ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

MODELAMIENTO POR COMPONENTES PRINCIPALES CON EL ENFOQUE DEL ANÁLISIS DE DATOS FUNCIONALES PARA PREDECIR EL RIESGO PAÍS DEL ECUADOR A TRAVÉS DEL INDICADOR DE BONOS DE MERCADOS EMERGENTES.

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

PROYECTO DE INVESTIGACIÓN

ALEXANDER GABRIEL CONSTANTE CACHIMUEL
alexandergcc98@outlook.com

Director: MIGUEL ALFONSO FLORES SÁNCHEZ, PH.D.
miguel.flores@epn.edu.ec

QUITO, FEBRERO 2022

DECLARACIÓN

Yo, ALEXANDER GABRIEL CONSTANTE CACHIMUEL, declaro bajo juramento que el trabajo aquí escrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración, cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su reglamento y por la normatividad institucional vigente.

Alexander Gabriel Constante Cachimuel

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por ALEXANDER GABRIEL CONSTANTE CACHIMUEL, bajo mi supervisión.

Miguel Alfonso Flores Sánchez, Ph.D.
Director del Proyecto

AGRADECIMIENTOS

Agradezco infinitamente a mi mamá, a mis hermanos y a mis abuelitas por estar pendientes de mí durante esta satisfactoria y, a la vez, complicada travesía académica, pues hasta este punto en el que me encuentro, no hubiese sido posible sin sus consejos, regaños, preocupaciones y bendiciones, para poder superarme profesionalmente. Los estimo mucho, ya que han sido y serán la fuente primordial para no rendirme y motivarme día a día.

A cada uno de los profesores que han compartido sus conocimientos y experiencias frente a las distintas áreas que tiene la Matemática; en especial, agradezco a mi tutor Miguel Flores, por su confianza y guía profesional, ya que con su apoyo y tiempo, he logrado desarrollar el presente proyecto de investigación.

A Gabriel, Diego, Erick, Esteban, Paúl, Dany y Vivi, por su buena amistad y ayuda, y haber plasmado agradables recuerdos durante esta etapa; y a todos los demás, con quienes de alguna manera he compartido y me han brindado su amistad y apoyo.

DEDICATORIA

A mi querida mamá, Marisol, por su enorme esfuerzo y dedicación para verme prosperar en la vida. La persona más guerrera que he conocido.

A mi abuelita Esther, por su gran apoyo y sabios consejos.

A mis hermanos: Blady, Santy y Ali, por su paciencia y compañía.

A mi cuñada Daysi, por su aprecio.

A mis sobrinos Thiago y Thadeo, por sus ocurrencias y cariño.

Índice general

Notaciones IX

Índice de figuras XI

Índice de tablas XIII

Resumen XIV

Abstract XV

1. Introducción **1**

1.1. Antecedentes 1

1.2. Justificación 3

1.3. Objetivos 4

1.3.1. General 4

1.3.2. Específicos 4

2. El Riesgo País **5**

2.1. Inicios de su importancia 5

2.2. Definición 6

2.2.1. ¿Cómo se mide el riesgo país? 6

2.2.2. EMBI 7

2.2.3. Cálculo del Indicador 8

2.3. Riesgo País del Ecuador 8

2.3.1. 2000 - 2009 9

2.3.2. 2010 - 2020 10
2.4. Enfoque del proyecto 11

3. Análisis de Datos Funcionales (FDA) 13
3.1. Datos Funcionales 13
3.2. Ajuste de datos discretos a funciones suaves 15
3.2.1. Bases de Fourier 16
3.2.2. Bases B-spline 17
3.2.3. Procedimiento de ajuste 19
3.2.4. Penalización por rugosidad 20
3.3. Funciones descriptivas 21
3.4. Detección de datos funcionales atípicos 22
3.4.1. Medidas de profundidad 23
3.4.2. Media recortada 25
3.4.3. Algoritmo de detección 25
3.5. Análisis de Componentes Principales Funcionales (FPCA) 26
3.5.1. Formulación del problema 27
3.5.2. Análisis de autofunciones 30
3.5.3. Método computacional 31
3.5.4. Interpretación 32
3.5.5. PCA vs FPCA 33
3.6. Modelo de predicción basado en FPCA 34
3.6.1. Modelamiento 34
3.6.2. Pronósticos 35
3.6.3. Suavizado exponencial 36
3.6.4. Bandas de Confianza Bootstrap 38
3.6.5. Nueva propuesta para pronosticar dos escenarios 39

4. Metodología FDA aplicado a datos históricos del indicador del Riesgo País del Ecuador 42
4.1. Descripción de la base de datos 42

4.1.1. Preprocesamiento 43
4.2. Representación a datos funcionales 45
4.2.1. Ajuste con base B-spline 46
4.3. Análisis exploratorio funcional 50
4.3.1. Área de las curvas 51
4.3.2. Primera y Segunda derivada 52
4.3.3. Detección de datos funcionales atípicos 54
4.3.4. Medidas estadísticas funcionales 58
4.4. Aplicación de la técnica FPCA 60
4.5. Desarrollo del modelo basado en FPCA bajo dos enfoques 66
4.5.1. Ajuste del modelo y verificación de supuestos 67
4.5.2. Estudio del mejor caso para el EMBI 2021 69
4.5.3. Estudio del peor caso para el EMBI 2021 74

5. Conclusiones y Recomendaciones **77**

Referencias **81**

A. Teoría complementaria **85**
A.1. Producto interno 85
A.2. Análisis de Componentes Principales (PCA) 87
A.2.1. Cálculo de las componentes 88
A.2.2. Elección del número de componentes 90

B. Implementación en el software `R` **91**
B.1. Análisis previo 91
B.2. Aplicación FDA 93

Notaciones

n	Número de datos funcionales
m	Número de puntos discretos de los datos
K	Número de funciones base
p	Número de componentes principales funcionales
$\mathcal{X}$	Variable aleatoria funcional $(v.f.)$
$X_i(\cdot)$	$i-$ésimo dato funcional
$\mathcal{T}$	Dominio de los datos funcionales
$L^2(\mathcal{T})$	Espacio de las funciones cuadrado integrables en $\mathcal{T}$
$\{\phi_k\}_{k\in\mathbb{N}}$	Conjunto de funciones base
$\boldsymbol{X}(\cdot)$	Vector de los n datos funcionales
$\boldsymbol{C}$	Matriz de coeficientes
SSE	Suma de errores cuadrados
λ	Parámetro de suavizamiento
$PENSSE$	Suma penalizada de errores cuadrados
R	Matriz de Penalización
S_λ	Matriz de suavizamiento
$\widehat{\mathcal{X}}(\cdot)$	Estimador de la media funcional de la $v.f.\,\mathcal{X}$
$s^2_{\mathcal{X}}(\cdot)$	Estimador de la varianza funcional de la $v.f.\,\mathcal{X}$
$v_{\mathcal{X}}(\cdot,\cdot)$	Estimador de la covarianza funcional de la $v.f.\,\mathcal{X}$
$\rho_{\mathcal{X}}(\cdot,\cdot)$	Estimador de la correlación funcional de la $v.f.\,\mathcal{X}$
FMD_{X_i}	Profundidad de Fraiman & Muniz del dato i
MD_{X_i}	Profundidad Modal del dato i
$\widehat{\mu}_{\alpha,TM}$	Media funcional $\alpha-$recortada
Θ	Número de muestras bootstrap
γ	Parámetro de bootstrap suavizado
$\mathcal{V}$	Operador de covarianza
$\alpha_j(\cdot)$	$j-$ésima componente principal funcional

λ_j	$j-$ésimo autovalor de $\mathcal{V}$
$f_{j,i}$	Puntuación del dato i en la componente j
HU	Método de Hyndman y Ullah
$\widehat{x}_{n+h\|n}$	Pronóstico puntual de x_{n+h}
α	Parámetro del suavizado exponencial
$X_{\cdot}^{+}$	Pronóstico de una función en el mejor caso
$X_{\cdot}^{-}$	Pronóstico de una función en el peor caso

Índice de figuras

2.1. Evolución del EMBI ecuatoriano por año 9

3.1. Funciones pertenecientes a una base de Fourier 17
3.2. Funciones pertenecientes a una base B-splines 19
3.3. Comportamiento exponencial para $\alpha = 0{,}5$ 37

4.1. Valores perdidos EMBI 44
4.2. Evolución del EMBI anual 45
4.3. Valores GCV según K y λ dados 47
4.5. Datos Funcionales del EMBI por año 51
4.6. Áreas aproximadas de los datos funcionales 52
4.7. Primera derivada de los datos 54
4.8. Segunda derivada de los datos 55
4.9. Profundidad Fraiman & Muniz 56
4.10. Profundidad Modal 56
4.11. Detección de datos funcionales atípicos 57
4.12. Media funcional, $\mu_{\mathcal{X}}(t)$ 58
4.13. Desviación Estándar funcional, $s_{\mathcal{X}}(t)$ 59
4.14. Covarianza - Correlación 59
4.15. Varianza de las FPC 60
4.16. Componentes Principales Funcionales 62
4.17. Interpretación de la primera FPC 64
4.18. Interpretación de la segunda FPC 65
4.19. Interpretación de la tercera FPC 65

4.20. Interpretación de la cuarta FPC 66
4.21. p-valores del test de rachas 67
4.22. p-valores del test de Ljung-Box 68
4.23. Media del error $\varepsilon_i(t)$, para cada t 69
4.24. Serie de las puntuaciones de los datos $X_i(t)$ para el mejor caso 70
4.25. Pronóstico de la función de las puntuaciones EMBI 2021 del mejor caso 71
4.26. Evolución de las puntuaciones EMBI en el 2021 hasta el 05 de agosto 72
4.27. Comparación de los pronósticos del mejor caso con los datos actuales del 2021 72
4.28. Banda de confianza Bootstrap para la función del mejor caso 2021_t^+, con un nivel de confianza del 90 % 73
4.29. Serie de las puntuaciones de los datos $X_i(t)$ para el peor caso 74
4.30. Pronóstico de la función de las puntuaciones EMBI 2021 del peor caso 75
4.31. Comparación de los pronósticos del peor caso con los datos del 2021 75
4.32. Banda de confianza Bootstrap para la función del peor caso 2021_t^-, con un nivel de confianza del 90 % 76

Índice de tablas

2.1. Promedio mensual del EMBI ecuatoriano 10

3.1. Notaciones del PCA y FPCA 34

4.1. Extracto de la información recuperada del BCE 43
4.2. Extracto de la base de datos ordenada 43
4.3. Valores perdidos por año, mes y día 44
4.4. Conjuntos de valores K y λ 46
4.5. Comparación de las profundidades FM y MD 57
4.6. Profundidad MD de cada dato funcional 57
4.7. Variabilidad explicada por cada FPC 61
4.8. Puntuaciones de los datos funcionales en cada FPC 61
4.9. Medidas de bondad de ajuste del modelo 69

Resumen

En el presente proyecto se investiga sobre la aplicación del Análisis de Datos Funcionales (FDA), una técnica relativamente nueva en el campo de la Estadística, abordando desde conceptos iniciales como: variable aleatoria funcional, dato funcional, ajuste de datos discretos por funciones suaves y medidas estadísticas de centralidad y variabilidad, para luego estudiar los enfoques funcionales de medidas de profundidad, adoptando el concepto de curva más central, y el método de componentes principales funcionales, útil para la elaboración de un modelo el cual pronostique una función completa. Esta técnica se la aplica a los datos diarios del riesgo país ecuatoriano para obtener el comportamiento futuro de las puntuaciones diarias para el año 2021; el riesgo país es un referente para atraer inversores y contribuir con el desarrollo de un país, determinado por el Indicador de Bonos de Mercados Emergentes (EMBI). Además, para el modelo, se proponen dos escenarios de pronóstico, basados en el mejor y peor caso y, así, comprender la tendencia futura del indicador en ambos escenarios y brindar soluciones frente a los impactos más graves que puede causar dicho indicador.

