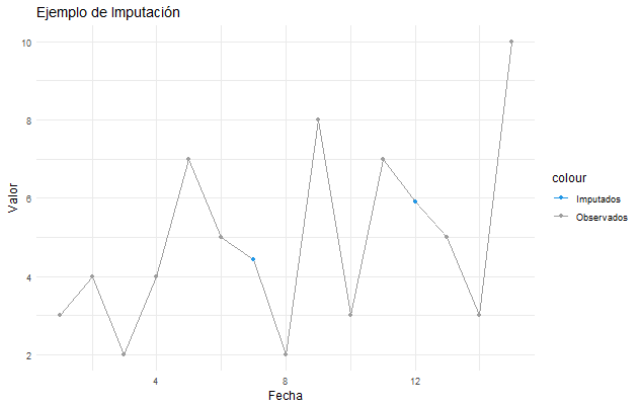


Figura 3.8: Ejemplo de imputación paso 3

3.5.1. Imputación clásica

La imputación clásica representa un desafío complejo en el contexto de esta investigación, sin embargo, se cuenta con la frecuencia de cada serie temporal. Este elemento clave facilita la imputación de los valores ausentes. En este escenario, se aprovechará la frecuencia de las observaciones para llevar a cabo el modelo armónico y posteriormente realizar la imputación.

En astronomía, la regresión armónica (ver anexo 4.3.2) ha sido una herramienta utilizada para modelar fenómenos astronómicos. La aproximación convencional suele implicar el uso de cuatro regresiones armónicas para capturar patrones oscilatorios.

Sin embargo, este enfoque convencional puede tener limitaciones, especialmente cuando se trata de la complejidad inherente de los fenómenos astronómicos. La elección de cuatro regresiones armónicas puede no ser óptima para todos los conjuntos de datos, y podría llevar a sobreajustes o subajustes.

Para abordar estas limitaciones, se propone la aplicación del criterio de validación cruzada generalizada (GCV), que proporcionaría una medida objetiva de la bondad de ajuste del modelo y permite la selección automática del número óptimo de regresiones armónicas. Este enfoque tiene la ventaja de adaptarse a la complejidad de los datos, evitando la necesidad de una elección arbitraria del número de armónicos.

Cálculo del GCV para la Elección del Mejor Modelo Armónico

La elección del modelo armónico óptimo es esencial para obtener predicciones precisas y significativas. La función recursiva ajuste_armonico descrita en el anexo 4.3.3 permite realizar ajustes armónicos y también evaluar el rendimiento de los modelos mediante el criterio de Validación Cruzada Generalizada (GCV).

El GCV se calcula para comparar 7 modelos en cada una de las curvas, cada modelo de 1 hasta 7 armónicos, permitiendo evaluar la eficacia de los modelos y seleccionar el que proporciona el mejor equilibrio entre ajuste y simplicidad.

Al analizar las tablas 4.4, 4.5 y 4.6 encontradas en el anexo 4.3.5, parece evidente que cada serie temporal requiere una configuración específica para ajustarse adecuadamente a los datos.

B293 Numero de armónicos	1	2	3	4	5	6	7
Cantidad de series	1491	1213	892	530	355	408	387

B294 Numero de armónicos	1	2	3	4	5	6	7
Cantidad de series	1737	964	950	723	484	502	434

B295 Numero de armónicos	1	2	3	4	5	6	7
Cantidad de series	1433	813	698	528	357	248	256

Tabla 3.5: Cantidad de armónicos seleccionados bajo el menor GCV - B293,B294 y B295

Se propone avanzar al siguiente paso considerando la cantidad de armónicos utilizados según el menor valor de GCV para cada serie. Este enfoque se comparará con la recomendación estándar de utilizar 4 armónicos para todas las series, como es sugerido por la convención astronómica. En los gráficos 4.3 ubicadas también en el anexo proporciona una representación visual del comportamiento del GCV para cada uno de los sectores.

Imputación basada en el menor GCV

Con el propósito de ilustrar el proceso de imputación de datos a cada una de las series, se tomará como ejemplo la serie con id 182251 perteneciente al sector B293. A continuación, se presenta la gráfica de imputación aplicada basado en el modelo óptimo según el GCV:

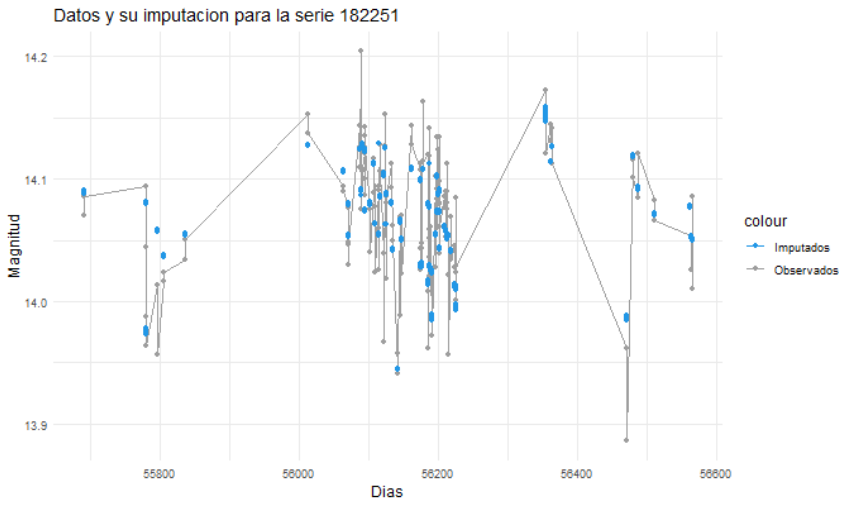


Figura 3.9: Imputación clásica basado en el menor GCV a la serie 182251

Como se puede apreciar en la figura 3.9, el color gris representa los datos observados, mientras que el azul indica la imputación en los espacios no observados. Es evidente que el proceso de imputación mantiene la coherencia en el comportamiento original de la serie.

Imputación clásica con 4 Armónicos para evaluar el desempeño

De igual forma se muestra la misma serie ejemplo 182251, pero esta vez utilizando 4 armónicos para la imputación.

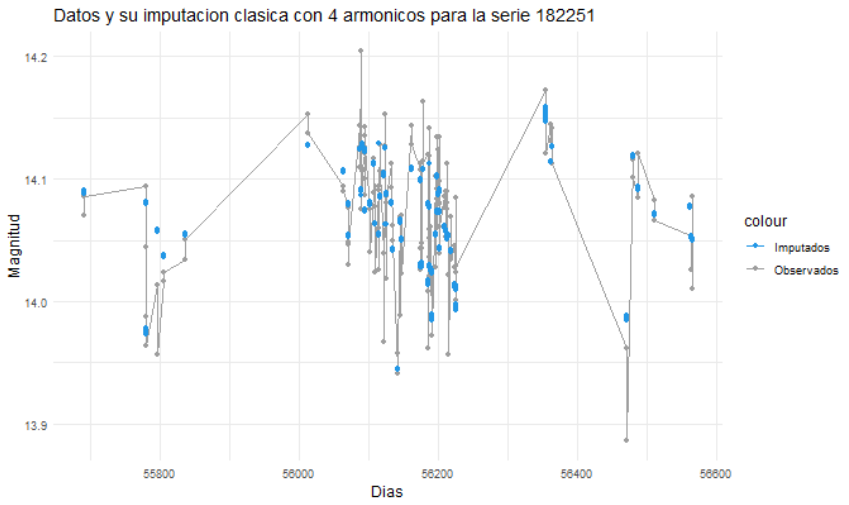


Figura 3.10: Imputación clásica basado en 4 armónicos a la serie 182251

Aunque los cambios no parezcan tener un impacto significativo (al menos no de manera visual), es posible que tengan repercusiones sutiles pero importantes posteriormente en el modelo de clasificación.

3.5.2. Imputación basada en el promedio diario

Esta imputación se implementó como una solución para abordar la problemática del elevado número de valores NA en los dos primeros grupos de datos, específicamente en B293 y B294.

El proceso de imputación se ejecutó de la siguiente manera: en primer lugar, se abordó la alta ocurrencia de NA truncando cada día de las series temporales, posteriormente, se calculó el promedio de la magnitud diaria para cada serie seleccionada.

En el anexo 4.3.8, se presentan tres tablas que reflejan el resultado de este proceso. Cada tabla representa un grupo de series temporales seleccionada.

Cabe señalar que las observaciones se llevaron a cabo durante 65 días distintos para los grupos B293 y B294, mientras que para el grupo B295 se extendió por 66 días durante los 5 años del proyecto, A continuación se presentan gráficos que comprenden el comportamiento de las series en cada uno de los sectores (B293, B294 y B295) después de haber realizado el promedio de la magnitud por día.
Se destaca que, a pesar de haber realizado el promedio diario, no se observan cambios sustanciales (al menos gráficamente) en el comportamiento general de las series

Cálculo del GCV para la Elección del Mejor Modelo Armónico

Se incluye a continuación una tabla resumen que explica la cantidad de armónicos seleccionados bajo el criterio del menor GCV para los datos promediados. Se recomienda ver detalles específicos sobre las series seleccionadas en la sección 4.3.9 ubicada en el anexo.

B293	Numero de armónicos	1	2	3	4	5	6	7
Cantidad de series		4443	567	157	75	24	8	2

B294	Numero de armónicos	1	2	3	4	5	6	7
Cantidad de series		4941	578	139	101	20	8	7

B295	Numero de armónicos	1	2	3	4	5	6	7
Cantidad de series		3406	587	221	73	29	13	4

Tabla 3.6: Cantidad de armónicos seleccionados bajo el menor GCV - promedio B293,B294 y B295

A pesar de las expectativas iniciales, el análisis de los valores de GCV derivados de las regresiones armónicas revela una tendencia inesperada en las tablas 4.10, 4.11, y 4.12. En la mayoría de las series temporales, los resultados sugieren que la elección óptima, según los criterios de GCV, es optar por una sola regresión armónica. En los casos donde el valor es mayor, la diferencia en la métrica no es significativa.

Como continuación a la sección anterior, se propone llevar a cabo un análisis comparativo entre la implementación de regresión armónica, una seleccionada según el menor valor de GCV, y la segunda de utilizar 4 regresiones armónicas.

Imputación basada en el menor GCV

Con el propósito de ilustrar el proceso de imputación de datos luego de haber promediado la magnitud por días, se tomará como ejemplo la serie con id 182251 perteneciente al sector B293. A continuación, se presenta la gráfica de imputación aplicada basado en el modelo óptimo según el GCV:

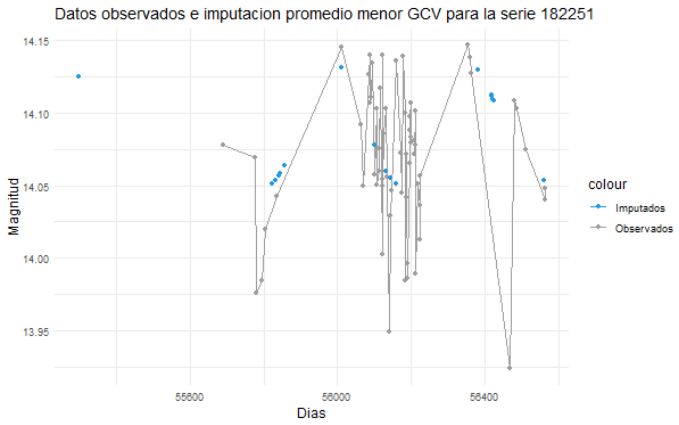


Figura 3.11: Imputación promedio basado en el menor GCV a la serie 182251

Imputación promedio con 4 Armónicos para evaluar el desempeño

De igual forma se muestra el mismo gráfico de imputación tras el promedio por día para la serie ejemplo 182251, pero esta vez utilizando 4 armónicos.

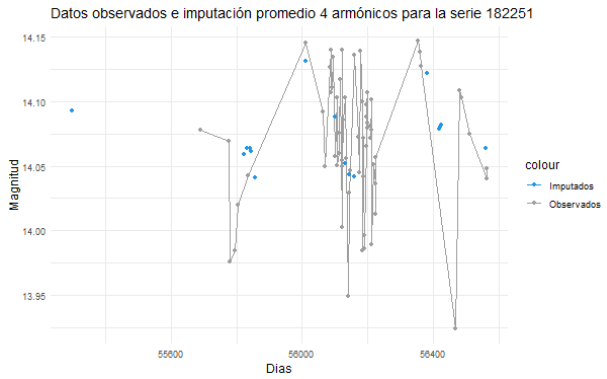


Figura 3.12: Imputación promedio basado en 4 armónicos la serie 182251

Es fácil darse cuenta que el numero de datos a imputar disminuye considerablemente, de hecho, en esta serie 182251 es un solo dato a imputar.

3.5.3. Imputación tras transformación fold

La imputación posterior a la transformación se justifica como una estrategia para abordar la sorprendente periodicidad en las variables de estrellas RRLyrae. Se busca contrastar el

enfoque de trabajar con series temporales en su forma original versus su transformación, con la expectativa de identificar mejoras significativas en la capacidad de los modelos para analizar y comprender la estructura de los datos.

Cálculo del GCV para la Elección del Mejor Modelo Armónico

El GCV se calcula para comparar 7 modelos en cada una de las curvas, cada modelo de 1 hasta 7 armónicos, permitiendo evaluar la eficacia de los modelos y seleccionar el que proporciona el mejor equilibrio entre ajuste y simplicidad.

fold	Numero de armónicos	1	2	3	4	5	6	7
Cantidad de series		8433	3199	2078	916	428	180	101

Tabla 3.7: Cantidad de armónicos seleccionados bajo el menor GCV - tras transformación

Al analizar la tabla 3.7, se hace evidente que la variación en el número de armónicos influye significativamente en la cantidad de curvas aceptadas como óptimas según el criterio de selección de modelos GCV. A medida que se incrementa el número de armónicos, se observa una tendencia clara: la cantidad de curvas aceptadas disminuye de manera consistente.

Imputación basada en el menor GCV

De la misma forma se toma el ejemplo la serie con id 182251 perteneciente al sector B293. A continuación, se presenta la gráfica de imputación aplicada basado en el modelo óptimo según el GCV:

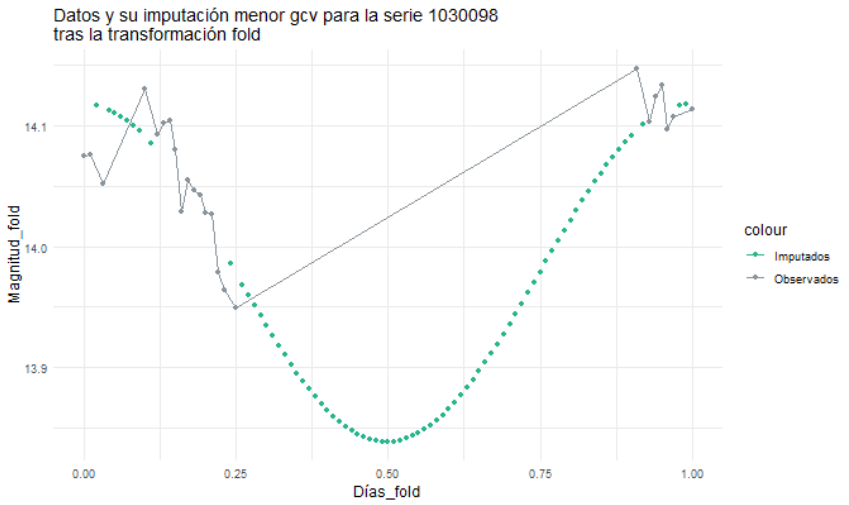


Figura 3.13: Imputación basado en el menor GCV a la serie 182251

Como se puede apreciar en la figura 3.13, el color gris representa los datos observados, mientras que el azul indica la imputación en los espacios no observados.