Palabras claves: FDA, componentes principales funcionales, riesgo país, EMBI.

Abstract

This project investigates the application of Functional Data Analysis (FDA), a relatively new technique in the field of Statistics, addressing from initial concepts such as: functional random variable, functional data, adjustment of discrete data from smooth functions and statistical measures of centrality and variability, to then study the functional approaches of depth measurements, adopting the concept of the most central curve, and the method of functional principal components, useful for the elaboration of a model which predicts a complete function. This technique is applied to the daily data of the ecuadorian country risk to obtain the future behavior of the daily scores for the year 2021; country risk is a benchmark for attracting investors and contributing to the development of a country, determined by the Emerging Markets Bond Index (EMBI). In addition, for the model, two forecasting scenarios are proposed, based on the best and worst case and, thus, understanding the future trend of the indicator in both scenarios and providing solutions to the most serious impacts that this indicator may cause.

Keywords: FDA, functional principal components, country risk, EMBI.

Capítulo 1

Introducción

1.1. Antecedentes

No es una tarea fácil reconocer con exactitud la economía de un país, en particular la del Ecuador, debido a los distintos factores que se involucran en la misma tales como: salud, educación, producción, exportación, pago de deudas externas, entre otros. Sin embargo, existen varios indicadores que perfilan la economía ecuatoriana, entre ellos: la inversión extranjera directa (IED), el cual ayuda a financiar las economías por su contribución a los sectores productivos (Calahorrano et al. [2020]); los precios internacionales de petróleo, que ha sido la principal fuente de ingresos del Ecuador, según los balances comerciales[1] del Banco Central; el PIB (producto interno bruto), considerado como un indicador que mide la evolución de la economía de un país (BCE [2021], Indicadores económicos - PIB nominal); o, uno de los indicadores poco conocidos y más importantes, el *riesgo país*, catalogado como el termómetro de la salud de una economía. Según Lapitz et al. [2005], este indicador mide la perspectiva de los mercados internacionales frente al cumplimiento de las obligaciones del país, acarreando impactos económicos, políticos y sociales como:

- el nivel de incertidumbre para conceder algún financiamiento al país;
- la capacidad del país en los pagos de un crédito;
- el aumento de las tasas de interés.

Así, cuanto más bajo este indicador, más confianza tendrán los inversionistas.

[1] https://www.bce.fin.ec/index.php/component/k2/item/299-evoluci%C3%B3n-de-la-balanza-comercial

El riesgo país se lo calcula principalmente por el Indicador de Bonos de Mercados Emergentes (EMBI), y a través del mismo se han realizado varios estudios, para por ejemplo, determinar variables macroeconómicas que influyen en dicho índice. Carrillo et al. [2021], concluyeron que indicadores como el WTI (precios del petróleo) y VIX (volatilidad del mercado financiero) tienen un efecto importante en el EMBI ecuatoriano, al igual que los EMBI's de Chile y Argentina; a partir de un modelo SVAR. A nivel de predicción, se puede mencionar el trabajo sobre el EMBI argentino de Matsumoto [2019], que al utilizar un modelo ARIMA, obtuvo dificultades en las predicciones por los cambios abruptos en los datos, de donde concluye que al realizar modelos más complejos se podrían controlar varios factores para las predicciones.

Por otra parte, aunque modelos con el enfoque multivariante han sido y son muy útiles para analizar y predecir información de la más variada índole, a finales de la década de los noventa se comienza a popularizar la técnica del Análisis de Datos Funcionales (FDA), teniendo como referencia principal la de Ramsay y Silverman [1997], para luego disponer de otros excelentes aportes como Ferraty y Vieu [2006] y Kokoszka y Reimherr [2017]. Así, con esta técnica, a los datos se los observan como funciones o curvas, para explorar la fluidez de los procesos que describen a los datos discretos, además que varios métodos multivariantes tienen su versión funcional. Ullah y Finch [2013] resumen una serie de artículos, tanto prácticos como teóricos que han utilizado el FDA, habiendo más énfasis en los métodos de: suavizamiento de datos, componentes principales funcionales, clustering, modelos de regresión funcional (FLM) y predicciones.

Con esto, se suele decir que el FDA es el análisis estadístico sobre muestras de curvas, cuyos objetivos son esencialmente similares al de otra área de la Estadística. Es decir, se realiza un análisis exploratorio, como por ejemplo, detectar datos funcionales atípicos siguiendo lo expuesto en Fraiman y Muniz [2001] y, posteriormente, realizar un estudio predictivo. Es así que el FDA se ha vuelto cada vez más frecuente en campos como: medicina, economía, climatología, finanzas, análisis de imágenes, lingüística, meteorología, biomecánica, entre otras; debido a la importancia de comprender las tendencias en los datos definidos en el tiempo o la frecuencia, que a su vez estos pueden estar en intervalos muy cortos y no equiespaciados, y lograr hallar información sobre tendencias futuras para la toma de decisiones, empleando por ejemplo, el modelo propuesto por Hyndman y Ullah [2007].

1.2. Justificación

Con el fin de obtener buenos pronósticos para el indicador del riesgo país diario del Ecuador y por la escasa bibliografía en cuanto a modelos de predicción del mismo, resulta interesante y adecuado utilizar esta novedosa técnica a dichos datos, para además, tener una mejor comprensión del comportamiento del EMBI por cada año, es decir, representar a cada EMBI anual por medio de una función y poder captar los patrones y variaciones que se presenten. El modelo a desarrollar se basará en el método de las componentes principales funcionales, las cuales recogen los modos de variación de los datos, propuesto por Hyndman y Ullah [2007] para predecir datos demográficos. Aún así, como los autores lo mencionan, este enfoque proporciona un marco de modelado que se adapte fácilmente para permitir otra información.

Este modelo se lo ha aplicado en: Hyndman y Shang [2009] sobre tasas de mortalidad femeninas francesas específicas por edad de 1816 a 2006 y en tasas de fecundidad australianas específicas por edad de 1921 a 2006; en Shang et al. [2010] sobre datos de mortalidad humana de catorce países desarrollados, donde además se presenta una mejora en la precisión de las predicciones, comparado con modelos multivariantes como el de Lee y Carter de 1992; en Andreozzi [2017] sobre datos de mortalidad de la Ciudad de Rosario en Argentina de 1980 a 2015, habiendo de igual forma mejoras en las predicciones; y, en varios documentos relacionados a demografía.

No obstante, una de las estructuras que puede haber en el modelo, consiste en asignar una mayor ponderación a los datos más recientes; por tanto, adaptando este modelo a los datos del indicador del riesgo país, en el presente proyecto se proponen dos escenarios de pronóstico, uno donde se prediga un comportamiento normal del indicador del riesgo país y otro donde se prediga un comportamiento en el peor de los casos, según el histórico a utilizar.

Además, gracias a los avances informáticos, el análisis y tratamiento de datos funcionales se vuelve una tarea más óptima, puesto que, en particular, en el software estadístico `R` se dispone de una gran cantidad de funciones dentro de los paquetes `fda` (Ramsay et al. [2009]) y `fda.usc` (Febrero-Bande y de la Fuente [2012]), para tratar desde el suavizamiento de los datos hasta la elaboración del modelo. Finalmente, de los resultados obtenidos con los datos del EMBI, se podrían brindar medidas interesantes para tener niveles más bajos del indicador o buscar soluciones para no llegar a puntuaciones muy altas, y que sean un referente para decisiones políticas o económicas futuras. De otro lado, al estudiarse una técnica poco conocida de la

Estadística, lo expuesto en el proyecto podría ser un referente para el entorno académico.

Así, el presente proyecto se lo estructura de la siguiente manera. En el Capítulo 2 se presenta la descripción del riesgo país, desde los inicios de su importancia hasta una breve reseña histórica del comportamiento del índice en el Ecuador. En el Capítulo 3 se expone la metodología para desarrollar un estudio exploratorio y un modelo bajo dos enfoques sobre un conjunto de datos, a través del Análisis de Datos Funcionales. En el Capítulo 4 se muestra la aplicación de la técnica a los datos diarios por año del índice del riesgo país. En el Capítulo 5 se exponen conclusiones y recomendaciones. Finalmente, en los Apéndices se incluye teoría complementaria en el desarrollo del proyecto y el código realizado en el software `R`.

1.3. Objetivos

1.3.1. General

Comprender el comportamiento del riesgo país del Ecuador y predecir el mismo bajo el mejor y peor escenario por medio del Indicador de Bonos de Mercados Emergentes (EMBI), aplicando un estudio exploratorio y la técnica de componentes principales desde el enfoque del análisis de datos funcionales, para obtener mejores decisiones frente a los impactos que dicho indicador conlleva.

1.3.2. Específicos

- Identificar los efectos de ascenso y descenso presentados en el indicador del riesgo país, relacionados con la economía del Ecuador.
- Ajustar los datos discretos de las puntuaciones del EMBI por medio de curvas para comprender el comportamiento de los datos.
- Detectar datos funcionales atípicos por medio de las medidas de profundidad para reducir la influencia anormal y determinar las medidas estadísticas funcionales básicas.
- Crear y validar el modelo a través de las componentes principales funcionales, para predecir las puntuaciones del EMBI del año 2021 con los enfoques del mejor y peor caso.