Imputación clásica con 4 Armónicos para evaluar el desempeño

De igual forma se muestra la misma serie ejemplo 182251, pero esta vez utilizando 4 armónicos para la imputación.

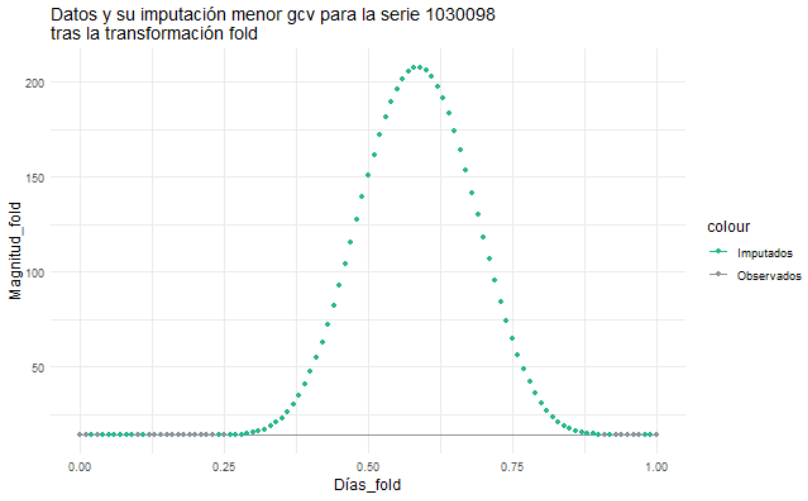


Figura 3.14: Imputación clásica basado en 4 armónicos a la serie 182251

Es evidente en 3.14 que la transformación de los datos y la omisión de la consideración de los armónicos basados en el menor criterio de validación cruzada generalizada (GCV) pueden conducir a un desajuste significativo y, en última instancia, a un mal ajuste del modelo de clasificación.

3.6. Regresión Logística funcional

Después imputados los datos, es esencial considerar las funciones que permiten la transformación de un objeto data.frame a un objeto fdata, como se detalla en el Anexo: Regresión Logística 4.5. Esta transformación permite adecuar el formato de la base de datos a los requisitos específicos de los métodos de la librería fda y fda.usc (Ramsay et. al., 2009 [12] Oviedo, 2020 [5]) .

Se procede a la partición de la base de datos en dos conjuntos distintos: uno de entrenamiento y otro de testeo. En este caso particular, se optó por seleccionar el 80 % de las curvas para el conjunto de entrenamiento y un 20 % curvas para el conjunto de testeo, con el comando sample.split(), que realiza una muestra aleatoria manteniendo la proporción respecto las categorías. Por consiguiente, se procede a crear transformación base del tipo bspline y fourier (ver anexo 4.5.3) con 4, 5, 6 y 7 bases a las curvas de entrenamiento, estas serán utilizadas para ingresarlas al modelo y entrenarlo. Posteriormente, se evaluarán criterios como el AIC (Criterio de Información de Akaike) y la devianza para la elección del mejor modelo en cada uno de los mecanismos, medir su poder de predicción, que permitirán comparar los modelos.

3.6.1. Regresión logística funcional imputación clásica - menor GCV

La primera base que se ingresa al modelo logístico funcional, es la base imputada clásica bajo el criterio de menor GCV.

El umbral de clasificación óptimo se refiere al punto en el que un modelo de clasificación determina si una observación pertenece a una clase o a otra. Para elegir este umbral, utilizamos la maximización del valor F1, que es una medida que combina precisión y sensibilidad, y es especialmente útil cuando hay un desequilibrio entre las clases en los datos.

En la Tabla 4.13, que se encuentra en el anexo, se presentan las métricas de rendimiento de cada modelo. Como los resultados de los modelos fueron muy similares, se opta por seleccionar tres modelos utilizando diferentes criterios: Devianza, AIC (Criterio de Información de Akaike) y AUC (Área bajo la curva ROC). Estos criterios nos ayudan a evaluar la

calidad del ajuste de los modelos y su capacidad predictiva.

El primer modelo consta de 4 bases bspline con una función de enlace probit y alcanza un AUC de 0.6527. Por otro lado, los otros dos modelos están compuestos por 7 bases bspline uno con una función de enlace probit con AUC de 0.654 y el otro con una función de enlace cloglog y un AUC de 0.6568. Estos resultados sugieren que la capacidad predictiva de los tres modelos es regular o moderada.

Después de seleccionar los tres modelos, examinamos sus métricas de rendimiento en un conjunto de testeo para evaluar cómo se desempeñan en la práctica.

Regresión logística funcional-Testeo						
Base	Numero	Enlace	Sensibilidad	Precision	especificidad	AUC
Bspl	4	Probit	0.8727	0.4398	0.416	0.6444
Bspl	7	Probit	0.8727	0.4369	0.413	0.6429
Bspl	7	Cloglog	0.8909	0.4312	0.406	0.6485

Tabla 3.8: Métricas de la regresión logística funcional base de testeo

La tabla 3.8 detalla las métricas que ayuda a comprender el rendimiento de los modelos en la base de testeo, de los tres modelos seleccionados durante el proceso de entrenamiento, siendo el último modelo el que exhibe un AUC mayor, por lo que seria la opción preferida entre los modelos en términos de su capacidad para clasificar correctamente en la base de prueba.

3.6.2. Regresión logística funcional imputación clásica - 4 armónicos

La segunda base ingresada al modelo logístico funcional, es la base imputación clásica bajo el criterio recomendado de 4 armónicos.

En la Tabla 4.5.5, que se encuentra en el anexo, se presentan las métricas de rendimiento de cada modelo al igual que la sección anterior. Estos criterios recomienda seleccionar 3 modelos, es importante destacar que los modelos seleccionados están asociados a una función de enlace probit.

El primer modelo consta de 4 bases bspline alcanzando un AUC de 0.647, el segundo modelo consta de 5 bases bspline, con un AUC de 0.6561, el 3 esta formados por 6 bases de fourier, con un AUC correspondiente a 0.6027.

Después de seleccionar los modelos, examinamos sus métricas de rendimiento en un conjunto de testeo para evaluar cómo se desempeñan en la práctica.

Regresión logística funcional - Testeo						
Base	Numero	Enlace	Sensibilidad	Precision	especificidad	AUC
Bspl	4	Probit	0.8727	0.4379	0.414	0.6434
Bspl	5	Probit	0.8727	0.4379	0.414	0.6434
Fourier	6	Probit	0.8181	0.453	0.433	0.6256

Tabla 3.9: Métricas de la regresión logística funcional base de testeo

La tabla 3.9 detalla las métricas que ayuda a comprender el rendimiento de los modelos en la base de testeo, de los tres modelos seleccionados durante el proceso de entrenamiento, siendo el modelo logístico creado con 4 bspline por parsimonia el que exhibe un AUC mayor, por lo que seria la opción preferida entre los modelos en términos de su capacidad para clasificar correctamente en la base de prueba.

3.6.3. Regresión logística funcional imputación Promedio - Menor GCV

La tercera base que se ingresa al modelo logístico funcional, es la base imputación promedio bajo el criterio de menor GCV.

En la Tabla 4.5.6, que se encuentra en el anexo, se presentan las métricas de rendimiento de cada modelo al igual que las secciones anteriores. Estos criterios recomienda seleccionar 2 modelos, es importante destacar que los 2 modelos seleccionados están asociados a bases de fourier y a la función de enlace probit.

El primer modelo consta de 4 bases respectivamente, alcanza un AUC de 0.5938 y el modelo consta de 6 bases, con un AUC de 0.5868. Estos resultados sugieren que la capacidad predictiva de los modelos es moderada o regular.

Después de seleccionar los modelos, examinamos sus métricas de rendimiento en un conjunto de testeo para evaluar cómo se desempeñan en la práctica.

Regresión logística funcional - Testeo						
Base	Numero	Enlace	Sensibilidad	Precision	especificidad	AUC
Fourier	4	Probit	0.856	0.3762	0.3546	0.6054
Fourier	6	Probit	0.7575	0.4766	0.464	0.6108

Tabla 3.10: Métricas de la regresión logística funcional base de testeo

La tabla 3.10 detalla las métricas que ayuda a comprender el rendimiento de los modelos en la base de testeo, de los modelos seleccionados durante el proceso de entrenamiento, siendo el modelo logístico creado con 6 bases de fourier el que exhibe un AUC mayor, por lo que seria la opción preferida entre los modelos en términos de su capacidad para clasificar correctamente en la base de prueba.

3.6.4. Regresión logística funcional imputación Promedio - 4 armónicos

La cuarta base que se ingresa al modelo logístico funcional, es la base imputación promedio bajo el criterio recomendado de 4 armónicos.

En la Tabla 4.5.7, que se encuentra en el anexo, se presentan las métricas de rendimiento de cada modelo al igual que las secciones anteriores. Estos criterios recomienda seleccionar 2 modelos, es importante destacar que los modelos seleccionados están asociados a bases de fourier y a la función de enlace probit.

El primer modelo consta de 4 bases, alcanza un AUC de 0.5938 y el segundo modelo consta de 6 bases, con un AUC de 0.5868. Estos resultados sugieren que la capacidad predictiva de los modelos es moderada o regular.

Después de seleccionar los modelos, examinamos sus métricas de rendimiento en un conjunto de testeo para evaluar cómo se desempeñan en la práctica.

Regresión logística funcional - Testeo						
Base	Numero	Enlace	Sensibilidad	Precision	especificidad	AUC
Fourier	4	Probit	0.9924	0.2539	0.2207	0.6066
Fourier	6	Probit	0.9848	0.2657	0.2333	0.6091

Tabla 3.11: Métricas de la regresión logística funcional base de testeo

La tabla 3.11 detalla las métricas que ayuda a comprender el rendimiento de los modelos en la base de testeo, de los modelos seleccionados durante el proceso de entrenamiento, siendo el modelo logístico creado con 6 bases de fourier el que exhibe un AUC mayor, por lo que seria la opción preferida entre los modelos en términos de su capacidad para clasificar correctamente en la base de prueba.

3.6.5. Regresión logística funcional tras transformación Fold - menor GCV

La quinta base ingresada al modelo logístico funcional, es la base tras la transformación bajo el criterio de menor GCV.

En la Tabla 4.5.8, que se encuentra en el anexo, se presentan las métricas de rendimiento de cada modelo al igual que las secciones anteriores. Estos criterios recomienda seleccionar 4 modelos, es importante destacar que los 4 modelos seleccionados están asociados a una función de enlace probit.

El primer modelo consta de 4 bases bspline alcanzando un AUC de 0.5032, el segundo modelo consta de 7 bases bspline, con un AUC de 0.5681, el tercer modelo consta de 4 bases de fourier, con un AUC de 0.5509 y el ultimo modelo consta con 7 bases de fourier con un AUC de 0.677. Estos resultados sugieren que la capacidad predictiva de los modelos es moderadamente mala, debido a que la clasificación es casi aleatoria.

Después de seleccionar los modelos, examinamos sus métricas de rendimiento (solo a modo de estudio) en un conjunto de testeo para evaluar cómo se desempeñan en la práctica.

Regresión logística funcional						
Base	Numero	Enlace	Sensibilidad	Precision	especificidad	AUC
Bspl	4	Probit	0.9398	0.3826	0.3575	0.6486
Bspl	7	Probit	0.4661	0.4985	0.5	0.5169
Fourier	4	Probit	0.9924	0.2557	0.2225	0.6075
Fourier	7	Probit	0.0075	0.875	0.9141	0.4608

Tabla 3.12: Métricas de la regresión logística funcional base de testeo

La tabla 3.12 detalla las métricas que ayuda a comprender el rendimiento de los modelos en la base de testeo, de los modelos seleccionados durante el proceso de entrenamiento,

siendo el modelo logístico creado con 4 bases bspline el que exhibe un AUC mayor, por lo que seria la opción preferida entre los modelos en términos de su capacidad para clasificar correctamente en la base de prueba, hay recalcar que el ultimo modelo tiene un auc de 0.4608, quiere decir que el modelo entrega deplorables clasificaciones a pesar de haber entregado resultados decentes en la base de entrenamiento.

3.6.6. Regresión logística funcional tras transformación Fold - 4 armónicos

La sexta y ultima base ingresada al modelo logístico funcional, es la base tras la transformación bajo el criterio recomendado de 4 armónicos.

En la Tabla 4.5.9, que se encuentra en el anexo, se presentan las métricas de rendimiento de cada modelo al igual que las secciones anteriores. Estos criterios recomienda seleccionar 2 modelos, es importante destacar que los 2 modelos seleccionados están asociados a una función de enlace probit.

El primer modelo consta de 4 bases bspline alcanzando un AUC de 0.52, el segundo modelo consta de 7 bases bspline, con un AUC de 0.5709. Estos resultados sugieren que la capacidad predictiva de los modelos es moderadamente mala, debido a que la clasificación es casi aleatoria.

Después de seleccionar los modelos, examinamos sus métricas de rendimiento (solo a modo de estudio) en un conjunto de testeo para evaluar cómo se desempeñan en la práctica.

Regresión logística funcional						
Base	Numero	Enlace	Sensibilidad	Precision	especificidad	AUC
Bspl	4	Probit	0.9924	0.0827	0.0417	0.5171
Bspl	7	Probit	0.9924	0.2148	0.1797	0.58613

Tabla 3.13: Métricas de la regresión logística funcional base de testeo

La tabla 3.13 detalla las métricas que ayuda a comprender el rendimiento de los modelos en la base de testeo, de los modelos seleccionados durante el proceso de entrenamiento, siendo el modelo logístico creado con 4 bases bspline el que exhibe un AUC mayor, por lo que seria la opción preferida entre los modelos en términos de su capacidad para clasificar correctamente en la base de prueba.

3.7. Comparación global de los modelos

Realizados ya la selección de los modelos, en la tabla 3.14 se muestran los modelos mas significativos para cada uno de los caminos de imputación realizado.

Base de datos	Representación	Función de enlace	AUC fit	AUC pred	Punto de corte
Clásica - Menor GCV	7 Bspline	cloglog	0.6568	0.6485	0.05
Clásica - 4 Armónicos	5 Bspline	probit	0.6561	0.6434	0.05
Promedio - Menor GCV	6 Fourier	probit	0.5868	0.6108	0.04
Promedio - 4 Armónicos	6 Fourier	probit	0.5938	0.6091	0.04
Fold - Menor GCV	4 Bspline	probit	0.5032	0.6486	0.04
Fold - 4 Armónicos	7 Bspline	probit	0.5709	0.58613	0.04

Tabla 3.14: Resumen de los modelos

A partir de la tabla anterior, es evidente que, a pesar de que el AUC más alto en la base de testeo es de 0.6486 debido a la transformación fold y la selección de armónicos para cada curva bajo el criterio de menor GCV, este resultado es inusual y se aparta del patrón observado en el resto de las curvas. En general, después de la transformación, las métricas AUC son considerablemente más bajas, con un promedio inferior a 0.58. Esto sugiere que el resultado obtenido puede ser atípico y posiblemente no replicable. De hecho, en la base de entrenamiento, la capacidad predictiva de este modelo es prácticamente aleatoria.

En un principio, se suponía que la transformación de los datos conduciría a una mejora en su tratamiento, una adaptación más eficiente al modelo y, por ende, una mejor capacidad de clasificación. Sin embargo, esto no se refleja en los modelos presentados.