Capítulo 2

El Riesgo País

2.1. Inicios de su importancia

Los inicios sobre la importancia del riesgo país se presenta en Gonzalez y Minguez [2005], donde se indica que en los primeros años de la década de 1960, la Organización para la Cooperación y Desarrollo Económicos (OCDE)[1] tenía una deuda externa total de aproximadamente cien millones de dólares, esto debido a que las tasas de interés estaban por debajo de la tasa de crecimiento de las exportaciones. Por tanto, los prestamistas, que en su mayoría eran parte del sistema bancario, no estaban conscientes del riesgo asociado a sus préstamos y para agosto de 1982, el análisis y estimación del riesgo país se lo consideró como uno de los problemas importantes dentro del sistema financiero internacional, pues México se había declarado imposibilitado de saldar la deuda externa de alrededor de ochenta mil millones de dólares. Es así que la evaluación del mismo es el mayor reto que los agentes afrontan al momento de realizar sus inversiones.

Además, Clark [2017] explica que la literatura sobre el riesgo país comenzó a ser ampliamente estudiada entre las décadas de los sesenta y setenta. Estos estudios se centraron en aspectos políticos, concretamente, en el riesgo de que los flujos de efectivo de una empresa que emplea una inversión internacional, se vean afectados de forma negativa por el accionar de un gobierno extranjero. Es decir, hubo preocupación por la influencia de gobiernos en empresas que realizaban negocios internacionales, pues en ese entonces, varios territorios comenzaron a cuestionarse de la importancia de que firmas poderosas operen en el lugar. Durante los años setenta

[1] Organismo fundado en 1961, compuesto por 38 países. Se basa en una cooperación internacional en cuanto a políticas sociales y económicas.

y ochenta, por crisis de deudas externas en varios países en desarrollo, gran parte de la literatura se trató sobre evaluar soluciones. Finalmente, durante las décadas de 1990 y 2000, varios estudios se basaron en las crisis bancaria y monetaria, como por ejemplo la crisis financiera de Ecuador entre 1998 y 1999, y la crisis hipotecaria de alto riesgo de 2008 en Estados Unidos, la cual afectó a gran parte del mundo. Por tanto, todo lo anterior fue estudiado y se sigue estudiando, dado que las crisis políticas, de deuda y financieras se han involucrado en el análisis del riesgo país.

2.2. Definición

El riesgo país se lo puede definir directamente como lo citado en Calahorrano et al. [2020], pues se indica que determina la predisposición de un Estado frente a sus obligaciones, orientado a preservar la rentabilidad de los inversionistas. También se lo relaciona como una medición de la capacidad de un Estado para cumplir con los pagos de sus intereses y de créditos al tiempo de su vencimiento.

Las inversiones son de suma importancia para el desarrollo y crecimiento de forma global de cada país, por ello, el riesgo país juega un papel crucial en esta proyección, pues estas solo llegarán al Estado en la medida que este indicador tenga un nivel bajo, es decir, cuanto más bajo sea, las entradas de inversiones serán mayor. Visto de otro modo, Lapitz et al. [2005] señalan que el riesgo país es el grado de incertidumbre para que a un país se le otorgue un préstamo; además, en la jerga de finanzas, a este indicador se lo considera también como el nivel de riesgo que tiene un inversionista en no poder recuperar las inversiones hechas en un país.

2.2.1. ¿Cómo se mide el riesgo país?

En la página oficial del Banco Central del Ecuador (BCE [2021]) se tiene que el riesgo país es un concepto económico que se lo ha estudiado tanto de forma académica como empírica a través de varias metodologías, desde sistemas que incorporan variables políticas, financieras y económicas hasta el uso de índices de mercado. Así, en Lapitz et al. [2005], se explica que existen diferentes empresas calificadores de riesgo (ECR) que brindan evaluaciones de este indicador, estimando en forma de probabilidad el incumplimiento de crédito o inversión internacional de las operaciones de un país, y poder compararlo con otros países. Por ejemplo, empresas como Institutional Investor o Euromoney, determinan niveles de riesgo por país en base al

análisis de varios factores cualitativos ponderados, como el marco macroeconómico en el que el emisor se encuentra. Por otra parte, Fitch-IBCA, Moody's o Standard & Poor's, empresas calificadoras de riesgo crediticio, establecen calificaciones a las emisiones de deuda pública para esta evaluación. De la misma referencia, se encuentran algunas metodologías de calificación del riesgo país, considerando un conjunto de criterios cuantitativos y cualitativos. Sin embargo, generalmente se utiliza un indicador de mercado que considera el rendimiento de deuda de un país, que se lo explica en la siguiente subsección.

2.2.2. EMBI

El Indicador de Bonos de Mercados Emergentes (EMBI) se utiliza generalmente para medir el riesgo país; el mismo fue introducido y calculado en 1990 por la consultora internacional JPMorgan Chase & Co.[2], la cual califica los precios y el movimiento de los países emergentes, en cuanto a títulos de deuda y bonos. Se basa en la tendencia de la deuda externa de cada país y, mientras una baja certeza de que el país asumirá con sus obligaciones, más alto será el EMBI del país y recíprocamente, Lazo [2021].

Bonos

Un bono es un título que representa el derecho a recibir pagos en períodos de tiempo futuros, denominados *flujos*, a cambio de dar un monto de dinero al instante de adquirir el mismo. Este título puede ser emitido por un gobierno local, municipio o empresa, con la intención de endeudarse; en el mismo se aclara: el monto (valor nominal del bono), llamado *principal*, el tipo de interés (*cupón*) ya sea fijo o variable, el plazo (periodo de vigencia), el tipo de moneda y la fecha de vencimiento. Visto de otra forma, un bono es un préstamo de dinero que realiza un inversionista a una empresa o país; también se lo denomina título de deuda.

Bonos del Tesoro de los Estados Unidos

También conocidos como T-bonds, son emitidos por el gobierno federal, vistos como las inversiones más confiables alrededor del mundo, pues gozan del respaldo

[2]Empresa de servicios financieros líder a nivel mundial con activos por más de dos billones de dólares. Líder en servicios financieros para consumidores y pequeñas empresas, en banca de inversión, en gestión de activos, etc.

y crédito del gobierno estadounidense, ya que ante cualquier situación grave, como inflación o guerras, a quienes invirtieron en el bono, el gobierno cumplirá con sus titulares.

2.2.3. Cálculo del Indicador

Al EMBI se lo relaciona como un margen de rentabilidad implícito en los Bonos del Tesoro de los EE.UU. por lo cual, el indicador es la resta, conocida como *spread*, entre las tasas de interés que costean los bonos definidos en dólares (emitidos por el país) y los Bonos del Tesoro de EE.UU. que son considerados "libres" de riesgo o que tienen menor riesgo en el mercado, Lapitz et al. [2005].

Este indicador se lo representa en puntos básicos (pb), donde cada 100 puntos expresan un punto porcentual. Por ejemplo, supongamos que los Bonos emitidos por un Estado están pagando un interés del 18 % y los Bonos del Tesoro están pagando un interés del 3 %, entonces se tiene que la puntuación del riesgo país es de

$$(18 - 3) \times 100 = 1500 \text{ puntos},$$

lo que quiere decir que el Bono del país que lo emitió, paga un 15 % más sobre la tasa del bono norteamericano

De lo anterior, se tiene que la suma entre la tasa del Bono del Tesoro y el EMBI del país determinado (en porcentaje), indica que es la mínima tasa que exigirá el inversionista para negociar en el país.

2.3. Riesgo País del Ecuador

En primer lugar, la evolución diaria del indicador del riesgo país (EMBI), desde el año 2000 hasta el año 2020, se muestra en la Figura 2.1, de donde claramente se pueden ver las diferencias en el comportamiento del indicador.

Por otra parte, en la Tabla 2.1, se encuentran las puntuaciones mensuales promedio del EMBI, con el fin de establecer una mejor idea del comportamiento del riesgo país en el Ecuador desde el 2000 al 2020. En la misma tabla, se ha señalado de color rojo a las puntuaciones que han superado los 3000 pb, de color azul a las puntuaciones de 1000 a 3000 pb, de color amarillo las que están entre 600 y 1000 pb, y las celdas no pintadas son las puntuaciones menores a 600 pb. Con esto, a continuación, se expone una pequeña reseña histórica de este indicador ecuatoriano.

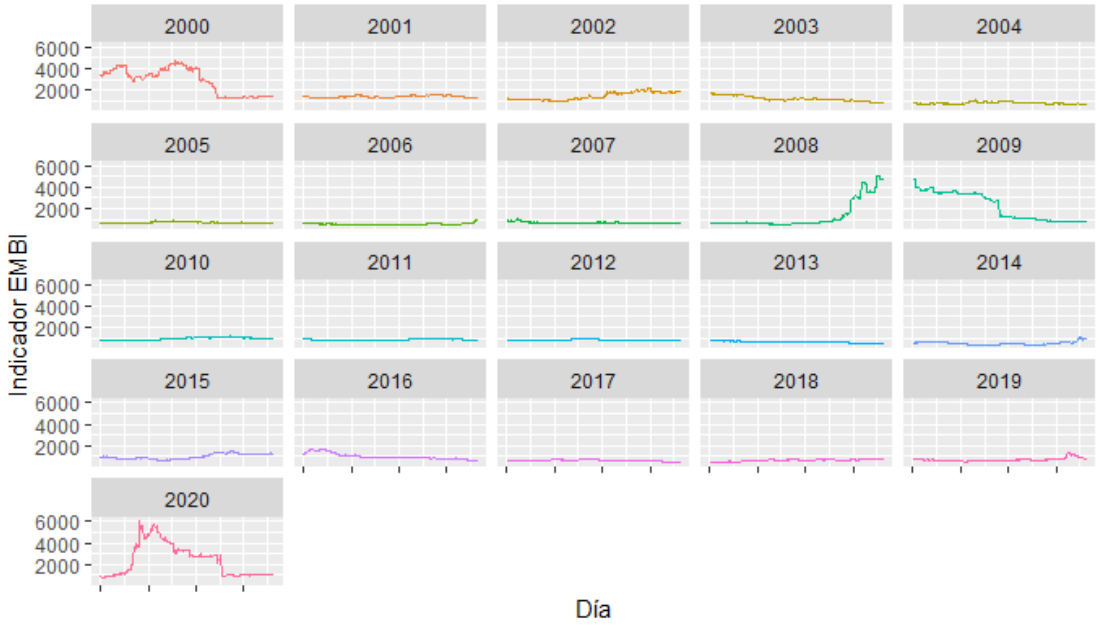


Figura 2.1: Evolución del EMBI ecuatoriano por año
Elaboración: Autor

2.3.1. 2000 - 2009

En el reporte de El Universo [2006], se indica que para ese entonces el índice del riesgo país más alto fue el registrado el 05 de junio del año 2000, con 4712 pb; esto debido a que el Ecuador atravesaba la crisis financiera a raíz de la dolarización. Para el año 2001 se logró financiar un 35 % del presupuesto nacional gracias a la exportación de crudo y hasta abril de 2002, se llega a un récord de 1018 pb, El Universo [2002]. Aunque, para el mes de mayo, hay una tendencia creciente a consecuencia de inestabilidades políticas y económicas, falta de acuerdos con el Fondo Monetario Internacional (FMI) y la época electoral, llegando a una puntuación alta de 2200 pb, en octubre.