Capítulo 4

Trabajo Futuro y Conclusiones

4.1. Trabajo Futuro y Conclusiones

Se ha realizado un análisis de los modelos de clasificación aplicados a curvas transformadas bajo diferentes técnicas de imputación, incluyendo imputación clásica, imputación diaria y el proceso de folding. Este análisis se llevó a cabo utilizando tanto el criterio de validación cruzada como la recomendación con 4 armónicos. Cada conjunto de datos imputado fue evaluado mediante dos técnicas de representación: Fourier y Bspline, combinadas con cuatro funciones de enlace distintas: Logit, Probit, Cauchit y Cloglog.

El propósito de esta comparación fue identificar los modelos y técnicas de transformación más apropiados para cada conjunto de datos, basándose en métricas clave como el AUC (Área bajo la curva) para los datos ajustados y las predicciones, así como el punto de corte utilizado para la clasificación.

Los resultados mostraron que, en la base de datos bajo imputación clásica con Menor GCV, el modelo con la representación de 7 bspline y la función de enlace Cloglog exhibió un rendimiento consistente, con un AUC de entrenamiento y predicción de 0.6568 y 0.6485 respectivamente, con un punto de corte de 0.05. En la base de datos bajo imputación clásica con la recomendación de 4 armónicos, el modelo con representación de 5 bspline y la función de enlace Probit también mostró un rendimiento consistente, aunque generalmente inferior al modelo bajo el criterio GCV, con un AUC de entrenamiento y predicción de 0.6561 y 0.6434 respectivamente, con un punto de corte de 0.05.

Por otro lado, al analizar el resto de las bases, se encontró que los modelos con la función de enlace Probit tendieron a superar en rendimiento a las otras funciones de enlace en términos de AUC de entrenamiento y predicción. Este resultado resalta la relevancia de la

elección de la función de enlace en la efectividad general del modelo.

Al analizar la influencia del tipo de imputación, se observó un aumento en el rendimiento en términos de AUC tanto en el entrenamiento como en la predicción, independientemente del criterio utilizado. Este hallazgo es significativo ya que, independientemente de la transformación aplicada, los modelos construidos sobre las bases promediadas o transformadas muestran un desempeño inferior en comparación con los modelos aplicados a la base de imputación clásica. Este resultado sugiere que para realizar cualquier tipo de transformación, es necesario considerar más características además de la frecuencia, con el fin de mejorar su rendimiento. Por otro lado, la tendencia a omitir información valiosa al realizar promedios en las magnitudes diarias también es una consideración importante. Además, es posible que al intentar explicar estos valores mediante la utilización de 4 a 7 bases, que al final se traducen en polinomios o funciones sinusoidales, se esté incurriendo en un sobreajuste al modelo. Este sobreajuste podría estar perjudicando el rendimiento general de los modelos.

En el estudio realizado por Elorrieta et al. (2016) [4], se presentaron resultados de AUC en su versión no funcional que superan significativamente los obtenidos en este análisis. Es importante destacar que estos resultados fueron obtenidos utilizando el criterio de GCV. Respondiendo a la pregunta "¿Se puede mejorar la clasificación en observaciones telescópicas mediante la aplicación de modelos de regresión logística funcional?", como se analizó en el informe, no se encontró ninguna mejora. Los modelos funcionales resultaron en malas clasificaciones. Sin embargo, aunque la evidencia sugiere lo contrario, uno podría considerar en una perspectiva aspirante que existe la posibilidad de mejorar.

Aunque los objetivos del proyecto fueron cumplidos, no como se esperaba, no se pierde la ilusión de que la regresión logística funcional podría ser una buena opción para la clasificación estelar. Quizás este enfoque se ve desfavorecido por la irregularidad de las mediciones. A pesar de los avances logrados en esta etapa de la investigación, se vislumbran diversas líneas de trabajo futuro que fortalecerían el propósito de este proyecto. Algunas de estas direcciones incluyen:

- Mejorar la programación de los modelos de regresión logística funcional para aumentar su rendimiento.
- Utilización de otros modelos de clasificación en su versión funcional.
- Utilización de otras tácticas de representación.
- Realiza una búsqueda sistemática de los hiperparámetros del modelo para encontrar la combinación óptima que maximice el rendimiento.

- Considerar métodos avanzados de optimización para mejorar la eficiencia del modelo en términos de clasificaciones. La aplicación de técnicas de optimización contribuirá a la fineza del modelo y, potencialmente, a una mejora sustancial en su rendimiento predictivo.

Estas sugerencias y líneas de trabajo futuro no solo consolidarán los cimientos establecidos hasta ahora, sino que también contribuirán significativamente al avance continuo de la investigación y al enriquecimiento del conocimiento en el campo de análisis de series temporales.

Referencias Bibliográficas

[1] Observatorio Europeo Austral. Eso, 2009.

[2] D. Minniti, P. W. Lucas, J. P. Emerson, R. K. Saito, M. Hempel, P. Pietrukowicz, A. V. Ahumada, M. V. Alonso, J. Alonso-García, J. I. Arias, R. M. Bandyopadhyay, R. H. Barbá, B. Barbuy, L. R. Bedin, E. Bica, J. Borissova, L. Bronfman, G. Carraro, M. Catelan, J. J. Clariá, N. Cross, R. de Grijs, I. Dékány, J. E. Drew, C. Fariña, C. Feinstein, E. Fernández Lajús, R. C. Gamen, D. Geisler, W. Gieren, B. Goldman, O. A. Gonzalez, G. Gunthardt, S. Gurovich, N. C. Hambly, M. J. Irwin, V. D. Ivanov, A. Jordán, E. Kerins, K. Kinemuchi, R. Kurtev, M. López-Corredoira, T. Maccarone, N. Masetti, D. Merlo, M. Messineo, I. F. Mirabel, L. Monaco, L. Morelli, N. Padilla, T. Palma, M. C. Parisi, G. Pignata, M. Rejkuba, A. Roman-Lopes, S. E. Sale, M. R. Schreiber, A. C. Schröder, M. Smith, L. Sodré Jr., M. Soto, M. Tamura, C. Tappert, M. A. Thompson, I. Toledo, and M. Zoccali. Vista variables in the via lactea (vvv): The public eso near-ir variability survey of the milky way. *Astronomy & Astrophysics*, 527:A81, 2010.

[3] Piotr Kokoszka and Matthew Reimherr. *Introduction to Functional Data Analysis*. CRC press, 2017.

[4] Felipe Elorrieta, Susana Eyheramendy, Andrés Jordán, István Dékány, Márcio Catelan, Rodolfo Angeloni, Javier Alonso-García, Rodrigo Contreras-Ramos, Felipe Gran, Gergely Hajdu, Néstor Espinoza, Roberto K. Saito, and Dante Minniti. A machine learned classifier for rr lyrae in the vvv survey. *Astronomy & Astrophysics*, 2016.

[5] Manuel Oviedo de la Fuente. *Utilies for Statistical Computing in Functional Data Analysis: The R Packahe fda.usc*. PhD thesis, Universidad de Santiago de Compostela, 2020.

[6] P. McCullagh and J. A. Nelder. Generalized linear models. *Chapman and Hall*, 1989.

[7] Michael H Zweig and Graham Campbell. Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577, 1993.

[8] Peter Meyer. *Julian Day Number (JDN) Calculation*, 1998.

[9] Felipe Elorrieta López. *Classification and Modeling of time series of astronomical data*. PhD thesis, Pontificia Universidad Católica de Chile, 2018.

[10] Felipe Elorrieta, Cesar Ojeda, Susana Eyheramendy, and Wilfredo Palma. *iAR: Irregularly Observed Autoregressive Models*, 2022. Versión: 1.2.0. Publicado: 2022-11-24. URL: https://CRAN.R-project.org/package=iAR.

[11] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA, 2021.

[12] James O. Ramsay, Giles Hooker, and Spencer Graves. *fda: Functional Data Analysis*. R package version 6.1.4, 2009.

Anexo

4.2. Anexo del capitulo 2

4.2.1. Análisis de componentes principales funcionales

En el análisis de datos funcionales, el análisis de Componentes Principales Funcionales (FPCA, por sus siglas en inglés) es una técnica que se utiliza para reducir la dimensionalidad de los datos y extraer información relevante de funciones continuas. FPCA es una extensión del análisis de componentes principales (PCA).

La estimación de las FPCA esta relacionada con la función de covarianza muestral. La idea es encontrar la función $\hat{v}_j$ tal que:

$$X_n(t) - \bar{X}_n(t) \approx \sum_{j=1}^{p} \hat{\varepsilon}_{nj}\hat{v}_j \ ; \ p < M$$

Para cada curva X_n, el coeficiente $\hat{\varepsilon}_j$ cuantifica la contribución de $\hat{v}_j$, siendo $\hat{v}_j$ una base ortonormal y se calculan a partir las observaciones una vez convertidas en datos funcionales. Una propiedad de las FPCA:

$$\int \hat{v}_j(t)\hat{v}_j(t)dt = \begin{cases} 0 & \text{si } i \neq j \\ 1 & \text{si } i = j \end{cases}$$

La idea de esta herramienta es poder explicar los datos funcionales a través de una conbinacion de variables ortonormales. La variabilidad total de la muestra sobre la función media muestral se descompone en la suma de variabilidades explicadas por cada estimación FPCA. El porcentaje de variabilidad explicada por $\hat{v}_j$ esta relacionada con el tamaño de las puntuaciones.

4.3. Anexo del Capitulo 3

4.3.1. Detalle de las series temporales seleccionadas

Se presenta 3 tablas detalladas que complementa la información visual proporcionada en el informe principal . La tabla incluye una descripción concisa de las series temporales seleccionadas para el análisis. Cada fila de la tabla representa una serie temporal específica, y se proporcionan detalles clave para facilitar la referencia y el entendimiento.

Grupo	Id Serie	Número de Observaciones	Número de NA
B293	107404	140	9930
B293	1097970	140	9930
B293	176166	205	9865
B293	182251	140	9930
B293	203213	140	9930
B293	288180	51	10019
B293	441506	139	9931
B293	459087	139	9931
B293	502801	70	10000
B293	575856	79	9991
B293	59582	140	9930
B293	632974	140	9930
B293	685197	141	9929
B293	740140	141	9929
B293	758258	141	9789
B293	850765	281	9930
B293	879883	140	9930
B293	891313	140	9930
B293	892679	140	9952
B293	948119	118	9932

Tabla 4.1: Detalles de las Series Temporales Seleccionadas (B293)

Podemos observar que las series, al ser superpuestas en la gráfica 3.3, muestran un solapamiento significativo a pesar de estar medidas en momentos distintos. Es evidente que estas observaciones fueron registradas posiblemente en el mismo día pero en momentos temporales diferentes.

Grupo	Id Serie	Número de Observaciones	Cantidad de NA
B294	1027476	137	9933
B294	1142242	69	10001
B294	1214828	97	9973
B294	1240416	70	10000
B294	1284347	67	10003
B294	137514	140	9930
B294	142749	140	9930
B294	155772	110	9960
B294	180570	135	9935
B294	209188	122	9948
B294	219467	141	9929
B294	282810	129	9941
B294	400687	126	9944
B294	522440	114	9956
B294	620976	141	9929
B294	653331	203	9867
B294	667629	140	9930
B294	790799	84	9986
B294	925859	140	9930
B294	927281	140	9930

Tabla 4.2: Resumen de Observaciones y NA para la Base B294

De manera similar en 4.1 y 4.2, podemos observar que en el grupo B294 se presenta la misma tendencia de superposición, sugiriendo que las observaciones se realizaron posiblemente en el mismo día pero en momentos temporales distintos. Si analizamos los sectores individualmente el sector B295 se observan mas curvas en el mismo momento.

Grupo	Id Serie	Número de Observaciones	Cantidad de NA
B295	1043071	103	9967
B295	1097787	116	9954
B295	1122103	139	9931
B295	121186	108	9962
B295	1212536	131	9939
B295	1288226	92	9978
B295	183661	110	9960
B295	217809	103	9967
B295	233286	252	9818
B295	257010	144	9926
B295	432401	144	9926
B295	493406	144	9926
B295	546742	73	9997
B295	55206	119	9951
B295	751004	141	9929
B295	838881	213	9857
B295	918015	125	9945
B295	918032	139	9931
B295	992284	114	9956
B295	996564	144	9926

Tabla 4.3: Resumen de Observaciones y Cantidad de NA para la Base B295

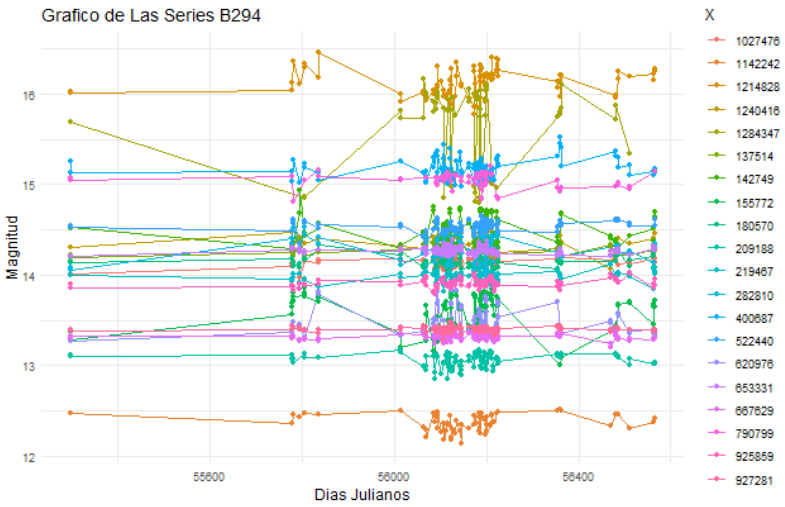


Figura 4.1: Comportamiento de las series correspondiente al grupo B294

Figura 4.2: Comportamiento de las series correspondiente al grupo B295

4.3.2. Regresión Armónica

La base de la regresión armónica radica en la capacidad de representar fenómenos cíclicos mediante funciones trigonométricas. En este contexto, las funciones seno y coseno son particularmente relevantes, el cual busca ajustar un modelo matemático que combine

estas funciones de manera ponderada para representar la serie de tiempo observada.

- Frecuencia: La frecuencia de los términos armónicos determina la velocidad con la que se repiten los ciclos en la serie de tiempo. Identificar la frecuencia correcta es esencial para capturar las variaciones periódicas presentes en los datos.
- Amplitud: La amplitud representa la magnitud de las oscilaciones armónicas y refleja la intensidad de las variaciones observadas.
- Fase: La fase determina el desplazamiento temporal de las funciones senoidales y cosenoidales. Este factor es crucial para alinear correctamente los componentes armónicos con los datos observados.

En el contexto astronómico, la regresión armónica se utiliza para analizar series temporales de observaciones estelares. Esto permite identificar patrones de variación.

Dada una serie temporal Y_t con respecto al tiempo t, un modelo de regresión armónica con K armónicos podría ser:

$$Y_t = \beta_0 + \sum_{k=1}^{K} (\beta_{1k} \cdot \sin(2\pi f_k t) + \beta_{2k} \cdot \cos(2\pi f_k t)) + \varepsilon_t$$

Donde:

- Y_t es el valor de la serie temporal en el tiempo t.
- β_0 es el término constante (intercepto).
- β_{1k} y β_{2k} son los coeficientes de amplitud para el seno y el coseno del armónico k, respectivamente.
- f_k es la frecuencia del armónico k.
- ε_t es el error en el tiempo t.

Este modelo busca capturar patrones repetitivos y oscilaciones en la serie temporal utilizando funciones sinusoidales. La estimación de los coeficientes se realiza mediante técnicas de regresión, como el método de mínimos cuadrados, para minimizar la suma de los cuadrados de los errores (ε_t).