Luego, en el año 2003, según el reporte de El Universo [2003], las puntuaciones aumentaron por atrasos en el envío y aprobación de reformas laborales y tributarias, y la ruptura con Pachakutik, visto en ese entonces como algo negativo ya que ocasionarían protestas. A partir del 2004 al 2007, se observa un comportamiento "normal" de las puntuaciones, llegando a 457 pb en mayo de 2006. Para el último trimestre del año 2008, se registra un riesgo país histórico de 5055 pb, superando lo del año 2000; este repunte ocurrió el 15 de diciembre, día en el que se venció el plazo del periodo de gracia para la liquidación de intereses de los bonos Global[3] 12 y cuando se

[3]Constituyeron un nuevo planteamiento para reestructurar las deudas ante la presencia de nuevas crisis en países de Latinoamérica a finales de los noventa e inicios de los 2000

	Ene.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ago.	Sep.	Oct.	Nov.	Dic.
2000	3672	4082	3117	3365	4007	4429	3790	2593	1296	1299	1350	1451
2001	1324	1244	1328	1474	1357	1306	1383	1473	1478	1541	1413	1260
2002	1145	1153	1086	998	1116	1314	1378	1742	1775	2005	1782	1797
2003	1619	1528	1449	1228	1109	1108	1172	1167	1105	1031	941	841
2004	731	747	723	753	920	884	921	835	804	731	722	711
2005	683	638	633	737	836	820	761	714	686	680	651	646
2006	624	568	561	513	488	517	504	494	576	564	525	740
2007	905	746	695	613	626	629	712	702	660	599	617	605
2008	633	642	669	601	562	569	662	698	878	1928	3697	4266
2009	3954	3627	3535	3414	3332	2734	1272	1132	1024	856	822	820
2010	761	813	822	816	900	963	1014	1037	1032	1016	996	928
2011	869	761	781	762	788	801	783	866	881	887	868	842
2012	808	790	803	807	850	892	856	805	753	766	815	827
2013	731	704	700	678	638	643	644	630	644	568	530	531
2014	567	610	551	453	355	370	421	397	360	415	558	824
2015	934	778	800	783	671	777	900	1164	1342	1334	1213	1235
2016	1518	1575	1187	1077	911	893	880	865	857	773	764	671
2017	612	609	614	705	665	711	689	641	629	573	541	470
2018	452	494	546	594	685	712	653	679	689	672	746	766
2019	712	659	612	552	573	590	578	702	648	763	979	938
2020	859	1170	3580	5033	4305	3287	2862	2780	1413	983	1028	1041

Tabla 2.1: Promedio mensual del EMBI ecuatoriano

debía pagar intereses de los Global 30. En general, el aumento de las puntuaciones se debieron a la caída del precio del petróleo (inferior a $ 35), la secuela de la crisis financiera mundial por EE.UU. y las declaraciones imprudentes del Gobierno sobre declarar deudas ilegítimas por falta de recursos, es decir, se declaró un bono en moratoria, El Universo [2008]. Para el año 2009, se presenta una tendencia a la baja del indicador en respuesta del Gobierno en cuanto al cupón de los bonos Global.

2.3.2. 2010 - 2020

Para los años posteriores, se opina que la calificación del indicador del riesgo país incide por la escasa inversión extranjera, falta de sostenibilidad en las tasas de crecimiento, falta de seguridad jurídica y la dependencia del petróleo, lo que implica que la volatilidad de este comportamiento seguirá hasta que no se maneje adecuadamente la situación económica y fiscal.

Valle y Aguirre [2020] exponen que del año 2010 al 2014, se logra una expansión en la economía ecuatoriana, debido a los altos precios extranjeros del crudo; aunque luego, se produjo una baja de la actividad económica por el desplome de los precios

de este commodity, lo que orientó a una crisis en los pagos de deuda en el 2015. En UTA [2021], se explica que en base a la información del BCE, en el año 2016, entre los meses de enero a abril, se registraron puntuaciones mayores a 1000 pb y, pese al terremoto de abril, el riesgo país fue mejorando a finales de año hasta llegar a un puntaje de 671 pb. Un comportamiento similar ocurrió en el año 2017 que, de enero con 612 pb, se alcanzó a diciembre 470 pb, gracias al repunte de la alza de los precios de petróleo y la solvencia de la demanda interna. En 2018, se tiene un comportamiento inverso que, de una apertura de aproximadamente 450 pb, cierra el año con alrededor de 766 pb, debido a la reducción de consumo final de los hogares y del Gobierno, además por la reducción en inversión.

De la misma referencia, se aclara que durante los tres primeros trimestres del año 2019, el riesgo país no fue mayor a los 750 pb, pero para octubre, debido a las manifestaciones en oposición a las normas optadas por el gobierno nacional, hubo un repunte del indicador por las grandes pérdidas económicas del Estado y las industrias; aunque no fueron cambios extremos como en el 2000 o 2008. No obstante, para finales de 2019 e inicios de 2020 se presentó una disminución del indicador en alrededor de 940 pb, pero para el mes de marzo se llegó a registrar más de 3800 pb, siendo el factor preponderante la pandemia que perjudicó a todo el mundo, trayendo consecuencias en la caída del petróleo. El Ecuador fue uno de los países más afectados en lo económico, debido a que es una economía que depende bastante de las exportaciones de petróleo y, en consecuencia, es vulnerable en las fluctuaciones de precios internacionales de este commodity, en especial, de China y Estados Unidos; de hecho, el 23 de marzo, se registró el puntaje más alto de este histórico, que fue de 6063 pb. Estos puntajes se mantuvieron elevados hasta el mes de agosto y, desde septiembre hasta diciembre, se obtuvo en promedio 1116 pb, gracias a que el Ecuador logró un acuerdo con el FMI para acceder a una línea de crédito de ayuda económica.

2.4. Enfoque del proyecto

Del comportamiento diario del EMBI ecuatoriano y de algunas de sus causas en cuanto al aumento o disminución del mismo, es importante consolidar un comportamiento futuro de este indicador, para así establecer mejores decisiones o recomendaciones en los casos más graves; aunque también, para puntuaciones que podrían catalogarse como normales, hallar una forma de reducirlas más y que de esa manera el Ecuador sea un mejor atractivo para los inversionistas. Así también, a estos

pronósticos se los podría relacionar, por ejemplo, con los precios internacionales del petróleo, pues como se observó en la sección anterior, este commodity tiene un efecto positivo y negativo en las puntuaciones del EMBI.

Por otra parte, aunque un modelo con un enfoque multivariante puede resultar útil para el pronóstico de futuras puntuaciones, en varios estudios sobre el análisis de datos funcionales se han encontrado resultados interesantes, debido a que se basa en un análisis sobre funciones o curvas que describen a los datos. Además de poder predecir una función completa bajo dos perspectivas que se proponen en el presente proyecto y que representarán la forma en la que se comportarían las puntuaciones futuras del indicador, para un horizonte de tiempo dado. Por tanto, en la siguiente sección, se abarca el marco teórico del análisis y modelamiento estadístico con el enfoque funcional.

Capítulo 3

Análisis de Datos Funcionales (FDA)

En el presente capítulo, se aborda una de las técnicas poco conocidas dentro del campo de la Estadística: el Análisis de Datos Funcionales (FDA, por sus siglas en inglés), con el objetivo de estudiar el marco teórico de la misma. Gracias a las innovaciones tecnológicas, en cuanto a registro de información y cálculos computacionales, el FDA ha tenido un buen recibimiento en la comunidad estadística, desarrollándose una gran variedad de técnicas donde al menos una de las variables de un conjunto de datos puede ser vista naturalmente como una función o curva suave. Las principales referencias que se utilizan en este capítulo son: Ramsay y Silverman [2005], Cuevas et al. [2006] y Shang [2013].

Asimismo, para una data funcional, desde el estudio exploratorio hasta la elaboración de modelos de predicción con este enfoque, se han desarrollado algoritmos de las distintas técnicas y se los han implementado en el software estadístico `R`, en particular, en las librerías `fda` y `fda.usc`, cuyas referencias, principalmente, vienen dadas en Ramsay et al. [2009] y Febrero-Bande y de la Fuente [2012], respectivamente.

3.1. Datos Funcionales

Comúnmente, en el caso multivariante, una variable aleatoria puede ser vista en tiempos distintos de un rango $\mathcal{T} = [t_1, t_m]$ con $m > 0$, con lo cual, una observación o dato puede expresarse en la familia aleatoria $\{X(t_j) : j = 1, ..., m\}$; sin embargo, los instantes consecutivos dentro del rango suelen ser pequeños, por ejemplo, datos que se registran a diario, por horas o incluso por minutos. Así, en Ferraty y Vieu [2006], se detalla que una forma de tener en cuenta estos diminutos espacios entre los datos,

es en considerar a los mismos como una observación de una familia continua

$$\mathcal{X} = \{X(t) \,:\, t \in \mathcal{T}\}.$$

Así entonces, de la misma referencia, se extraen las siguientes definiciones:

DEFINICIÓN 3.1. *Una variable aleatoria funcional (v.f.) es una variable aleatoria $\mathcal{X}$ cuyos valores se encuentran en un espacio infinito dimensional o espacio funcional. Un dato funcional es una observación $X \in \mathcal{X}$.*

DEFINICIÓN 3.2. *Un conjunto de datos funcionales es una muestra $\{X_1, ..., X_n\}$ tomada de una variable funcional $\mathcal{X}$.*

De la definición 3.1, surge la pregunta: ¿en qué espacio funcional puede tomar los valores una variable funcional $\mathcal{X}$? En la monografía de Kokoszka y Reimherr [2017], se detalla que la mayoría de las herramientas para el FDA se fundamentan de la teoría de los espacios de Hilbert y, usualmente, los datos funcionales son una colección de funciones $X \in L^2$ definidas en un intervalo común. Por tanto, se presenta la definición del espacio funcional L^2:

DEFINICIÓN 3.3. *Sea $\mathcal{T} = [a, b] \subset \mathbb{R}$. El espacio de las funciones cuadrado integrables se define como:*

$$L^2(\mathcal{T}) = \left\{ X : \mathcal{T} \to \mathbb{R} \,:\, \int_{\mathcal{T}} X(t)^2 \, dt < \infty \right\}.$$

Además, el producto interno se define por:

$$\langle X(t),\, Y(t) \rangle = \int_{\mathcal{T}} X(t) Y(t) \, dt,$$

y la L^2-norma sigue la siguiente relación: $\|X\| = \sqrt{\langle X,\, X \rangle}$.