La regresión armónica proporciona una metodología robusta para analizar y comprender la variabilidad temporal en datos astronómicos. Su aplicación adecuada no solo revela patrones cíclicos, sino que también facilita la predicción de eventos futuros, contribuyendo así a la interpretación detallada de las observaciones estelares.

4.3.3. Función recursiva 'ajuste_armonico'

Se presenta una función recursiva 'ajuste_armonico' implementada en el lenguaje de programación R. La función toma como entrada un conjunto de datos ('data') y el número de armónicos ('num_armonicos') que se utilizarán en el ajuste.

La función realiza los siguientes pasos:

1. Identificación de los valores de las series (`X`) en el conjunto de datos.

2. Iteración sobre estos valores y ajuste armónico para cada subconjunto de datos correspondiente a un valor específico de `X`.

3. Creación de una fórmula para el modelo armónico en función del número de armónicos especificado.

4. Ajuste del modelo de regresión lineal usando la función `lm`.

5. Evaluación del rendimiento del ajuste mediante la métrica GCV (Generalized Cross-Validation).

6. Generación de predicciones y comparación con los valores observados.

7. Imputación de los datos, solamente si magnitud arroja un NA

La función devuelve una lista de resultados que incluye un marco de datos con las predicciones y métricas de ajuste para cada valor único de `X`. En caso de que no haya observaciones para ningún valor de `X`, la función devuelve `NULL`. A continuación, se presenta el código de la función:

Comando Software R: Funcion ajuste_armonico

```
ajuste_armonico = function(data, num_armonicos) {
  cl = makeCluster(detectCores())
  registerDoParallel(cl)
  unique_X_values = unique(data$X)
  lista_resultados = foreach(X_value = unique_X_values,.packages = c("data.table","dplyr
      "), .combine = 'c') %dopar% {
    serie  =  data[X == X_value, .SD, .SDcols = -"X"]
    if (nrow(serie) == 0) {
      warning(paste("No hay observaciones para X=", X_value))
      return(NULL)
    }
    serie_melt = melt(serie, variable.name = "dias", value.name = "magnitud")[2:(.N)]
    f1_prueba  =  serie[, f1]
    serie_melt$dias  =  as.numeric(as.character(serie_melt$dias))
    dias  =  serie_melt$dias
    formula_armonica  =  as.formula(paste("magnitud ~", paste(c(
      paste("sin(2*pi*dias*f1_prueba*", 1:num_armonicos, ")", sep = ""),
      paste("cos(2*pi*dias*f1_prueba*", 1:num_armonicos, ")", sep = "")
    ), collapse = " + ")))
    armonico  =  lm(formula_armonica, data = serie_melt)
    nueva_df  =  serie_melt[, c("magnitud", "dias")]
    nueva_df$X  =  rep(X_value, times = nrow(nueva_df))
    nueva_df$predicciones  =  predict(armonico, newdata = data.frame(dias = nueva_df$
        dias, magnitud = nueva_df$magnitud))
    nueva_df$magnitud  =  ifelse(is.na(nueva_df$magnitud), nueva_df$predicciones, nueva_
        df$magnitud)
    predicciones  =  predict(armonico, newdata = data.frame(dias = nueva_df$dias,
        magnitud = nueva_df$magnitud))
    valores_observados  =  nueva_df$magnitud
    traza_H  =  sum(hatvalues(armonico))
    n  =  length(valores_observados)
    gcv  =  (n / (n - traza_H)) * sum((valores_observados - predicciones)^2 / (1 -
        hatvalues(armonico))^2)
    return(list(list(nueva_df = nueva_df, gcv = gcv)))
  }
  lista_resultados <- lista_resultados[sapply(lista_resultados, function(x) nrow(x$nueva
      _df) > 0)]
  stopCluster(cl)
  if (length(lista_resultados) > 0) {
    resultados <- do.call(rbind, lista_resultados)
    return(resultados)
  } else {
    return(NULL)
  }
}
```

Nota: Esta función está implementada a través de un algoritmo de separación de núcleos, optimizando la creación y ajuste de los armónicos dividiendo la tarea en la cantidad de núcleos disponibles del equipo. El tiempo de espera dependerá de la cantidad de bases, la cantidad de tiempos a imputar y principalmente de la cantidad de núcleos del computador, debido a que la función contempla el trabajo en paralelo con la librería doParallel.

.

4.3.4. Función recursiva ”corte_optim”

Esta función implementada en el lenguaje de programación R toma como entrada un conjunto de datos referente a los datos ajustados del modelo de regresión funcional (’res_fitted_values’) y la variable dicotómica utilizada como respuesta en el ajuste del modelo de regresión (’rrab_bin_ind’).

Comando Software R: Función corte_optimo

```
corte_optim <- function(res_fitted_values, rrab_bin_ind) {
  precision <- numeric()
  sensibilidad <- numeric()
  especificidad <- numeric()
  youden <- numeric()
  lista <- numeric()
  for (i in 0:700) {
    lista[i+1] <- i*0.001
    predicciones <- ifelse(res_fitted_values > 0.001 * i, 1, 0)
    tabla_contingencia <- table(predicciones, rrab_bin_ind)
    if (nrow(tabla_contingencia) == 2 && ncol(tabla_contingencia) == 2) {
      sensibilidad <- c(sensibilidad, tabla_contingencia[2, 2] / sum(rrab_bin_ind == 1))
      especificidad <- c(especificidad, tabla_contingencia[1, 1] / sum(rrab_bin_ind ==
          0))
      precision <- c(precision,(tabla_contingencia[1, 1] +tabla_contingencia[2, 2])/ (
          sum(rrab_bin_ind == 0)+sum(rrab_bin_ind == 1)))
      youden <- c(youden, sensibilidad[length(sensibilidad)] + especificidad[length(
          especificidad)] - 1)
    } else {
      sensibilidad <- c(sensibilidad, NA)
      especificidad <- c(especificidad, NA)
      youden <- c(youden, NA)
      precision <- c(precision,NA)
    }}
  max_youden <- max(youden, na.rm = TRUE)
  valor_lista <- lista[which.max(youden)]
  sensibilidad<-sensibilidad[which.max(youden)]
  precision <- precision[which.max(youden)]
  return(list(max_youden = max_youden, valor_lista = valor_lista, sensibilidad=
      sensibilidad,precision=precision))
}
```

La función devuelve una lista de resultados que incluye los siguientes valores: Punto de corte óptimo para la discriminación,especificidad sensibilidad, precisión e índice de youden.

4.3.5. Tablas GCV

Se presenta la tabla resumen que muestra los resultados del cálculo del GCV para cada serie temporal seleccionada previamente. La tabla incluirá información sobre el número de armónicos utilizados y los valores correspondientes de GCV. Este análisis facilitará la identificación del modelo armónico más adecuado para la posterior imputación de datos.

GCV	Cantidad de Armónicos - B293						
Serie	1	2	3	4	5	6	7
59582	0.09381788294	0.09645989919	0.09927683365	0.10247708056	0.10483335803	0.10627098472	0.10655325682
107404	0.54705331215	0.37656562622	0.34124452317	0.34221579678	0.34277890516	0.33989228446	0.33334525541
176166	1.21738299614	1.15705335269	1.18139044913	1.19297712420	1.22228035466	1.25498588054	1.28927420164
182251	0.23218429434	0.18341379176	0.18246531836	0.17976602926	0.18396949172	0.18374771501	0.19095016769
203213	0.29417119600	0.30377120169	0.29212734670	0.29199912792	0.30074105696	0.33576771811	0.37806473567
288180	0.04177609442	0.04477030998	0.04513351708	0.04612070498	0.04793936874	0.05208708890	0.05563560729
441506	0.32415751108	0.32681979758	0.33226862577	0.33432927255	0.31921385390	0.33414132376	0.34776404986
459087	1.38566595723	1.13421914609	1.07832288792	1.09850416298	1.13209017333	1.13751043649	1.17206193526
502801	0.12731882874	0.08553126051	0.06812553805	0.05882357997	0.05637071283	0.06128301887	0.06051856955
575856	1.80539065788	1.54248183362	1.61943812786	1.68897319489	2.20854622415	3.16480775706	2.67774960506
632974	0.06112558955	0.06245953662	0.06222950860	0.05999719670	0.06254383945	0.06560118541	0.06909004440
685197	0.09378062849	0.09274219571	0.08757915777	0.09020156926	0.08815125267	0.08948308036	0.09356392018
740140	0.57920468547	0.58492535029	0.59207454919	0.58618071886	0.59323473107	0.61232224713	0.62686129074
758258	0.21396407374	0.21444618417	0.20754312890	0.20980512467	0.20544517448	0.20775523690	0.21057070339
850765	0.09720875393	0.09357292447	0.09612521946	0.09728921979	0.10040426512	0.09933234499	0.10338614366
879883	0.08435080854	0.07106655798	0.07329084073	0.07496005269	0.07084228523	0.07267580783	0.07217442347
891313	0.46321404249	0.44689677469	0.45815491815	0.45269361898	0.45622350945	0.45614164585	0.46905752353
892679	4.06527175547	3.60363531020	3.73569585074	4.10172147961	4.19920697207	4.36193293460	4.34734642721
948119	0.47401774744	0.45811296317	0.47022357277	0.47329940759	0.48404613497	0.49548751868	0.51049452273
1097970	0.26110685743	0.15820985223	0.15030252116	0.14838020952	0.15241173867	0.15221614482	0.15619483172

Tabla 4.4: Cálculo de GCV para cada serie - Grupo B293

GCV	Cantidad de Armónicos - B294						
Serie	1	2	3	4	5	6	7
137514	0.21326634943	0.21788221596	0.20789652401	0.21348460776	0.21916704285	0.2214910491	0.2274167989
142749	0.94899484298	0.95490218728	0.89006841552	0.89806782774	0.87005063343	0.81567384	0.8258284002
155772	2.24776661688	2.23988703102	2.25673823398	2.19853523378	3.14083040884	40045.4069	897375186.7
180570	0.17962576817	0.12550417306	0.12072522004	0.11923020808	0.12001245891	0.1153736125	0.1159908676
209188	0.49846428967	0.5038764255	0.52020301752	0.52778590599	0.55037358864	0.5698486097	0.5798173959
219467	0.1833742093	0.18492132983	0.1847700549	0.1894731172	0.19195017946	0.1976455918	0.2042690235
282810	1.43042870388	1.22643481094	1.27613907938	1.67558568022	4.7471216005	28.88579878	166.218502
400687	0.67092660818	0.65787057017	0.52190847897	0.51259637064	0.51327851464	0.5288331633	0.5418896603
522440	0.21285576793	0.21931736301	0.21698791794	0.2243854374	0.2139199265	0.2119714191	0.2170629707
620976	1.72015506329	1.71286622566	1.57589909016	1.46972313747	1.50690741148	1.51553405	1.564801497
653331	0.18993001381	0.19061780294	0.18702959528	0.18305948132	0.1860992143	0.1891609564	0.1926225672
667629	0.08887217551	0.09109457237	0.09078828935	0.09369687314	0.09688945406	0.0985811579	0.101182172
790799	0.37938619555	0.34956430945	0.34545332852	0.36394801875	0.3842970386	0.3989148203	0.3984512967
925859	0.16323326426	0.15893912603	0.16251028019	0.16054211022	0.16056445345	0.1619174286	0.164899119
927281	0.07113965274	0.07215201154	0.07413715586	0.07725727467	0.0786025262	0.08334134697	0.08714595775
1027476	0.2506276364	0.24883768408	0.25645515537	0.2457880456	0.24795156732	0.2560630282	0.2548723791
1142242	0.32337689938	0.31069190442	0.29924430087	0.30761587708	0.31705175693	0.3246984012	0.3273425895
1214828	0.97078020005	0.99352934885	1.00707114706	1.05225862041	1.07291495544	1.063219761	1.137759591
1240416	0.51350210931	0.54599152503	0.56100905319	0.57879263946	0.61751462816	0.647897185	0.7134704787
1284347	3.50278593331	1.34997872853	1.20541204616	1.13460950191	1.17407825015	1.052658475	1.109075046

Tabla 4.5: Calculo de GCV para cada serie - Grupo B294

GCV	Cantidad de Armónicos - B294						
Serie	1	2	3	4	5	6	7
55206	0.61273237408	0.62613288388	0.61781772817	0.65266619827	0.6783575482	0.7323538966	0.7484523224
121186	0.09057212174	0.09034671446	0.09502316299	0.09866440400	0.1042366629	0.1073383545	0.1138739839
183661	0.64693348307	0.59721657616	0.56998449728	0.50335305606	0.5225019891	0.5269718816	0.5391333859
217809	1.04373926901	0.85435753588	0.76628968457	0.55673434244	0.5611771500	0.5680459585	0.6501349884
233286	1.28642442716	1.02444143201	1.03791647940	0.85023331816	0.7796480968	0.8122060648	0.8628657346
257010	0.91433129437	0.91507084644	0.71061517422	0.74011367781	0.7980164145	1.0351108800	1.507511951
432401	0.13226825115	0.12413183197	0.12579887375	0.12311650807	0.1243781415	0.1260517370	0.1275835256
493406	0.13272467796	0.08401382743	0.08550350581	0.08795621548	0.0879935451	0.0915747104	0.0951187678
546742	0.24904237583	0.21823441771	0.22911140903	0.22288469857	0.2362508001	0.2500448065	0.2660333829
751004	0.34501394907	0.32759030284	0.30138170911	0.30958415289	0.3030705715	0.3148052128	0.3250520964
838881	0.25302179761	0.25913248365	0.26046094567	0.24882548638	0.2363207828	0.2367000543	0.2425708167
918015	1.22579200156	1.27359807622	1.35312210542	11.96923711639	1701.150766	69866.63361	20678742.21
918032	0.08536777066	0.08551497018	0.08588625601	0.08735667469	0.0866606042	0.0858839881	0.08778744898
992284	0.16614733453	0.17267104430	0.17844000875	0.18400763127	0.1910472793	0.1988290849	0.2066577721
996564	0.70196887350	0.59624432075	0.58347565414	0.60963900423	0.6406547733	0.6715923758	0.7070522272
1043071	0.90152554098	0.88953047022	0.84797778946	0.87327115357	0.8640954788	0.8982208238	0.8838203298
1097787	0.88225293034	0.91022160942	0.87993506255	0.92309379114	0.9552244404	1.0088029250	1.0450916560
1122103	0.43114670878	0.40410452356	0.39610753438	0.39458332793	0.3759679901	0.4065083837	0.4197909375
1212536	0.19758581741	0.20412364815	0.20465345556	0.20594113069	0.2098072916	0.2139664628	0.2220185807
1288226	0.47672652708	0.48020485582	0.50207603474	0.50861105297	0.5205725923	0.5274975511	0.5668929077