OBSERVACIÓN. El espacio L^2 sobre un dominio $\mathcal{T}$ es un espacio de Hilbert.

Para complementar lo anterior, en el capítulo dos de Ramsay y Silverman [1997], se enfatiza en la notación del producto interno para considerar las conexiones y similitudes entre el contexto multivariante y funcional, pues como se recalca, siempre hay que tomar en cuenta la naturaleza de los datos, ya sean vectores o funciones; por lo cual, lo importante es cómo se define el producto interno para que se adapte a los tipos de datos. Así, en el apéndice A.1 se presenta una breve descripción del producto interno.

Ahora, es importante tener en cuenta que por lo general, en la práctica, un dato

funcional (observado en un intervalo finito $\mathcal{T}$) se presenta de forma discretizada, típicamente en instantes de tiempo $t_1, ..., t_m \in \mathcal{T}$ con $m > 0$, por lo cual, la primera tarea en el FDA consiste en adaptar la estructura funcional de dicho conjunto discreto de observaciones. Este procedimiento se presenta en la siguiente sección y se fundamente, esencialmente, del libro de Ramsay y Silverman [2005].

3.2. Ajuste de datos discretos a funciones suaves

Supongamos que se dispone de un conjunto finito de datos discretos (t_j, x_j) con $t_j \in \mathcal{T}$ $(j \in \{1, ..., m\})$, siendo m el número de puntos disponibles que representarán a un dato funcional de una variable $\mathcal{X}$. Entonces, el primer paso consistirá en ajustar los datos x_j a una función X, tal que pueda ser calculada para todo $t \in \mathcal{T}$, es decir, el objetivo es en convertir los datos discretos a funciones suaves, donde el término *funcional* sobre los datos observados se enfatiza en el comportamiento intrínseco de los mismos más que a su estructura explícita. Ahora, frecuentemente no solo se dispone de una función X, sino de una muestra de n datos funcionales de una variable $\mathcal{X}$; por tanto, una función X_i, con $i \in \{1, ..., n\}$, depende de los datos discretos x_{ij} medidos en los instantes $t_{ij} \in \mathcal{T}$, con $j \in \{1, ..., m_i\}$. Con esto, es importante notar que las n curvas están sobre un intervalo común $\mathcal{T}$ y los tiempos discretizados pueden ser o no equiespaciados.

OBSERVACIÓN. Para el presente proyecto, por la estructura de los datos funcionales, se tiene que cada dato está discretizado bajo los mismos tiempos, es decir, para cada i se tienen los tiempos t_j con $j \in \{1, ..., m\}$. Este tipo de estructura se encuentra en varias referencias.

Entonces, para ajustar un dato funcional, se asume que existe una función suave (cuadrado integrable) que explica a los datos, a lo cual, uno de los métodos principales para realizar estas representaciones es expresar a los datos por medio de una expansión o combinación lineal de *funciones base*; sin embargo, por la presencia de ruido que puede existir en los datos, suponemos que un dato funcional $i \in \{1, ..., n\}$ se lo observa a través del siguiente modelo:

$$x_{ij} = X_i\left(t_j\right) + \varepsilon_{ij}, \qquad j \in \{1, ..., m\} \tag{3.1}$$

y de forma vectorial visto como:

$$\boldsymbol{x}_i = X_i\left(\boldsymbol{t}\right) + \boldsymbol{\varepsilon}_i$$

donde los residuos $\boldsymbol{\varepsilon_i}$ son independientes con $X_i(\boldsymbol{t})$. Por tanto, la meta principal del suavizamiento a los datos es en eliminar o reducir el error y obtener una estimación lo más fidedigna de la función X_i. Que una función tenga una o más derivadas, se entiende que la misma es suave.

Con esto, para la representación de las funciones, nos basamos en dos etapas:

1. elegir un conjunto conocido $\{\phi_k\}_{k\in\mathbb{N}}$ de funciones base; y,
2. construir un vector de coeficientes, tal que aproxime a la función a través de una expansión de los elementos ϕ_k.

Hay que tomar en cuenta que en la práctica se utiliza un número finito $K \in \mathbb{N}$ de funciones base para aproximar X_i, es decir,

$$X_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t) \equiv \boldsymbol{c}_i^{\intercal}\boldsymbol{\phi}(t). \tag{3.2}$$

donde $\boldsymbol{c}_i \in \mathbb{R}^K$ son los coeficientes de la combinación lineal de X_i y $\boldsymbol{\phi}$ es un vector cuyos elementos son las K funciones base ϕ_k.

OBSERVACIÓN. Considerando toda la muestra de datos funcionales de una variable $\mathcal{X}$, (3.2) se lo puede ver matricialmente como:

$$\boldsymbol{X}(t) = \boldsymbol{C}\boldsymbol{\phi}(t), \tag{3.3}$$

con $\boldsymbol{X}(t)$ el vector de las n funciones $X_i(t)$ y $\boldsymbol{C} \in \mathbb{R}^{n\times K}$ la matriz de coeficientes.

Así, para el paso 1, puesto que estamos trabajando en el espacio L^2, existen diferentes tipos de bases que pueden considerarse; por tanto, se detallan principalmente dos de estas ya que se suelen utilizar con mayor frecuencia en el FDA, Ramsay et al. [2009].

3.2.1. Bases de Fourier

En ciertas ocasiones, las funciones se repiten durante un período τ, por lo cual, para mantener la periodicidad en los datos, se utiliza la base de Fourier dada por:

$$\phi_0(t)=1, \quad \phi_1(t)=\text{sen}(\omega t), \quad \phi_2(t)=\cos(\omega t), \quad \phi_3(t)=\text{sen}(2\omega t), \quad \phi_4(t)=\cos(2\omega t), \quad \cdots$$

donde la constante ω se relaciona con τ por: $\omega = 2\pi/\tau$.

Si los valores de t_j son equiespaciados y la longitud del intervalo $\mathcal{T}$ asigna el periodo, se tiene la siguiente base de funciones ortonormales, para todo $\rho \geq 1$:

$$\phi_0(t) = \frac{1}{\sqrt{\tau}}, \qquad \phi_{2\rho-1}(t) = \frac{1}{\sqrt{\tau/2}}\,\text{sen}(\rho\omega t), \qquad \phi_{2\rho}(t) = \frac{1}{\sqrt{\tau/2}}\cos(\rho\omega t)$$

Además, por la estructura de la base, se observa que después de ϕ_0, las funciones se organizan en pares sucesivos de sen / cos cuyos argumentos se multiplican por $1, 2, ...$ y hasta algún entero ℓ, por tal motivo, se recomienda $K = 1 + 2\ell$ funciones base.

Dentro del paquete `fda` se dispone de la función `create.fourier.basis`, así por ejemplo, en la Figura 3.1 se muestran las cinco primeras funciones definidas en un intervalo $[0, 365]$ y de periodo $\tau = 365$.

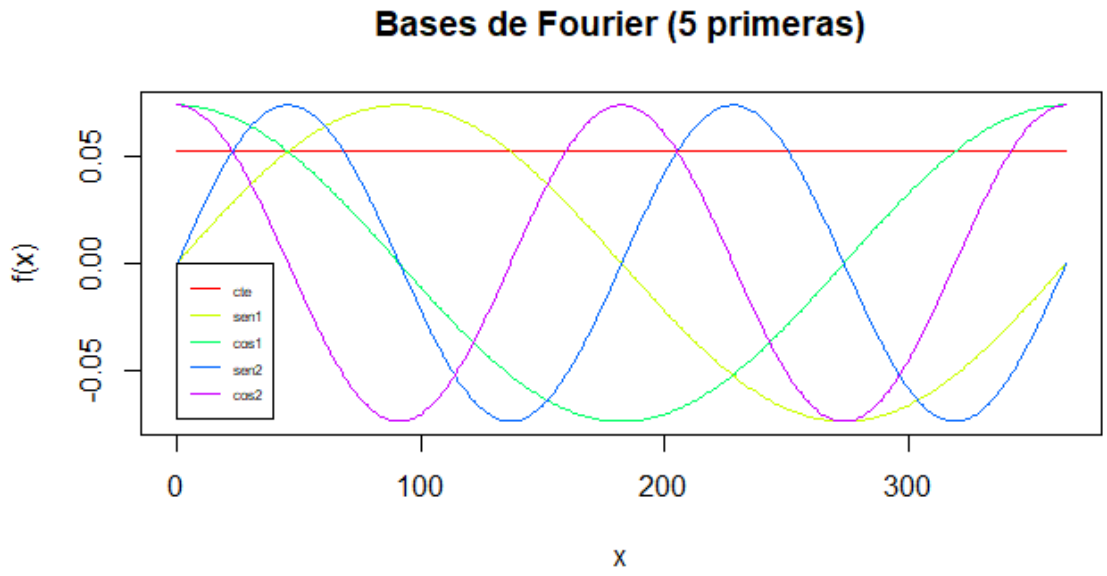


Figura 3.1: Funciones pertenecientes a una base de Fourier
Elaboración: Autor

3.2.2. Bases B-spline

Las funciones spline son polinomios por partes que se ajustan en un intervalo $\mathcal{T}$ y, comúnmente son utilizados para el caso de datos funcionales no periódicos. Entonces, para construir splines se necesita dividir al intervalo de observaciones en L subintervalos divididos por puntos τ_ℓ, $\ell = 1, ..., L-1$, denominados puntos de ruptura (*breakpoints*) o nodos, aunque cabe aclarar que dichos nombres, bajo otro

contexto, no son equivalentes; y, en cada intervalo, se define un polinomio de orden m fijo (splines).

OBSERVACIÓN. El *orden* m de un polinomio $p(x)$ se refiere al número de constantes que lo definen, es decir, el grado de $p(x)$ más uno (el término constante).

En resumen, un sistema de funciones spline se define por la secuencia de nodos y el orden de los segmentos polinomiales y, a menudo, se emplean splines de orden $m = 4$, es decir, cada subintervalo está compuesto por segmentos polinomiales cúbicos.

Una vez visto cómo se define un spline, se necesita especificar cómo se construyen; para esto, se hace uso de una base de funciones f_k con las siguientes propiedades:

- Cada f_k es un spline definido por un orden m y una sucesión de nodos τ.
- Un spline definido por m y τ se lo expresa por una combinación lineal de las funciones base.