Tabla 4.6: Calculo de GCV para cada serie - Grupo B295

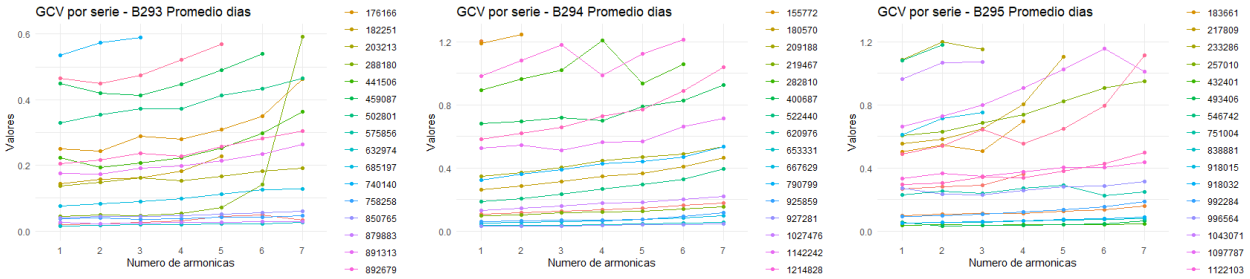
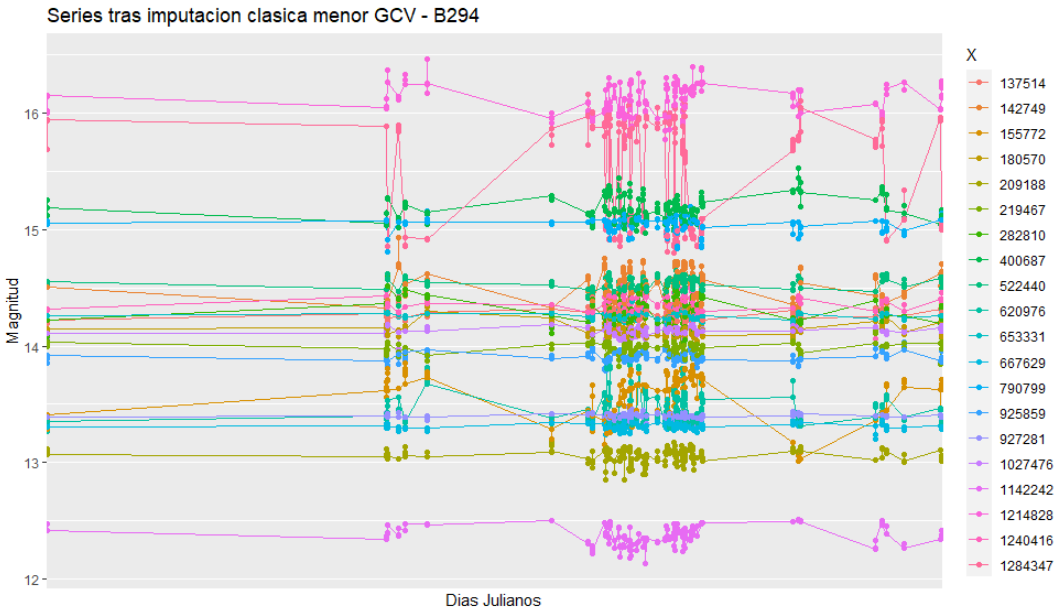
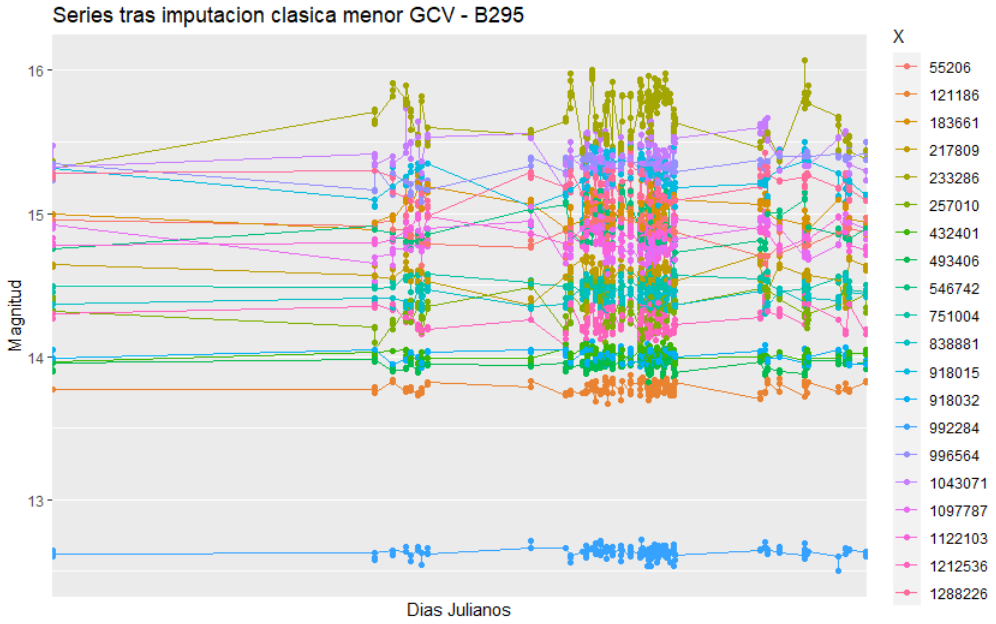


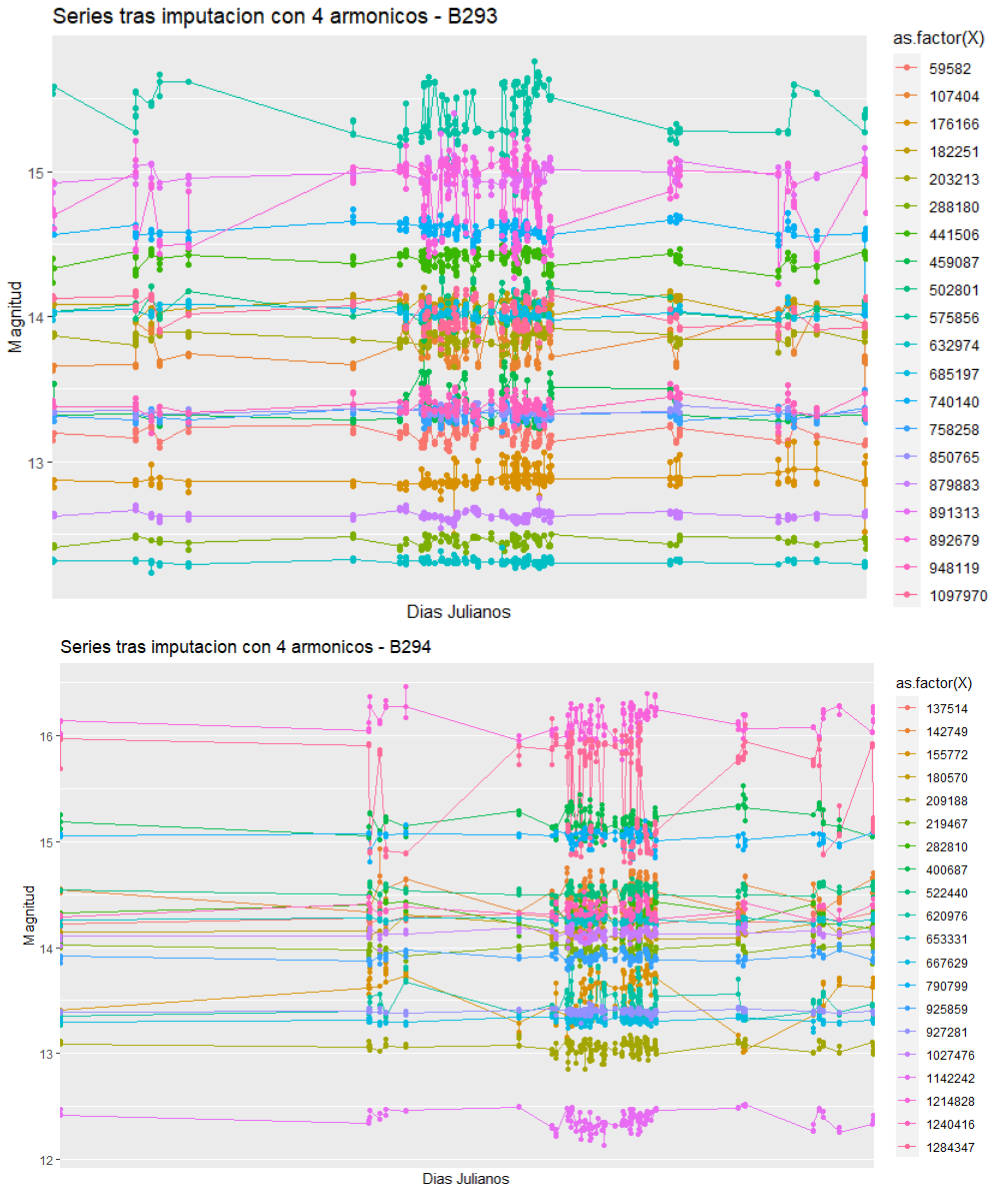
Figura 4.3: Comportamiento de GVC bajo imputación clásica

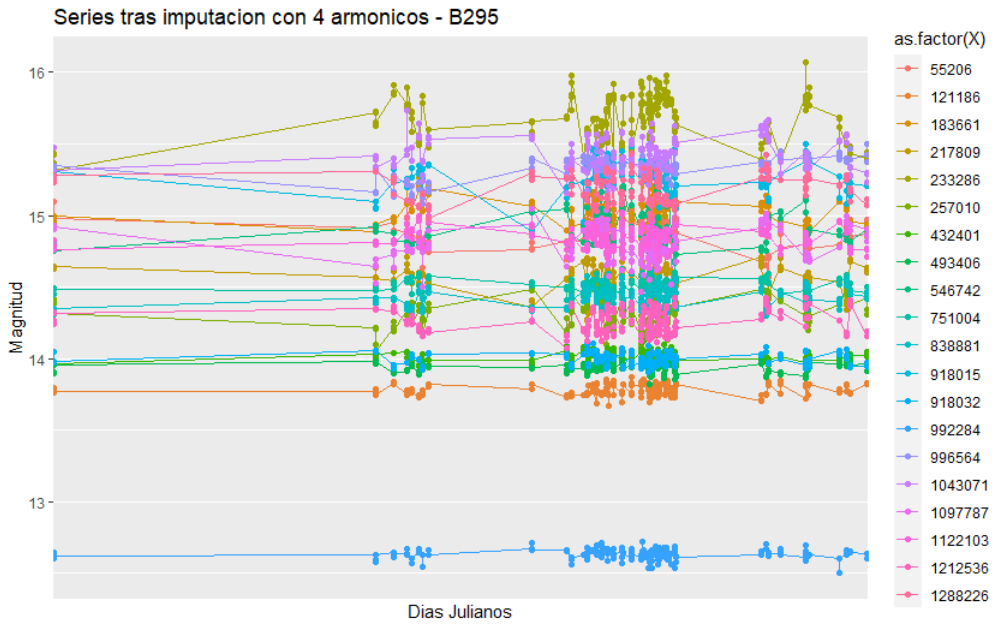
4.3.6. Análisis de Imputación con Enfoque en el Menor Valor de GCV


Series tras imputacion clasica menor GCV - B294
Magnitud
Dias Julianos
X
137514
142749
155772
180570
209188
219467
282810
400687
522440
620976
653331
667629
790799
925859
927281
1027476
1142242
1214828
1240416
1284347


Series tras imputacion clasica menor GCV - B295
Magnitud
Dias Julianos
X
55206
121186
183661
217809
233286
257010
432401
493406
546742
751004
838881
918015
918032
992284
996564
1043071
1097787
1122103
1212536
1288226

4.3.7. Comparación de Estrategias de Imputación: Evaluación de Desempeño con 4 Armónicos





4.3.8. Detalle de las series temporales seleccionadas con días promediados

Grupo	Id Serie	Número de Observaciones	Número de NA
B293	107404	66	0
B293	1097970	66	1
B293	176166	66	0
B293	182251	66	1
B293	203213	66	1
B293	288180	66	33
B293	441506	66	1
B293	459087	66	1
B293	502801	66	1
B293	575856	66	23
B293	59582	66	1
B293	632974	66	1
B293	685197	66	0
B293	740140	66	0
B293	758258	66	0
B293	850765	66	1
B293	879883	66	1
B293	891313	66	1
B293	892679	66	8
B293	948119	66	2

Tabla 4.7: Detalles de las Series Temporales Seleccionadas (B293)

Grupo	Id Serie	Número de Observaciones	Cantidad de NA
B294	1027476	66	2
B294	1142242	66	2
B294	1214828	66	12
B294	1240416	66	1
B294	1284347	66	27
B294	137514	66	1
B294	142749	66	0
B294	155772	66	11
B294	180570	66	3
B294	209188	66	6
B294	219467	66	0
B294	282810	66	4
B294	400687	66	4
B294	522440	66	10
B294	620976	66	0
B294	653331	66	1
B294	667629	66	1
B294	790799	66	18
B294	925859	66	1
B294	927281	66	1

Tabla 4.8: Resumen de Observaciones y NA para la Base B294

Grupo	Id Serie	Número de Observaciones	Cantidad de NA
B295	1043071	67	9
B295	1097787	67	8
B295	1122103	67	3
B295	121186	67	9
B295	1212536	67	5
B295	1288226	67	27
B295	183661	67	13
B295	217809	67	12
B295	233286	67	11
B295	257010	67	0
B295	432401	67	0
B295	493406	67	0
B295	546742	67	0
B295	55206	67	7
B295	751004	67	0
B295	838881	67	0
B295	918015	67	5
B295	918032	67	0
B295	992284	67	10
B295	996564	67	0

Tabla 4.9: Tabla con los nuevos valores para la Base B295

4.3.9. Cálculo del GCV para la Elección del Mejor Modelo Armónico

Las siguientes tablas resumen los resultados obtenidos para las diferentes configuraciones de regresiones armónicas aplicadas al haber promediado las observaciones diarias de las series.

GCV	Cantidad de Armónicos - B293						
Serie	1	2	3	4	5	6	7
59582	0.02032796438	0.02245287550	0.02544096141	0.03121055313	0.04377886620	0.04942091778	0.03224631235
107404	1.65451494714	1.75390295655	1.90208469318	1.97244490733	2.23863495083	1.69069533699	1.82546027774
176166	0.25089852248	0.24424765831	0.28817213478	0.28081605094	0.30906577503	0.35038868176	0.46386103261
182251	0.14494160759	0.15798562341	0.16151800406	0.18142384032	0.22803456432	0.62116103640	0.36341771101
203213	0.13721568781	0.14866492904	0.16286117998	0.15212315269	0.16760982935	0.18308800755	0.19140217636
288180	0.04476335158	0.04986460729	0.04683974169	0.05301431794	0.07112257504	0.14264502691	0.59123528841
441506	0.22261504309	0.19475282187	0.20840344177	0.22256143187	0.25286128558	0.29878768514	0.36230068388
459087	0.44850194253	0.41916440344	0.41205004994	0.44696751300	0.49006556847	0.53979499406	0.61850032530
502801	0.32935495865	0.35386648056	0.37185027628	0.37344901441	0.41226574681	0.43251385752	0.46441733381
575856	1.50028831220	1.68373062634	1.99361773139	2.24836897118	2.83281153926	3.50932789466	4.66048979703
632974	0.01591364171	0.01729167480	0.01877774762	0.01940580680	0.02146686276	0.02300842188	0.02555506839
685197	0.07722661924	0.08368110886	0.08875935030	0.09926529830	0.11183065233	0.12592020841	0.12821914562
740140	0.53557496014	0.57295387527	0.59067330992	0.63914668446	0.68430859152	0.72604123080	0.78413997471
758258	0.03756352948	0.03958371418	0.03509081741	0.03790929313	0.04149909713	0.04260604993	0.04702395687
850765	0.03981036354	0.04357851299	0.04533314358	0.04764136183	0.05107823048	0.05560054249	0.06063905402
879883	0.02705875467	0.02194784851	0.02423193969	0.02672210481	0.02595906822	0.02805514642	0.02892553915
891313	0.17461034306	0.17298003009	0.19157223755	0.19793567000	0.21491580963	0.23342026781	0.26307692056
892679	3.72327859436	3.82862921112	4.09041651858	3.88296059290	4.34365365491	4.88798478136	5.18926759524
948119	0.20421513242	0.21595375334	0.23578891453	0.22824940951	0.25599131556	0.28255137788	0.30352587823
1097970	0.46524109625	0.44879360942	0.47492627586	0.52214551472	0.56870485092	0.64463886898	0.64140373544