Si bien hay varias maneras de construir una base con las características anteriores, la base más popular es la B-spline de orden m, la cual es positiva en no más de m intervalos adyacentes. En de Boor [2001], se encuentra una extensa teoría sobre splines.

La herramienta más útil para definir las funciones base B-spline, en cuanto a implementación computacional, es por medio de la fórmula recursiva de *Cox-de Boor*, la cual se define de la siguiente manera:

Sea $\{\tau_\ell\}_{\ell=1}^{L-1}$ el conjunto de nodos creciente de un intervalo $[\tau_0, \tau_L]$. La $k-$ésima función base B-spline de orden m, denotada por $B_{k,m}$, se define como,

$$B_{k,0}(t) = \begin{cases} 1, & \text{si } t \in [\tau_k, \tau_{k+1}); \\ 0, & \text{otro caso.} \end{cases}$$

$$B_{k,m}(t) = \frac{t - \tau_k}{\tau_{k+m} - \tau_k} B_{k,m-1}(t) + \frac{\tau_{k+m+1} - t}{\tau_{k+m+1} - \tau_{k+1}} B_{k+1,m-1}(t).$$

Así por ejemplo, en la Figura 3.2 se muestran las cinco primeras bases generadas por splines de orden 2 en un intervalo $[0, 3]$ con tres nodos equiespaciados, utilizando la función `create.bspline.basis`

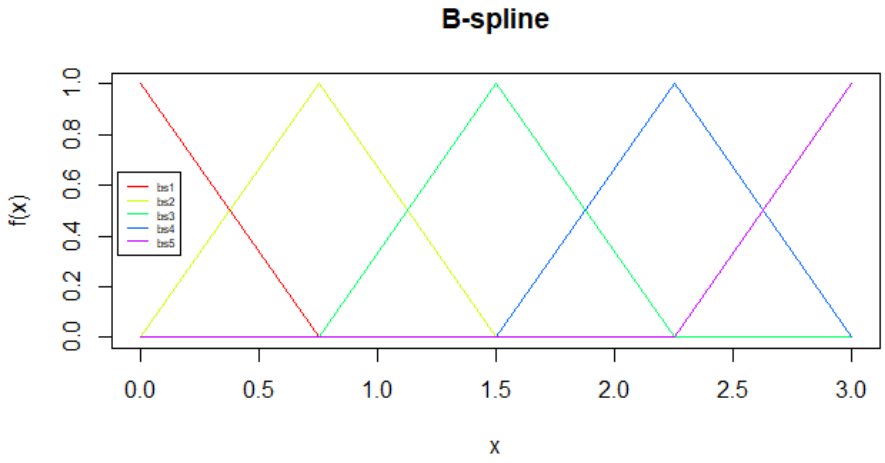


Figura 3.2: Funciones pertenecientes a una base B-splines
Elaboración: Autor

3.2.3. Procedimiento de ajuste

Del modelo (3.1) definido, utilizaremos un suavizamiento por mínimos cuadrados ordinarios para estimar los coeficientes $\boldsymbol{c}_i$ de la combinación lineal de funciones base ϕ_k, tales que aproximen a X_i. Regularmente, se define el ajuste de datos por el criterio de minimizar la suma de errores cuadrados,

$$SSE_i = \sum_{j=1}^{m} \left(x_{ij} - X_i(t_j)\right)^2 = \sum_{j=1}^{m} \left(x_{ij} - \boldsymbol{c_i}^{\mathsf{T}} \boldsymbol{\phi}(t_j)\right)^2. \tag{3.4}$$

Notando por $\boldsymbol{x}_i$ al vector de observaciones del dato funcional i y $\boldsymbol{\Phi}$ a la matriz de orden $m \times K$ que contiene los valores $\phi_k(t_j)$; a (3.4), de forma matricial, se lo ve como:

$$\begin{aligned} SSE_i &= (\boldsymbol{x}_i - \boldsymbol{\Phi c}_i)^{\mathsf{T}} (\boldsymbol{x}_i - \boldsymbol{\Phi c}_i) \\ &= \|\boldsymbol{x_i} - \boldsymbol{\Phi c}_i\|^2 . \end{aligned}$$

Por tanto, tomando la derivada respecto al parámetro $\boldsymbol{c}_i$, se tiene que la solución (estimación) de dicho vector es,

$$\widehat{\boldsymbol{c}}_i = \left(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^{\mathsf{T}} \boldsymbol{x}_i \tag{3.5}$$

y, en consecuencia, el ajuste del vector de datos $\widehat{X}_i(t_j) \approx x_{ij}$ viene dado por:

$$\widehat{X}_i = \boldsymbol{\Phi}\widehat{\boldsymbol{c}}_i = S\,\boldsymbol{x}_i \tag{3.6}$$

con $S = \boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathsf{T}} \in \mathbb{R}^{m\times m}$ conocida como la *matriz de suavizamiento*.

Con esto, se dispone de la metodología para el suavizamiento de los datos. No obstante, no queremos que el ajuste sea perfecto, debido a que la función sería demasiado rugosa; ni pésimo, debido a que se tendría un comportamiento desorientado a los datos; por estas razones, una opción poderosa a emplear es la llamada penalización por rugosidad o simplemente regularización.

3.2.4. Penalización por rugosidad

Un criterio de ajuste para lograr la suavidad en una función, es a través del cuadrado de la segunda derivada de la función X en t, la cual la denotaremos por $[\mathtt{D}^2X(t)]^2$ y, a menudo, es relacionada con la *curvatura* en t, puesto que por ejemplo, la segunda derivada de una recta (función lineal) es cero y, la misma no presenta curvatura. Por consiguiente, para medir la rugosidad de la función, se integra el cuadrado de su segunda derivada, es decir, se utiliza la siguiente penalización:

$$PEN_2(X) = \int \left[\mathtt{D}^2X(t)\right]^2 \, dt. \tag{3.7}$$

Entonces, la expresión (3.8) representa la suma penalizada de errores cuadrados y se la minimiza de modo que regule la suavidad respecto al ajuste de la curva que describe a la función.

$$PENSSE_\lambda(\boldsymbol{x}_i|\boldsymbol{c}_i) = [\boldsymbol{x}_i - \boldsymbol{\Phi c}_i]^\mathtt{T}[\boldsymbol{x}_i - \boldsymbol{\Phi c}_i] + \lambda PEN_2(X_i), \tag{3.8}$$

λ representa el parámetro de suavizamiento que cuantifica la tasa de intercambio entre el ajuste a los datos (SSE_i, primer término) y la variación de la función X_i ($PEN_2(X_i)$, segundo término).

Ahora, mediante el suavizado por regularización, se presenta la forma explícita de estimar $\boldsymbol{c}_i$; para esto, en primer lugar, notemos que (3.7) se puede reescribir de la siguiente manera:

$$\begin{aligned} PEN_2(X) &= \int \left[\mathtt{D}^2X(t)\right]^2 \, dt \\ &= \int \left[\mathtt{D}^2\boldsymbol{c}^\mathtt{T}\boldsymbol{\phi}(t)\right]^2 \, dt \\ &= \int \left[\boldsymbol{c}^\mathtt{T}\,\mathtt{D}^2\boldsymbol{\phi}(t)\right] \left[\mathtt{D}^2\boldsymbol{\phi}(t)^\mathtt{T}\boldsymbol{c}\right] \, dt \\ &= \boldsymbol{c}^\mathtt{T}R\boldsymbol{c} \end{aligned}$$

con $R \in \mathbb{R}^{K\times K}$ la matriz de penalización por rugosidad (matriz de Gram), que con-

tiene los productos entre cada par de las funciones base, es decir,

$$R = \int \mathtt{D}^2\boldsymbol{\phi}(t)\mathtt{D}^2\boldsymbol{\phi}(t)^\intercal\, dt$$
$$= \begin{pmatrix} \int [\mathtt{D}^2\phi_1(t)]^2\, dt & \int \mathtt{D}^2\phi_1(t)\mathtt{D}^2\phi_2(t)\, dt & \cdots & \int \mathtt{D}^2\phi_1(t)\mathtt{D}^2\phi_K(t)\, dt \\ \int \mathtt{D}^2\phi_2(t)\mathtt{D}^2\phi_1(t)\, dt & \int [\mathtt{D}^2\phi_2(t)]^2\, dt & \cdots & \int \mathtt{D}^2\phi_2(t)\mathtt{D}^2\phi_K(t)\, dt \\ \vdots & \vdots & \ddots & \vdots \\ \int \mathtt{D}^2\phi_K(t)\mathtt{D}^2\phi_1(t)\, dt & \int \mathtt{D}^2\phi_K(t)\mathtt{D}^2\phi_2(t)\, dt & \cdots & \int [\mathtt{D}^2\phi_K(t)]^2\, dt \end{pmatrix}$$

Así pues, (3.5) y (3.6), respectivamente, se estimarían bajo las siguientes expresiones:

$$\widehat{\boldsymbol{c}}_i = (\boldsymbol{\Phi}^\intercal\boldsymbol{\Phi} + \lambda R)^{-1}\boldsymbol{\Phi}^\intercal\boldsymbol{x}_i \tag{3.9}$$
$$\widehat{X}_i = \boldsymbol{\Phi}(\boldsymbol{\Phi}^\intercal\boldsymbol{\Phi} + \lambda R)^{-1}\boldsymbol{\Phi}^\intercal\boldsymbol{x}_i := S_\lambda\, \boldsymbol{x}_i. \tag{3.10}$$

El uso de la matriz S_λ también es útil para el cálculo del valor de los grados de libertad para el ajuste definido por λ, y sigue la relación $\mathtt{df}(\lambda) = \mathrm{tr}\,(S_\lambda)$; además, es importante realizar las siguientes observaciones sobre el parámetro λ:

- Para valores de λ suficientemente grandes ($\lambda \to \infty$), $\mathtt{D}^2(X)$ será esencialmente cero, con lo que implicaría que X será un polinomio de grado uno. Visto de otra forma, $PENSSE_\lambda$ tendrá más énfasis en la suavidad de los datos que en el ajuste.
- Para valores de λ muy pequeños ($\lambda \to 0$), los datos se ajustan lo más cerca posible del conjunto de bases seleccionado, es decir, existe menos penalización y X se acerca a un interpolante de datos.

Para finalizar con esta sección, no existe una regla fija para elegir un número finito de bases K y el parámetro λ, la decisión radica en el objetivo de estudio y sobre todo en los datos. En el paquete `fda.usc` se ha implementado la función `optim.basis` basada en validación cruzada (CV) y validación cruzada generalizada (GCV) para elegir adecuadamente un número razonable de bases y un parámetro de penalización, Febrero-Bande y de la Fuente [2012].

3.3. Funciones descriptivas

Convertida la data discreta a datos funcionales bajo el uso sutil de una base, comenzamos a trabajar con los datos funcionales $X_i(t)$ con $t \in \mathcal{T} = [t_1, t_m]$, $i =$

$1, ..., n$ de una variable funcional $\mathcal{X}$. Entonces, se definen las funciones muestrales descriptivas de centralidad y variabilidad.

- **Media:** el estimador de $\mu_{\mathcal{X}}(t) = E\left[\mathcal{X}(t)\right]$ es,

$$\widehat{\mathcal{X}}(t) = \widehat{\mu}_{\mathcal{X}}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t) \tag{3.11}$$

- **Varianza:** el estimador de $\mathrm{Var}_{\mathcal{X}}[t] = E\left[(\mathcal{X}(t) - \mu_{\mathcal{X}}(t))^2\right]$ es,

$$s^2_{\mathcal{X}}(t) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i(t) - \widehat{\mathcal{X}}(t)\right)^2 \tag{3.12}$$

- **Desviación estándar:** $s_{\mathcal{X}}(t) = \sqrt{s^2_{\mathcal{X}}(t)}$

Además, la función de covarianza, que viene dada por

$$\mathrm{Cov}_{\mathcal{X}}(t_1, t_2) = E\left[(\mathcal{X}(t_1) - \mu_{\mathcal{X}}(t_1))(\mathcal{X}(t_2) - \mu_{\mathcal{X}}(t_2))\right]$$

nos provee información de cómo los datos de las funciones en t_1 se relacionan con los de t_2, y es estimado por:

$$v_{\mathcal{X}}(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n}[X_i(t_1) - \widehat{\mathcal{X}}(t_1)][X_i(t_2) - \widehat{\mathcal{X}}(t_2)] \tag{3.13}$$

y la función de correlación asociada viene dada por:

$$\rho_{\mathcal{X}}(t_1, t_2) = \frac{v_{\mathcal{X}}(t_1, t_2)}{s_{\mathcal{X}}(t_1) \cdot s_{\mathcal{X}}(t_2)}.$$

De ello, es frecuente decir que el estimador de la distribución funcional central es (3.11); mientras que el estimador de la escala y estructura de variación de la distribución funcional es (3.13).

3.4. Detección de datos funcionales atípicos

Tal como en el caso multivariante, para la aplicación de técnicas del FDA, las estimaciones pueden estar gravemente sesgadas si existen datos funcionales atípicos que contaminan la muestra; por tanto, es necesario disminuir dicha influencia para aplicar técnicas posteriores.

Estos datos atípicos pueden encontrarse fuera del rango de la gran mayoría de los datos, conocidos como *valores atípicos de magnitud* o tienen una forma muy distinta al de otros datos, aunque pertenezcan al mismo rango, denominados *valores atípicos de forma*. Así, se han propuesto distintas metodologías para la detección de datos funcionales atípicos como en Hyndman y Shang [2010]; sin embargo, en varios estudios se utilizan las medidas de profundidad para utilizar la noción de curva más central, además que para el presente proyecto serán muy útiles en el procedimiento del modelamiento.

3.4.1. Medidas de profundidad

El concepto de profundidad conduce a un orden natural de centro hacia afuera de los datos de una muestra, de manera que los datos que se acercan más al centro consten de una alta profundidad, por ejemplo en el caso multivariante, la mediana suele ser el punto más profundo del conjunto de datos. Con ello, se utiliza el criterio de que la profundidad y lo atípico son nociones inversas, de modo que si la función correspondiente tiene una profundidad significativamente baja, entonces se lo considera un outlier del conjunto de datos, Febrero-Bande y de la Fuente [2012].

El concepto de profundidad surgió como una metodología dentro de la literatura de robustez para medir qué tan profundo o central es un dato; es así que, el uso de las medidas de profundidad en los datos funcionales, consiste en determinar la centralidad de una curva X_i dentro del conjunto de los datos funcionales.

A continuación, se describen dos medidas importantes para nuestro estudio, con base de Cuevas et al. [2006] y Cuevas et al. [2007].

Profundidad de Fraiman y Muniz (FM)

Los autores a quienes se les atribuye el nombre de esta medida de profundidad, la establecieron con la idea de responder cuál de las curvas permanece más en el "medio de la muestra", lo cual corresponde a una noción intuitiva de la mediana.

DEFINICIÓN 3.4. *Sea* $\{X_i(t),\, t \in \mathcal{T}\}_{i=1}^{n}$ *una muestra funcional de una v.f.* $\mathcal{X}$*, cuya función de distribución empírica en* t *se define por:*

$$F_{n,t}(X_i(t)) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}\left\{X_k(t) \leq X_i(t)\right\}.$$

Considere la profundidad univariante de i en t, respecto a la muestra, dada por:

$$Z_i(t) = 1 - \left| \frac{1}{2} - F_{n,t}(X_i(t)) \right|,$$

entonces para cada $i \in \{1, 2, ..., n\}$, se define la profundidad funcional de Fraiman y Muniz del dato i como:

$$FMD_{X_i} = \int_{\mathcal{T}} Z_i(t)\, dt. \tag{3.14}$$

La integral de la ecuación (3.14) se la estima usando sumas de Riemann (Febrero et al. [2008]) pues se utilizan los datos discretos; esta estimación es:

$$FMD_{X_i} = \sum_{j=2}^{m} (t_j - t_{j-1}) \cdot Z_i\left(t_j\right), \qquad i = 1, ..., n.$$

Así, se tiene que la mediana funcional ($FM-$mediana) coincide con el dato más profundo, es decir, es la función $X_{\widehat{\mu}_{FM}}(t)$ con el más alto FMD, donde

$$\widehat{\mu}_{FM} = \arg\max \left\{ FMD_{X_i}, \, i = 1, ..., n \right\}. \tag{3.15}$$

Profundidad Modal

Ahora, se logra extender el concepto de moda a un marco funcional, por lo que esta medida de profundidad se basa en qué tan rodeadas están las curvas con respecto a una distancia y seleccionar la trayectoria más densamente "envuelta" por las demás curvas del conjunto de datos. Formalmente, se define de la siguiente manera:

DEFINICIÓN 3.5. *Sea $\{X_i(t),\, t \in \mathcal{T}\}_{i=1}^{n}$ una muestra funcional de una v.f. $\mathcal{X}$, considere una función kernel $K\colon \mathbb{R}^+ \to \mathbb{R}^+$ y un parámetro de suavizado h, denominado ancho de banda. Entonces, la profundidad funcional de la Moda MD para el dato i, se define por:*

$$MD_{X_i} = \sum_{k=1}^{n} K\left(\frac{\|X_i - X_k\|}{h} \right).$$

Comúnmente se toma h como el 15vo percentil del conjunto de distancias entre las curvas, es decir, de la distribución empírica $\{\|X_i - X_k\| \,:\, i, k = 1, ..., n\}$, Febrero et al. [2008].

Los métodos de Kernel se utilizan de forma intensiva debido a que son una forma útil de hacer una ponderación local; la idea principal de la ponderación alrededor de X_i es en atribuir a cada X_k un peso teniendo en cuenta la norma L^2 y, cuanto más distante X_k de X_i, menor es la ponderación (Ferraty y Vieu [2006]). Además, se

sugiere utilizar el Kernel asimétrico de Gauss, es decir,

$$K(t) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right), \quad t > 0.$$

Asimismo, se tiene que la moda (indistintamente mediana) funcional es el dato más profundo, es decir, es la curva $X_{\widehat{\mu}_{MOD}}(t)$ con el MD más alto, donde

$$\widehat{\mu}_{MOD} = \arg\max \left\{ MD_{X_i}, \, i = 1, ..., n \right\}. \tag{3.16}$$

3.4.2. Media recortada

Fraiman y Muniz [2001] introdujeron un estimador de ubicación robusto del centro de la distribución de los datos funcionales a partir de las medidas de profundidad, la cual es la media funcional $\alpha-$recortada ($\alpha-$trimmed), definiéndola como el promedio de las $n - [\alpha n]$ curvas más centradas, con $0 \leq \alpha \leq \frac{n-1}{n}$ y $[\cdot]$ la función parte entera.

DEFINICIÓN 3.6. *Sean $X_{(1)}X_{(2)}, ..., X_{(n)}$ los datos funcionales ordenados descendentemente bajo un criterio de profundidad, es decir, $X_{(1)}$ y $X_{(n)}$ son las curvas más y menos profundas, respectivamente. La media recortada para un α dado, se define como:*

$$\widehat{\mu}_{\alpha,TM} = \frac{1}{n - [\alpha n]} \sum_{i=1}^{n-[\alpha n]} X_{(i)}.$$

De forma similar, se define la desviación estándar $\alpha-$recortada como sigue:

$$\widehat{s}_{\alpha,TSD} = \sqrt{\frac{1}{n - [\alpha n]} \sum_{i=1}^{n-[\alpha n]} \left(X_{(i)} - \widehat{\mu}_{\alpha,TM}\right)^2}.$$

OBSERVACIÓN. La mediana y la moda son curvas que pertenecen al conjunto de datos, mientras que las medias y medias recortadas son combinaciones lineales de todas las curvas.

3.4.3. Algoritmo de detección

Dentro del paquete `fda.usc` se dispone de la función `outliers.depth.trim` que, como su nombre lo señala, es un procedimiento para descubrir datos funcionales atípicos, basado en la búsqueda de curvas con profundidades bajas y remuestreo. Para esta búsqueda se utiliza el procedimiento *bootstrap suavizado*, el cual trabaja

con la covarianza de los datos. El algoritmo se lo describe a continuación, Febrero et al. [2008]:

Dada una $v.f.\ \mathcal{X}$ con una muestra de n datos funcionales,

1. Calcular las profundidades funcionales $\{D(X_i)\}_{i=1}^{n}$, bajo una de las medidas definidas previamente.
2. Sin considerar el $\alpha\,\%$ de las curvas menos profundas, conseguir Θ muestras bootstrap de n funciones. A estas muestras se las escriben por $X_{i^*}^{\theta}$ con $i^* = 1, ..., n$ y $\theta = 1, ..., \Theta$.
3. Determinar muestras de bootstrap suavizado $Y_{i^*}^{\theta} = X_{i^*}^{\theta} + \Gamma_{i^*}^{\theta}$, con $\Gamma_{i^*}^{\theta}$ tales que $\Gamma_{i^*}^{\theta}(t_1), ..., \Gamma_{i^*}^{\theta}(t_m)$ se distribuyan normalmente con media cero y matriz de covarianza $\gamma\Sigma_{\mathcal{X}}$, siendo $\Sigma_{\mathcal{X}}$ la matriz de covarianza de $X(t_1), ..., X(t_m)$ y γ el parámetro de bootstrap suavizado que maneja la variación de las remuestras.
4. Obtener un valor K^{θ} para cada remuestra $\theta = 1, ..., \Theta$ relacionado al $k-$ésimo percentil de la distribución $\{D(Y_{i^*}^{\theta})\}_{i^*=1}^{n}$ empírica.
5. Tomar el límite K como la mediana de los Θ valores K^{θ}.
6. Si se encuentran curvas tales que $D(X_j(t)) \leq K$, se las consideran curvas atípicas (datos funcionales atípicos) y comúnmente se eliminan de la muestra.