Tabla 4.10: Calculo de GCV para cada serie con magnitud promediada - Grupo B293

GCV	Cantidad de Armonicos						
Serie	1	2	3	4	5	6	7
137514	0.10724481698	0.11605252795	0.12730804057	0.13462645896	0.1469869233	0.1638109348	1.770382135e-01
142749	1.20238982620	1.30388132558	1.41834770325	1.45203245076	1.582305323	1.662665653	1.835350341
155772	1.19189267600	1.24785543271	2.50103908043	11.69294913839	1381.860176	29327019.57	298344587000
180570	0.26006754739	0.28660731973	0.31317910381	0.34564074002	0.3662137755	0.4065533129	0.4632433455
209188	0.34843242702	0.37293004581	0.40183296622	0.44475741400	0.4713224553	0.4893895412	0.5340224040
219467	0.09881382521	0.10447421814	0.11415688200	0.12044260874	0.1279245237	0.1399914529	0.1546235113
282810	0.89285293581	0.96627390527	1.02294467153	1.21172243057	0.9371324774	1.059948339	2.410343995
400687	0.68226650557	0.69472257030	0.72107320229	0.69980478119	0.7906940002	0.8299354764	0.9261375782
522440	0.18519902984	0.20766276819	0.23492162179	0.26547549476	0.2936279219	0.3279591003	0.3925146875
620976	1.40178590797	1.50759095969	1.61788612742	1.71136606319	1.881529439	1.884779085	2.064731491
653331	0.03502526079	0.03730553986	0.03850354599	0.03971501024	0.04359431915	0.04822461688	0.05378388251
667629	0.05126802842	0.05550237130	0.05871235629	0.06552139127	0.07398415164	0.08536875775	0.09692359696
790799	0.32521328461	0.36290211633	0.38996925380	0.42854642027	0.4413089224	0.4687620837	0.5351730476
925859	0.06115444064	0.06275008208	0.06899984997	0.07151726000	0.07593169868	0.09312936928	0.1143432193
927281	0.02939517456	0.02925478945	0.03222044450	0.03536406158	0.03885960634	0.04000083312	0.04443568708
1027476	0.13228994781	0.14606814097	0.15745547360	0.17540168386	0.1833206909	0.2034748577	0.2204445747
1142242	0.52806493709	0.54494376979	0.51156646692	0.56553338136	0.5704514467	0.6607861305	0.7125617445
1214828	0.98471044852	1.08155355921	1.18313295918	0.98645238911	1.124093135	1.215490334	1.386501150
1240416	0.58465934138	0.61844337641	0.65928100178	0.73114052085	0.7696028350	0.8890636758	1.042008111
1284347	8.79647465470	9.58741339746	11.27477249150	11.60931910751	44.0142066	239.1529549	5459.012878

Tabla 4.11: Calculo de GCV para cada serie con magnitud promediada- Grupo B294

GCV	Cantidad de Armonicos						
Serie	1	2	3	4	5	6	7
55206	0.26739592385	0.28096033430	0.28979029968	0.36361971324	2.22570609595	4.111474237	9.684064114
121186	0.09837892519	0.10585047292	0.10969610812	0.11216148326	0.12362367882	0.1364802458	0.1595454952
183661	0.50308412563	0.54629380540	0.50916003269	0.69553331406	1.67131213747	9.399570493	0.5624768585
217809	0.55466685276	0.58206277018	0.65074981303	0.80291360848	1.10635185039	3.241382691	10.32355219
233286	1.08885220550	1.19948817238	1.15192408388	1.28779766413	1.43500131330	1.630893360	2.015528904
257010	0.60457779435	0.63027005732	0.68472785210	0.73912684208	0.82306478682	0.9067799296	0.9499559970
432401	0.03758574710	0.03867786415	0.03784197553	0.03954486458	0.04112832287	0.04301205483	0.04563648530
493406	0.05387926030	0.03289701589	0.03484103267	0.03815263155	0.04162506005	0.04650171376	0.06400978525
546742	1.08027650950	1.18196105286	1.26899992003	1.42178626689	1.55703401049	1.787411348	1.944413717
751004	0.23117597339	0.25080217010	0.24045296893	0.26993520818	0.29185915620	0.2263170054	0.2502519017
838881	0.04865767652	0.05321635691	0.05701768174	0.06327977822	0.06798191225	0.07599613995	0.07865031241
918015	0.61103184752	0.71557000625	0.75048846768	3.79551515924	20.12292373304	2514.180711	18601.81680
918032	0.05122514774	0.05573398147	0.06098007827	0.06622052520	0.07234098991	0.08085203082	0.08911512516
992284	0.09190968347	0.09931021035	0.10752196538	0.11978763991	0.13537190480	0.1517341009	0.1872738737
996564	0.27147558206	0.23322298577	0.22874651892	0.25785177127	0.28025992665	0.2861197062	0.3142213572
1043071	0.96607265005	1.06603099902	1.07438212357	1.26161053768	1.37710051719	1.538002446	1.629461090
1097787	0.66144711327	0.72803378557	0.80081063849	0.90829074331	1.02467647525	1.157035702	1.011032015
1122103	0.33513521526	0.36387757836	0.34612867683	0.37544674362	0.40467410384	0.4057944136	0.4354854589
1212536	0.29454237635	0.30568910287	0.34096737475	0.33967715004	0.38217796516	0.4265219462	0.4998365493
1288226	0.48959166526	0.54176187618	0.64178209614	0.55655093247	0.65021066347	0.7933205482	1.114221273

Tabla 4.12: Calculo de GCV para cada serie con magnitud promediada - Grupo B295

4.3.10. Análisis de Imputación con Enfoque en el Menor Valor de GCV a series promediadas diariamente

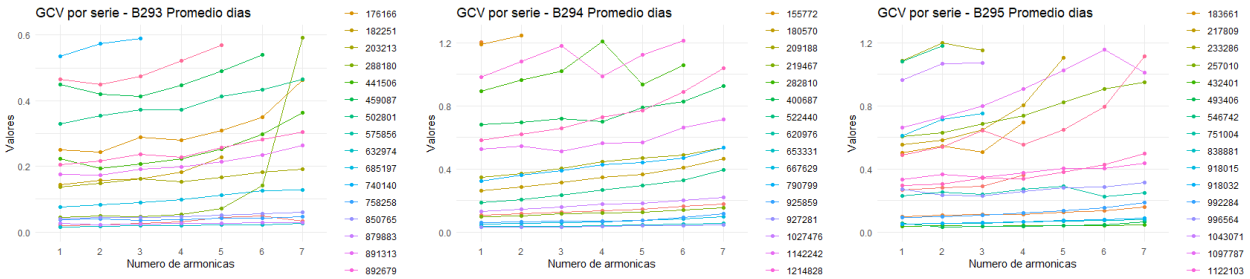


Figura 4.4: Comportamiento de GVC

A pesar de las expectativas iniciales, el análisis de los valores de GCV derivados de las regresiones armónicas revela una tendencia inesperada en las tablas 4.10, 4.11, y 4.12. En la mayoría de las series temporales, los resultados sugieren que la elección óptima, según los criterios de GCV, es optar por una sola regresión armónica. En los casos donde el valor es mayor, la diferencia en la métrica no es significativa.

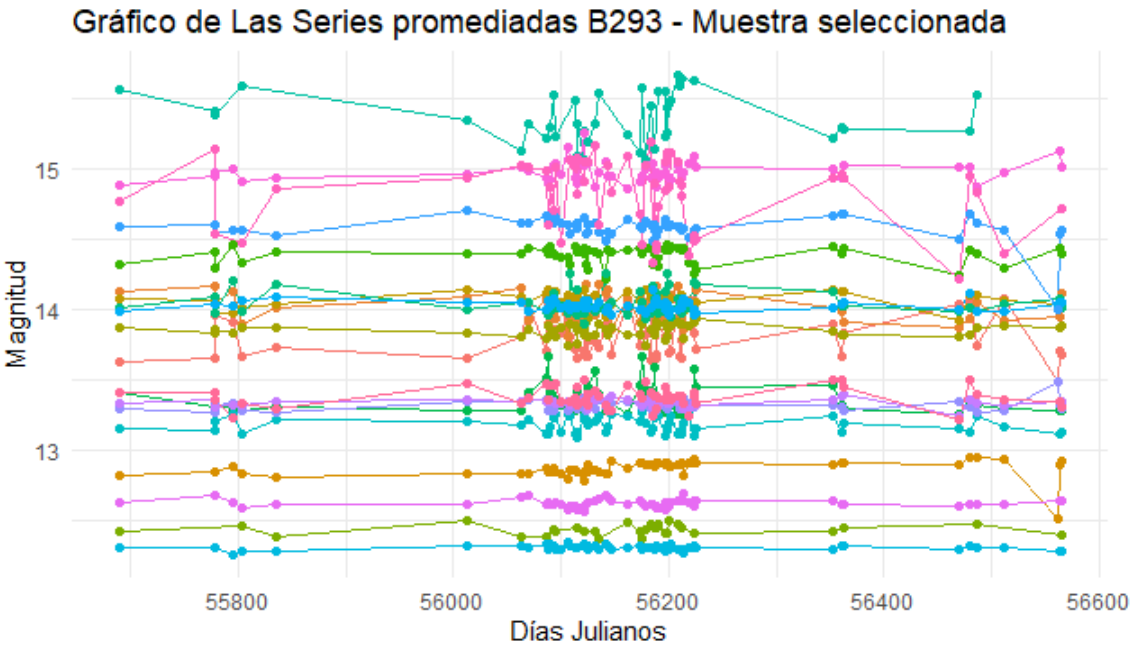


Gráfico de Las Series promediadas B293 - Muestra seleccionada
Magnitud
15
14
13
55800
56000
56200
56400
56600
Días Julianos

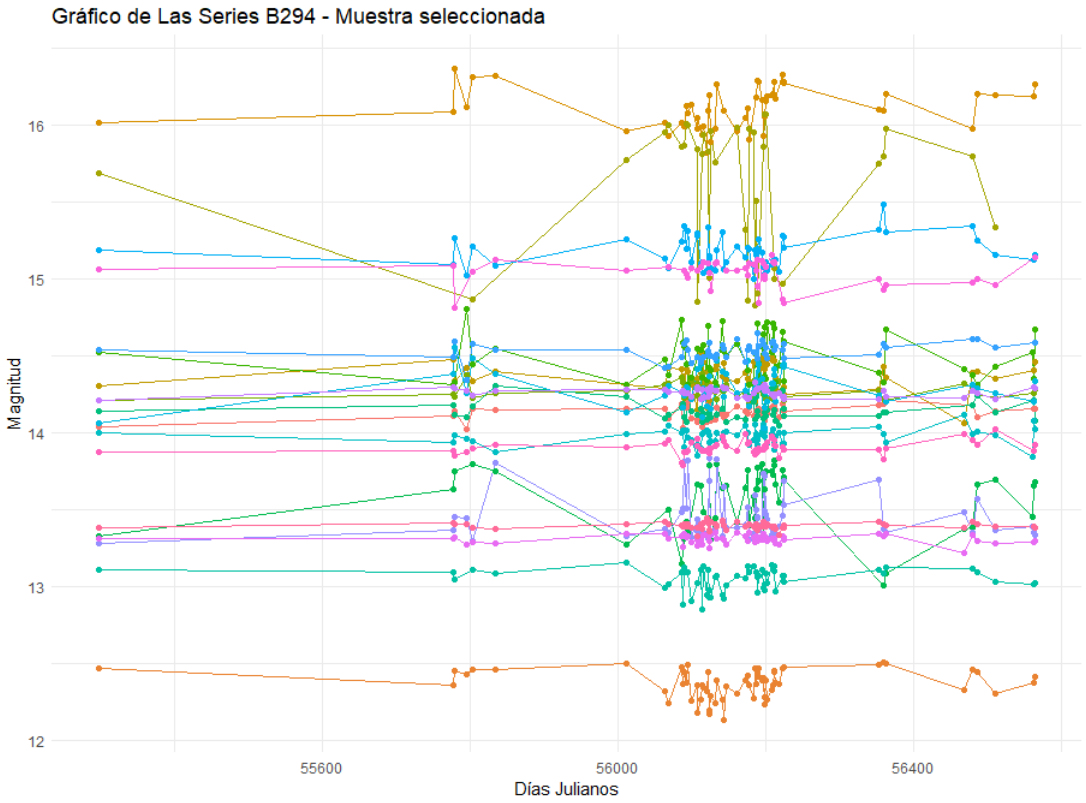


Gráfico de Las Series B294 - Muestra seleccionada
Magnitud
16
15
14
13
12
55600
56000
56400
Días Julianos

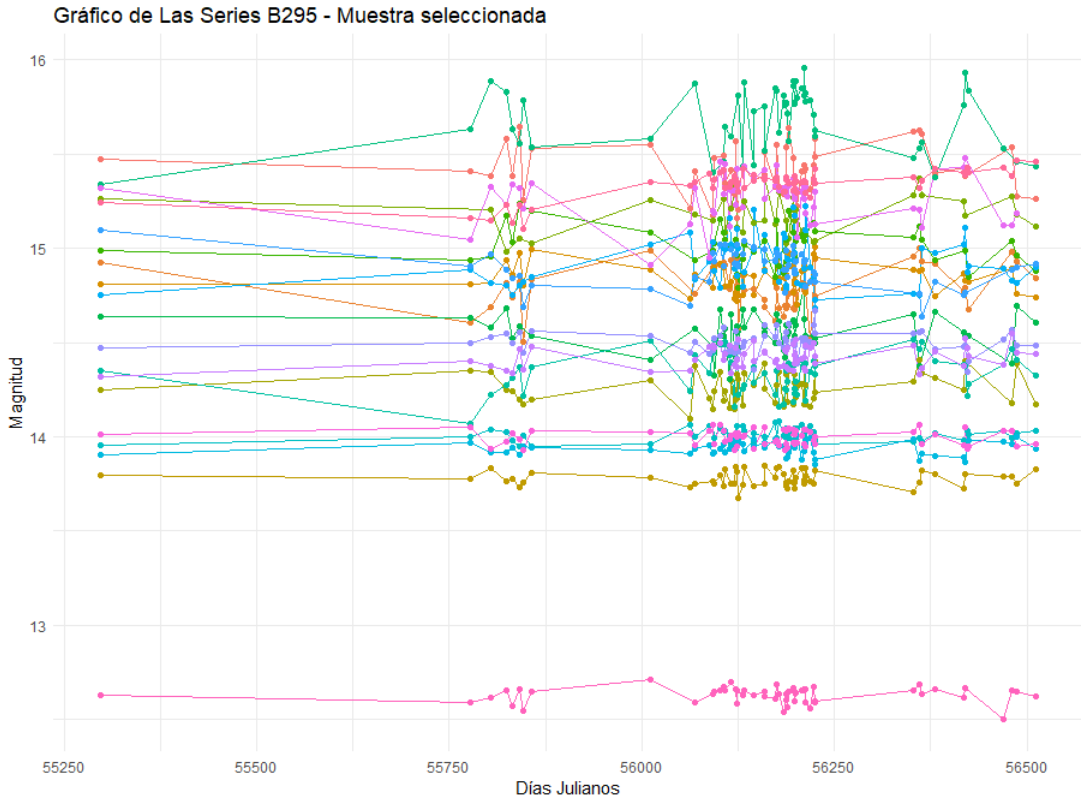


Figura 4.5: Comportamiento de las series tras calcular el promedio diario

4.4. Transformación Fold

4.4.1. Función recursiva ”transformacion()”

Esta función transformacion en R toma como argumento un data.frame, la cual recorre todas las curvas realizando a cada una la transformación folded, devuelviendo el mismo data.frame pero con las nuevas columnnas de los tiempos transformados. A continuación, se presenta el código de la función:

Comando Software R: Función tranformacion()

```
transformacion <- function(dt) {
  X <- unique(dt$X)
  for (valor in X) {
    filas_valor <- dt[dt$X == valor, ]
    resultado <- foldlc(filas_valor[c(4,5,6)], unique(filas_valor$f1), plot =
        FALSE)$folded
    resultado <- resultado[order(as.numeric(rownames(resultado))),]
    dt[dt$X == valor, "dias_fold"] <- resultado[1]
    dt[dt$X == valor, "magnitud_fold"] <- resultado[2]
    dt[dt$X == valor, "error_fold"] <- resultado[3]
  }
  return(dt)
}
```

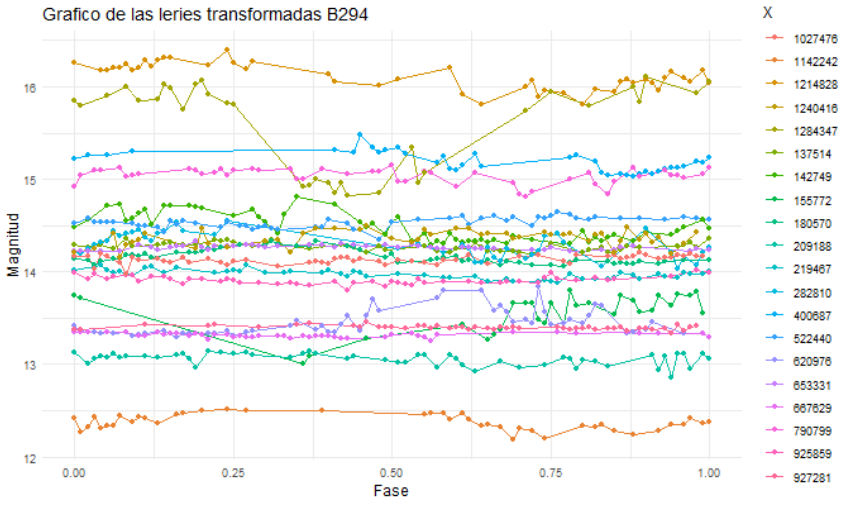


Figura 4.6: Comportamiento de la transformación correspondiente al grupo B294

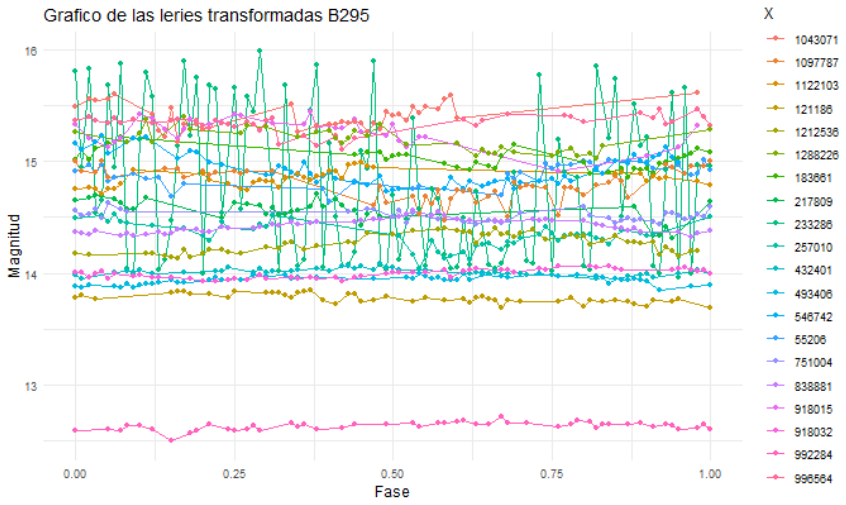


Figura 4.7: Comportamiento de la transformación correspondiente al grupo B295

4.4.2. Tablas GCV

4.5. Regresión logística

Como bien se explica en la sección 3.1 las bases de datos que ya han sido imputadas tienen que necesariamente ser transformadas a datos funcionales, a través de representaciones bases (B-spline, pc y Fourier 2.1.2).

El primer obstáculo que siempre se tendrá al analizar datos funcionales es encontrar una representación adecuada para los datos. Típicamente, el conjunto de datos funcionales $\{X_1, \ldots, X_n\}$ se evalúa en un número de puntos de discretización $t_1, \ldots, t_m$ que pueden observarse de manera no equiespaciada.

4.5.1. Función fdata()

El paquete fda.usc evita la transformación de bases realizada por el paquete fda y define un objeto llamado fdata como una lista de los siguientes argumentos:

- **data**: típicamente una matriz de dimensión (n x m) que contiene un conjunto de n curvas discretizadas en m puntos o argvals.
- **argvals**: ubicaciones de los puntos de discretización.

- **rangeval**: rango de puntos de discretización, por defecto: rango(argvals).
- **names**: lista opcional con tres componentes: main, un título general; xlab, un título para el eje x; y ylab, un título para el eje y.

Esta función retorna una lista que corresponde a la clase fdata.

4.5.2. Función optim.base()

por otro lado existe la función optim.base (antes llamada min.basis) que permite encontrar el numero óptimo de bases en la estimación de datos funcionales mediante GCV, esta función permite realizar el ajuste mediante alguna penalización. los argumentos a ingresar son:

- **fdataobj**: objeto de clase fdata.
- **lambda**: parámetro de penalización.
- **numbasis**: numero de bases a evaluar.
- **type.basis**: Tipo de base a transformar, por defecto es bspline.

retorna valores correspondiente a GCV de cada base, gcv.optim, el numero óptimo de base, entre otras cosas.

4.5.3. Transformación a datos funcionales

4.5.3.1. Representación base de fourier y bspline

La función create.fourier.basis() o create.fdata.basis() permite generar una base de Fourier adaptada a los datos funcionales. los argumentos son:

- **rangeval**: Rango de valores permitidos en la función.
- **nbasis**: numero de bases.
- **period**: periodo t aplicado.

retorna un objeto basisfd.

4.5.4. Imputación clásica menor GCV

Regresión logística funcional - Base de entrenamiento									
Bspl	Enlace	Devianza	AIC	Sensibilidad	Especificidad	Precisión	F1	Youden	AUC
4	Logit	1734.182	1744.182	0.9009	0.4078	0.4337	0.5856	0.3087	0.6544
	Probit	1733.424	1743.424	0.8873	0.4181	0.4427	0.5907	0.3054	0.6527
	Cauchit	1736.705	1746.705	0.9594	0.3345	0.3674	0.5313	0.294	0.647
	Cloglog	1734.302	1744.302	0.9009	0.4068	0.4328	0.5847	0.3077	0.6539
5	Logit	1734.182	1746.182	0.9009	0.4078	0.4337	0.5856	0.3087	0.6544
	Probit	1733.423	1745.423	0.8828	0.4183	0.4427	0.5897	0.3012	0.6506
	Cauchit	1736.688	1748.688	0.9594	0.3328	0.3657	0.5296	0.2922	0.6461
	Cloglog	1734.302	1746.302	0.9009	0.4066	0.4325	0.5845	0.3075	0.6538
6	Logit	1734.096	1748.096	0.9009	0.4088	0.4347	0.5864	0.3097	0.6549
	Probit	1733.316	1747.316	0.8828	0.4206	0.4449	0.5916	0.3034	0.6517
	Cauchit	1736.661	1750.661	0.9639	0.33	0.3634	0.5278	0.294	0.647
	Cloglog	1734.219	1748.219	0.9054	0.4081	0.4342	0.5869	0.3135	0.6568
7	Logit	1734.078	1750.078	0.9009	0.4096	0.4354	0.5871	0.3105	0.6553
	Probit	1733.295	1749.295	0.8873	0.4206	0.4451	0.5928	0.3079	0.654
	Cauchit	1736.643	1752.643	0.9639	0.3293	0.3627	0.527	0.2932	0.6466
	Cloglog	1734.202	1750.202	0.9054	0.4083	0.4344	0.5871	0.3137	0.6569
Fourier	Enlace	Devianza	AIC	Sensibilidad	Especificidad	Precisión	F1	Youden	AUC
4	Logit	1734.166	1746.166	0.9009	0.4078	0.4337	0.5856	0.3087	0.6544
	Probit	1733.403	1745.403	0.8828	0.4198	0.4442	0.591	0.3027	0.6514
	Cauchit	1736.684	1748.684	0.9594	0.333	0.366	0.5299	0.2925	0.6463
	Cloglog	1734.286	1746.286	0.9009	0.4073	0.4333	0.5851	0.3082	0.6541
5	Logit	1734.166	1746.166	0.9009	0.4078	0.4337	0.5856	0.3087	0.6544
	Probit	1733.403	1745.403	0.8828	0.4198	0.4442	0.591	0.3027	0.6514
	Cauchit	1736.684	1748.684	0.9594	0.333	0.366	0.5299	0.2925	0.6463
	Cloglog	1734.286	1746.286	0.9009	0.4073	0.4333	0.5851	0.3082	0.6541
6	Logit	1734.132	1750.132	0.8963	0.4088	0.4344	0.5852	0.3052	0.6526
	Probit	1733.364	1749.364	0.8783	0.4208	0.4449	0.2992	0.5906	0.6496
	Cauchit	1736.665	1752.665	0.9594	0.3343	0.3672	0.5311	0.2937	0.6469
	Cloglog	1734.253	1750.253	0.8963	0.4071	0.4328	0.5837	0.3034	0.6517
7	Logit	1734.132	1750.132	0.8963	0.4088	0.4344	0.5852	0.3052	0.6526
	Probit	1733.364	1749.364	0.8783	0.4208	0.4449	0.5906	0.2992	0.6496
	Cauchit	1736.665	1752.665	0.9594	0.3343	0.3672	0.5311	0.2937	0.6469
	Cloglog	1734.253	1750.253	0.8963	0.4071	0.4328	0.5837	0.3034	0.6517

Tabla 4.13: Métricas de la regresión logística funcional con distintos números de bspline y funciones de enlace - Menor GCV

4.5.5. Imputación clásica 4 Armónicos

Regresión logística funcional - Base de entrenamiento									
Bspl	Enlace	Devianza	AIC	Sensibilidad	Especificidad	Precisión	F1	Youden	AUC
4	Logit	1734.147	1744.147	0.8918	0.4091	0.4344	0.5843	0.3009	0.6505
	Probit	1733.352	1743.352	0.8738	0.4201	0.4439	0.5888	0.2939	0.647
	Cauchit	1736.748	1746.748	0.9504	0.3365	0.3688	0.5314	0.287	0.6435
	Cloglog	1734.272	1744.272	0.8918	0.4066	0.4321	0.5821	0.2984	0.6492
5	Logit	1734.091	1746.091	0.8963	0.4098	0.4354	0.5861	0.3062	0.6531
	Probit	1733.279	1745.279	0.8918	0.4203	0.4451	0.5938	0.3122	0.6561
	Cauchit	1736.729	1748.729	0.9549	0.3378	0.3702	0.5336	0.2927	0.6464
	Cloglog	1734.221	1746.221	0.9009	0.4081	0.434	0.5858	0.309	0.6545
6	Logit	1733.445	1747.445	0.8243	0.4243	0.4453	0.5783	0.2486	0.6243
	Probit	1732.586	1746.586	0.8108	0.4326	0.4524	0.5808	0.2434	0.6217
	Cauchit	1736.314	1750.314	0.9234	0.3435	0.3740	0.5324	0.267	0.6335
	Cloglog	1733.586	1747.586	0.8288	0.4226	0.4439	0.5782	0.2514	0.6257
7	Logit	1732.33	1748.33	0.7792	0.4361	0.4541	0.5738	0.2153	0.6077
	Probit	1731.413	1747.413	0.7612	0.4458	0.4624	0.5753	0.2071	0.6036
	Cauchit	1735.584	1751.584	0.8783	0.3538	0.3814	0.5318	0.2322	0.6161
	Cloglog	1732.487	1748.487	0.7837	0.4343	0.4527	0.5739	0.2181	0.6091
Fourier	Enlace	Devianza	AIC	Sensibilidad	Especificidad	Precisión	F1	Youden	AUC
4	Logit	1734.083	1746.083	0.8873	0.4111	0.4361	0.5848	0.2984	0.6492
	Probit	1733.281	1745.281	0.8693	0.4223	0.4458	0.5894	0.2917	0.6459
	Cauchit	1736.715	1748.715	0.9594	0.3375	0.3702	0.5343	0.297	0.6485
	Cloglog	1734.21	1746.21	0.8918	0.4098	0.4352	0.5849	0.3017	0.6509
5	Logit	1734.083	1746.083	0.8873	0.4111	0.4361	0.5848	0.2984	0.6492
	Probit	1733.281	1745.281	0.8693	0.4223	0.4458	0.5894	0.2917	0.6459
	Cauchit	1736.715	1748.715	0.9594	0.3375	0.3702	0.5343	0.297	0.6485
	Cloglog	1734.21	1746.21	0.8918	0.4098	0.4352	0.5849	0.3017	0.6509
6	Logit	1731.888	1747.888	0.7882	0.4291	0.4479	0.5713	0.2173	0.6087
	Probit	1730.824	1746.824	0.7657	0.4396	0.4567	0.5722	0.2053	0.6027
	Cauchit	1735.557	1751.557	0.9144	0.34	0.3702	0.5271	0.2544	0.6272
	Cloglog	1732.083	1748.083	0.7882	0.4263	0.4453	0.5691	0.2146	0.6073
7	Logit	1731.888	1747.888	0.7882	0.4291	0.4479	0.5713	0.2173	0.6087
	Probit	1730.824	1746.824	0.7657	0.4396	0.4567	0.5722	0.2053	0.6027
	Cauchit	1735.557	1751.557	0.9144	0.34	0.37029	0.5271	0.2544	0.6272
	Cloglog	1732.083	1748.083	0.7882	0.4263	0.4453	0.5691	0.2146	0.6073

Tabla 4.14: Métricas de la regresión logística funcional con distintos números de base y funciones de enlace - 4 armónicos

4.5.6. Imputación promedio Menor GCV

Se emplearán 2 transformaciones distintas: bases de B-spline y bases de Fourier. Estas representaciones son evaluadas mediante cuatro funciones de enlace diferentes: logit,

probit, cauchit y cloglog. El propósito es identificar el modelo que logre la clasificación más efectiva.

Regresión logística funcional - Base de entrenamiento									
Bspl	Enlace	Devianza	AIC	Sensibilidad	Especificidad	Precisión	F1	Youden	AUC
4	Logit	4382.1101	4392.1101	1	0.0065	0.0494	0.0941	0.0065	0.5032
	Probit	4381.8556	4391.8556	1	0.0077	0.0505	0.0962	0.0077	0.5038
	Cauchit	4382.9802	4392.9802	1	0.0037	0.0467	0.0893	0.0037	0.5018
	Cloglog	4382.1509	4392.1509	1	0.0062	0.0491	0.0937	0.0062	0.5031
5	Logit	4381.7009	4393.7009	0.9962	0.0862	0.1255	0.2229	0.0825	0.5412
	Probit	4381.4205	4393.4205	0.9924	0.1292	0.1665	0.2852	0.1217	0.5608
	Cauchit	4382.7197	4394.7197	1	0.0037	0.0467	0.0893	0.0037	0.5018
	Cloglog	4381.7395	4393.7395	0.9962	0.0797	0.1192	0.213	0.0759	0.5379
6	Logit	4381.5858	4395.5858	0.9887	0.1388	0.1338	0.2359	0.0911	0.5455
	Probit	4381.2919	4395.2919	0.9924	0.2296	0.1755	0.2981	0.1275	0.5637
	Cauchit	4382.6499	4396.6499	1	0.0038	0.0468	0.0894	0.0038	0.5019
	Cloglog	4381.6276	4395.6276	0.9962	0.0879	0.1271	0.2255	0.0841	0.542
7	Logit	4381.3829	4397.3829	0.9943	0.1067	0.145	0.2531	0.101	0.5505
	Probit	4381.0808	4397.0808	0.9868	0.1495	0.1856	0.3125	0.1363	0.5681
	Cauchit	4382.5032	4398.5032	1	0.0046	0.0476	0.0909	0.0046	0.5023
	Cloglog	4381.4246	4397.4246	0.9962	0.1011	0.1398	0.2452	0.0974	0.5487
Fourier	Enlace	Devianza	AIC	Sensibilidad	Especificidad	Precisión	F1	Youden	AUC
4	Logit	4342.76	4354.76	0.8809	0.2933	0.3187	0.468	0.1742	0.5871
	Probit	4340.428	4352.428	0.8544	0.333	0.3555	0.5021	0.1875	0.5938
	Cauchit	4351.669	4363.669	0.9905	0.0786	0.1179	0.2107	0.0691	0.5346
	Cloglog	4343.177	4355.177	0.8884	0.2852	0.3112	0.4609	0.1736	0.5868
5	Logit	4342.76	4354.76	0.8809	0.2933	0.3187	0.468	0.1742	0.5871
	Probit	4340.428	4352.428	0.8544	0.333	0.3555	0.5021	0.1875	0.5938
	Cauchit	4351.669	4363.669	0.9905	0.0786	0.1179	0.2107	0.0691	0.5346
	Cloglog	4343.177	4355.177	0.8884	0.2852	0.3112	0.4609	0.1736	0.5868
6	Logit	4308.016	4324.016	0.7296	0.4344	0.4471	0.5545	0.1641	0.5821
	Probit	4300.156	4316.156	0.7107	0.4628	0.4735	0.5683	0.1735	0.5868
	Cauchit	4339.528	4355.528	0.9678	0.1313	0.1674	0.2854	0.0992	0.5496
	Cloglog	4310.548	4326.548	0.7466	0.4218	0.4358	0.5504	0.1685	0.5843
7	Logit	4308.016	4324.016	0.7296	0.4344	0.4471	0.5545	0.1641	0.5821
	Probit	4300.156	4316.156	0.7107	0.4628	0.4735	0.5683	0.1735	0.5868
	Cauchit	4339.528	4355.528	0.9678	0.1313	0.1674	0.2854	0.0992	0.5496
	Cloglog	4310.548	4326.548	0.7466	0.4218	0.4358	0.5504	0.1685	0.5843

Tabla 4.15: Métricas de la regresión logística funcional con distintos números de base y funciones de enlace

4.5.7. Imputación promedio 4 armónicos

Regresión logística funcional									
Bspl	Enlace	Devianza	AIC	Sensibilidad	Especificidad	Precisión	F1	Youden	AUC
4	Logit	4347.983	4357.983	0.8431	0.2768	0.3012	0.4439	0.1199	0.56
	Probit	4346.643	4356.643	0.8298	0.3115	0.3338	0.4761	0.1413	0.5707
	Cauchit	4354.038	4364.038	0.9716	0.0737	0.1124	0.2016	0.0454	0.5227
	Cloglog	4348.247	4358.247	0.8487	0.2698	0.2948	0.4376	0.1186	0.5593
5	Logit	4334.478	4346.478	0.809	0.3286	0.3493	0.488	0.1377	0.5689
	Probit	4332.327	4344.327	0.8034	0.3504	0.3699	0.5066	0.1538	0.5769
	Cauchit	4347.097	4359.097	0.9489	0.1299	0.1652	0.2814	0.0788	0.5394
	Cloglog	4335.066	4347.066	0.8185	0.3212	0.3426	0.4831	0.1397	0.5699
6	Logit	4330.848	4344.848	0.792	0.3528	0.3717	0.506	0.1449	0.5725
	Probit	4327.781	4341.781	0.7769	0.3763	0.3936	0.5225	0.1532	0.5766
	Cauchit	4345.157	4359.157	0.93	0.1475	0.1812	0.3034	0.0775	0.5388
	Cloglog	4331.515	4345.515	0.7958	0.3468	0.3662	0.5016	0.1427	0.5714
7	Logit	4329.16	4345.16	0.7844	0.3582	0.3765	0.5088	0.1427	0.5714
	Probit	4325.749	4341.749	0.7731	0.3824	0.3992	0.5265	0.1555	0.5778
	Cauchit	4344.379	4360.379	0.9243	0.1523	0.1856	0.3091	0.0766	0.5383
	Cloglog	4329.888	4345.888	0.7863	0.3526	0.3713	0.5044	0.139	0.5695
Bspl	Enlace	Devianza	AIC	Sensibilidad	Especificidad	Precisión	F1	Youden	AUC
4	Logit	4342.76	4354.76	0.8809	0.2933	0.3187	0.468	0.1742	0.5871
	Probit	4340.428	4352.428	0.8544	0.333	0.3555	0.5021	0.1875	0.5938
	Cauchit	4351.669	4363.669	0.9905	0.0786	0.1179	0.2107	0.0691	0.5346
	Cloglog	4343.177	4355.177	0.8884	0.2852	0.3112	0.4609	0.1736	0.5868
5	Logit	4342.76	4354.76	0.8809	0.2933	0.3187	0.468	0.1742	0.5871
	Probit	4340.428	4352.428	0.8544	0.333	0.3555	0.5021	0.1875	0.5938
	Cauchit	4351.669	4363.669	0.9905	0.0786	0.1179	0.2107	0.0691	0.5346
	Cloglog	4343.177	4355.177	0.8884	0.2852	0.3112	0.4609	0.1736	0.5868
6	Logit	4308.016	4324.016	0.7296	0.4344	0.4471	0.5545	0.1641	0.5821
	Probit	4300.156	4316.156	0.7107	0.4628	0.4735	0.5683	0.1735	0.5868
	Cauchit	4339.528	4355.528	0.9678	0.1313	0.1674	0.2854	0.0992	0.5496
	Cloglog	4310.548	4326.548	0.7466	0.4218	0.4358	0.5504	0.1685	0.5843
7	Logit	4308.016	4324.016	0.72967	0.4344	0.4471	0.55452	0.1641	0.5821
	Probit	4300.156	4316.156	0.7107	0.4628	0.4735	0.5683	0.1735	0.5868
	Cauchit	4339.528	4355.528	0.9678	0.1313	0.1674	0.2854	0.0992	0.5496
	Cloglog	4310.548	4326.548	0.7466	0.4218	0.4358	0.5504	0.1685	0.5843

Tabla 4.16: Métricas de la regresión logística funcional con distintos números de base y funciones de enlace

4.5.8. Transformación fold Menor GCV

A continuación se presenta una tabla resumen de los resultados de la regresión logística funcional utilizando transformación bspline, utilizando de 4 a 7 bases con la función de

enlace logit, probit, cauchit y cloglog para la base tras imputación tras transformación Fold-Menor GCV:

Regresión logística funcional									
Bspl	Enlace	Devianza	AIC	Sensibilidad	Especificidad	Precisión	F1	Youden	AUC
4	Logit	4382.1101	4392.1101	1	0.0065	0.0494	0.0941	0.0065	0.5032
	Probit	4381.8556	4391.8556	1	0.0077	0.0505	0.0962	0.0077	0.5038
	Cauchit	4382.9802	4392.9802	1	0.0037	0.0467	0.0893	0.0037	0.5018
	Cloglog	4382.1509	4392.1509	1	0.0062	0.0491	0.0937	0.0062	0.5031
5	Logit	4381.7009	4393.7009	0.9962	0.0862	0.1255	0.2229	0.0825	0.5412
	Probit	4381.4205	4393.4205	0.9924	0.1292	0.1665	0.2852	0.1217	0.5608
	Cauchit	4382.7197	4394.7197	1	0.0037	0.0467	0.0893	0.0037	0.5018
	Cloglog	4381.7395	4393.7395	0.9962	0.0797	0.1192	0.213	0.0759	0.5379
6	Logit	4381.5858	4395.5858	0.9887	0.1388	0.1338	0.2359	0.0911	0.5455
	Probit	4381.2919	4395.2919	0.9924	0.2296	0.1755	0.2981	0.1275	0.5637
	Cauchit	4382.6499	4396.6499	1	0.0038	0.0468	0.0894	0.0038	0.5019
	Cloglog	4381.6276	4395.6276	0.9962	0.0879	0.1271	0.2255	0.0841	0.542
7	Logit	4381.3829	4397.3829	0.9943	0.1067	0.145	0.2531	0.101	0.5505
	Probit	4381.0808	4397.0808	0.9868	0.1495	0.1856	0.3125	0.1363	0.5681
	Cauchit	4382.5032	4398.5032	1	0.0046	0.0476	0.0909	0.0046	0.5023
	Cloglog	4381.4246	4397.4246	0.9962	0.1011	0.1398	0.2452	0.0974	0.5487
Bspl	Enlace	Devianza	AIC	Sensibilidad	Especificidad	Precisión	F1	Youden	AUC
4	Logit	4382.1101	4392.1101	0.9962	0.06615	0.1063	0.1921	0.0623	0.5311
	Probit	4381.8556	4391.8556	0.9924	0.1094	0.1476	0.2570	0.1019	0.5509
	Cauchit	4382.9802	4392.9802	1	0.0033	0.0464	0.0887	0.0033	0.5016
	Cloglog	4382.1509	4392.1509	0.9962	0.0597	0.1002	0.1821	0.0560	0.5280
5	Logit	4381.7009	4393.7009	0.9962	0.0661	0.1063	0.1921	0.0623	0.5311
	Probit	4381.4205	4393.4205	0.9924	0.1094	0.1476	0.2570	0.1019	0.5509
	Cauchit	4382.7197	4394.7197	1	0.0033	0.0464	0.0887	0.0033	0.5016
	Cloglog	4381.7395	4393.7395	0.9962	0.0597	0.1002	0.1821	0.0560	0.5280
6	Logit	4381.5858	4395.5858	0.9962	0.1026	0.1412	0.2473	0.0988	0.5494
	Probit	4381.2919	4395.2919	0.9887	0.1468	0.1831	0.3090	0.1355	0.5677
	Cauchit	4382.6499	4396.6499	1	0.0042	0.0472	0.0902	0.0042	0.5021
	Cloglog	4381.6276	4395.6276	0.9962	0.0964	0.1352	0.2382	0.0926	0.5463
7	Logit	4381.3829	4397.3829	0.9962	0.1026	0.1412	0.2473	0.0988	0.5494
	Probit	4381.0808	4397.0808	0.9887	0.1468	0.1831	0.3090	0.1355	0.5677
	Cauchit	4382.5032	4398.5032	1	0.004	0.0472	0.0902	0.0042	0.5021
	Cloglog	4381.4246	4397.4246	0.9962	0.996	0.1352	0.2382	0.0926	0.5463

Tabla 4.17: Métricas de la regresión logística funcional con distintos números de base y funciones de enlace

4.5.9. Transformación fold 4 armónicos

A continuación se presenta una tabla resumen de los resultados de la regresión logística funcional utilizando transformación bspline, utilizando de 4 a 7 bases con la función de enlace logit, probit, cauchit y cloglog para la base tras imputación tras transformación Fold-4 armónicos:

Regresión logística funcional									
Bspl	Enlace	Devianza	AIC	Sensibilidad	Especificidad	Precisión	F1	Youden	AUC
4	Logit	4379.1207	4389.1207	1	0.0365	0.0781	0.1449	0.0365	0.5182
	Probit	4378.1306	4388.1306	1	0.04	0.0814	0.1506	0.04	0.52
	Cauchit	4381.9830	4391.9830	1	0.0205	0.0628	0.1182	0.0205	0.5102
	Cloglog	4379.3066	4389.3066	1	0.0361	0.0777	0.1442	0.0361	0.5180
5	Logit	4378.0141	4390.0141	1	0.0488	0.0899	0.165	0.0488	0.5244
	Probit	4376.7383	4388.7383	0.9962	0.0856	0.1249	0.2220	0.0819	0.5409
	Cauchit	4381.5599	4393.5599	1	0.0191	0.0615	0.1159	0.0191	0.5095
	Cloglog	4378.2752	4390.2752	1	0.0423	0.0836	0.1544	0.0423	0.5211
6	Logit	4377.4132	4391.4132	0.9962	0.0757	0.1154	0.2069	0.0719	0.5359
	Probit	4376.0560	4390.0560	0.9962	0.1212	0.1590	0.2743	0.1175	0.5587
	Cauchit	4381.2690	4395.2690	1	0.0199	0.0622	0.1171	0.0199	0.5099
	Cloglog	4377.6747	4391.6747	0.9962	0.0678	0.1079	0.1947	0.0640	0.5320
7	Logit	4377.2071	4393.2071	0.9962	0.1083	0.1466	0.2556	0.1045	0.5522
	Probit	4375.8627	4391.8627	0.9906	0.1513	0.1875	0.3153	0.1419	0.5709
	Cauchit	4381.1178	4397.1178	1	0.0194	0.0617	0.1163	0.0194	0.5097
	Cloglog	4377.4588	4393.4588	0.9962	0.1015	0.1401	0.2457	0.0977	0.5488
Bspl	Enlace	Devianza	AIC	Sensibilidad	Especificidad	Precisión	F1	Youden	AUC
4	Logit	4378.2219	4390.2219	1	0.0379	0.0794	0.1472	0.0379	0.5189
	Probit	4376.9624	4388.9624	1	0.0418	0.0831	0.1535	0.0418	0.5209
	Cauchit	4381.6751	4393.6751	1	0.0201	0.0624	0.1176	0.0201	0.51
	Cloglog	4378.4804	4390.4804	1	0.0370	0.0786	0.1458	0.0370	0.5185
5	Logit	4378.2219	4390.2219	1	0.0379	0.0794	0.1472	0.0379	0.5189
	Probit	4376.9624	4388.9624	1	0.0418	0.0831	0.1535	0.0418	0.5209
	Cauchit	4381.6751	4393.6751	1	0.0201	0.0624	0.1176	0.0201	0.51
	Cloglog	4378.4804	4390.4804	1	0.0370	0.0786	0.1458	0.0370	0.5185
6	Logit	4377.3353	4393.3353	0.9962	0.1032	0.1417	0.2482	0.0994	0.5497
	Probit	4376.0085	4392.0085	0.9924	0.1452	0.1818	0.3074	0.1377	0.5688
	Cauchit	4381.1808	4397.1808	0.0609	0.0185	0.0609	0.1148	0.0185	0.5092
	Cloglog	4377.5838	4393.5838	0.9962	0.0964	0.1352	0.2382	0.0926	0.5463
7	Logit	4377.3353	4393.3353	0.9962	0.1032	0.1417	0.2482	0.0994	0.5497
	Probit	4376.0085	4392.0085	0.9924	0.1452	0.1818	0.3074	0.1377	0.5688
	Cauchit	4381.1808	4397.1808	1	0.0185	0.0609	0.1148	0.0185	0.5092
	Cloglog	4377.5838	4393.5838	0.9962	0.0964	0.1352	0.2382	0.0926	0.5463

Tabla 4.18: Métricas de la regresión logística funcional con distintos números de bspline y funciones de enlace