3.5. Análisis de Componentes Principales Funcionales (FPCA)

Las matrices de varianza-covarianza y correlación pueden ser complicadas de interpretarlas y no siempre brindan una característica comprensible de la estructura de la variabilidad de los datos observados directamente, por lo que el Análisis de Componentes Principales otorga una forma de observar dicha estructura, siendo también un método útil para hallar patrones existentes en los datos. Esto ocurre tanto para el caso multivariante como para el caso funcional.

Esta técnica fue uno de los métodos iniciales adaptados al contexto funcional y varios autores describen una relación peculiar de dicha extensión, viéndola simplemente como un cambio a funciones, operadores lineales, operadores de covarianza y productos escalares en espacios funcionales, en lugar de vectores, matrices, matrices de covarianza y productos escalares en espacios vectoriales, respectivamente; es decir, se "intercambian" sumatorios por integrales, Chávez et al. [2015]. Así entonces, en el apéndice A.2, se encuentra una breve descripción de la metodología de la mencionada técnica con el enfoque multivariante y, a continuación, se describe la técnica con el enfoque funcional.

3.5.1. Formulación del problema

El propósito del FPCA es en encontrar un conjunto de funciones ortonormales pertenecientes al espacio $L^2(\mathcal{T})$, llamadas *funciones de peso* (componentes principales), tales que maximicen la varianza durante cada una. Así, para formular el problema, en primer lugar, es habitual asumir que los datos se encuentran centrados; para esto, se sustrae la media de los datos funcionales, es decir, se la resta a cada dato $\mathsf{X}_i(t) = X_i(t) - \hat{\mu}_{\mathcal{X}}(t)$. Luego, para resolver el problema y obtener las componentes, se realiza una descomposición ortogonal a la función de covarianza. Entonces, planteamos el problema:

Primera Componente

De Ramsay y Silverman [2005], suponiendo que se dispone de una muestra de datos funcionales (centrados) $\mathsf{X}_1(t), ..., \mathsf{X}_n(t)$ con $t \in \mathcal{T}$, para determinar la primera componente principal, necesitamos encontrar una *función de peso* $\alpha_1 \in L^2(\mathcal{T})$, de donde las puntuaciones, pesos o aportaciones de cada dato funcional sobre esta componente vienen dadas por:

$$f_{1,i} = \langle \alpha_1, \mathsf{X}_i \rangle = \int_{\mathcal{T}} \alpha_1(t)\mathsf{X}_i(t)\, dt := \int \alpha_1 \mathsf{X}_i, \quad \forall i \in \{1, ...n\}.$$

Con esto, $\alpha_1(t)$ se la obtiene al maximizar el promedio de las puntuaciones al cuadrado, sujeto a la restricción análoga de vector unitario, es decir, se resuelve el

siguiente problema:

$$\begin{cases} \text{Max } \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} f_{1,i}^2 = \dfrac{1}{n}\sum_{i=1}^{n}\left[\int \alpha_1(t)\cdot \mathsf{X}_i(t)\,dt\right]^2 \\ \|\alpha_1\|^2 = \displaystyle\int \alpha_1^2 = 1. \end{cases}$$

Generalización

Para el cálculo de la $j-$ésima componente principal funcional, se determina $\alpha_j(t)$ de forma similar, incluyendo las restricciones de ortogonalidad, por tanto, se resuelve el problema:

$$\begin{cases} \text{Max } \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} f_{j,i}^2 = \dfrac{1}{n}\sum_{i=1}^{n}\left[\int \alpha_j(t)\cdot \mathsf{X}_i(t)\,dt\right]^2 \\ \|\alpha_j\|^2 = \displaystyle\int \alpha_j^2 = 1 \\ \langle \alpha_\ell,\, \alpha_j\rangle = \displaystyle\int \alpha_\ell \cdot \alpha_j = 0 \qquad\qquad \ell < j \end{cases} \tag{3.17}$$

Notemos que al estar los datos centrados, la función objetivo anterior también corresponde a maximizar la varianza que recoge cada componente de los datos funcionales, tal como lo exponen Deręgowski y Krzyśko [2011], pues bajo la notación que llevamos, se la ve como:

$$\text{Var}\left\langle \alpha_j,\, \mathcal{X}\right\rangle = \text{Var}\left[\int \alpha_j(t)\mathcal{X}(t)\,dt\right] = \frac{1}{n}\sum_{i=1}^{n}\left[\int \alpha_j(t)\mathsf{X}_i(t)\,dt\right]^2$$

o equivalentemente, se la puede describir por:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\left[\int \alpha_j(t)\mathsf{X}_i(t)\,dt\right]^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[\int \alpha_j(s)\mathsf{X}_i(s)\,ds\right]\left[\int \alpha_j(t)\mathsf{X}_i(t)\,dt\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\int\int \alpha_j(s)\mathsf{X}_i(s)\mathsf{X}_i(t)\alpha_j(t)\,ds\,dt \\ &= \int\int \alpha_j(s)\left(\frac{1}{n}\sum_{i=1}^{n}\mathsf{X}_i(s)\mathsf{X}_i(t)\right)\alpha_j(t)\,ds\,dt \\ &= \int\int \alpha_j(s)\cdot \nu_{\mathcal{X}}(s,t)\cdot \alpha_j(t)\,ds\,dt. \end{aligned} \tag{3.18}$$

Del problema planteado, nos aseguramos que las componentes principales sean ortogonales y que cada función de peso defina el modo de variación más importante en las curvas, es decir, que cada vez se localice una nueva forma de variación que no ha sido hallada antes. En resumen, deseamos hallar un conjunto de p funciones

ortonormales α_j, tal que la expansión para cada dato i, en términos de las nuevas funciones base, se aproxime lo más cerca posible, es decir,

$$\widetilde{\mathsf{X}}_i(t) = \sum_{j=1}^{p} f_{j,i} \cdot \alpha_j(t).$$

Así también, el problema se lo puede ver como un criterio de ajuste para cada curva y se puede utilizar la integración de los errores al cuadrado:

$$\left\|\mathsf{X}_i - \widetilde{\mathsf{X}}_i\right\|^2 = \int \left[\mathsf{X}_i(t) - \widetilde{\mathsf{X}}_i(t)\right]^2 dt$$

y de manera general, deseamos encontrar una base ortonormal $\alpha_1(t), ..., \alpha_p(t)$, con $p < K$, tal que se minimice:

$$\widehat{S}^2 = \sum_{i=1}^{n} \left\|\mathsf{X}_i - \sum_{j=1}^{p} f_{j,i} \cdot \alpha_j\right\|^2,$$

Por esta última razón, las funciones $\alpha_j(t)$ se las llaman funciones ortonormales empíricas, ya que están determinadas por los datos que se utilizan para la expansión. Además, cabe aclarar que utilizando cualquiera de los dos criterios, se halla el mismo conjunto de funciones de peso (Horváth y Kokoszka [2012]), pues se tiene:

$$\begin{aligned}
\left\|\mathsf{X}_i - \sum_{j=1}^{p} f_{j,i} \cdot \alpha_j\right\|^2 &= \left\langle \mathsf{X}_i - \sum_{j=1}^{p} f_{j,i} \cdot \alpha_j,\, \mathsf{X}_i - \sum_{j=1}^{p} f_{j,i} \cdot \alpha_j \right\rangle \\
&= \langle \mathsf{X}_i, \mathsf{X}_i \rangle - 2 \sum_{j=1}^{p} f_{j,i} \underbrace{\langle \alpha_j, \mathsf{X}_i \rangle}_{f_{j,i}} + \sum_{j=1}^{p} f_{j,i}^2 \underbrace{\langle \alpha_j, \alpha_j \rangle}_{\|\alpha_j\|^2 = 1} \\
&= \|\mathsf{X}_i\|^2 - \sum_{j=1}^{p} f_{j,i}^2
\end{aligned}$$

es decir,

$$\widehat{S}^2 = \sum_{i=1}^{n} \|\mathsf{X}_i\|^2 - \sum_{j=1}^{p} \sum_{i=1}^{n} f_{j,i}^2$$

por lo cual, para minimizar $\widehat{S}^2$, necesitamos maximizar $\sum_{i=1}^{n} f_{j,i}^2$, el cual es el mismo criterio visto anteriormente.

3.5.2. Análisis de autofunciones

Es conocido que para el caso mutivariante, el problema de maximización para hallar las componentes principales se resume en el análisis de autovectores y autovalores sobre la matriz $\text{Cov}(X_p, X_q)$; afortunadamente, para la versión funcional, gracias a la teoría espectral, esta correspondencia se la puede mantener.

Para el caso funcional, se extiende el concepto de matriz de covarianza de la muestra funcional mediante el operador integral de Hilbert de covarianza, o simplemente *operador de covarianza*, $\mathcal{V} \in \mathcal{L}\left(L^2(\mathcal{T})\right)$, como:

$$(\mathcal{V}f)(t) = \int_{\mathcal{T}} v_{\mathcal{X}}(t,s) \cdot f(s)\, ds, \qquad t \in \mathcal{T},\, f \in L^2(\mathcal{T}) \tag{3.19}$$

donde la función $v_{\mathcal{X}}(t,s)$ representa el núcleo del operador $\mathcal{V}$, la cual es: continua; simétrica, pues $v_{\mathcal{X}}(t,s) = v_{\mathcal{X}}(s,t)$; y, definida no negativa, pues la expresión (3.18) es mayor o igual a cero. Por lo tanto, se sigue que el operador $\mathcal{V}$ es simétrico y definido no negativo, Horváth y Kokoszka [2012].

Así entonces, la solución del problema (3.17) se la obtiene al resolver la ecuación integral de Fredholm (auto-ecuación) dada por:

$$\begin{aligned} \int_{\mathcal{T}} v_{\mathcal{X}}(t,s) \cdot \alpha_j(s)\, ds &= \lambda_j \alpha_j(t) \\ \left\langle v_{\mathcal{X}}(t,\cdot),\, \alpha_j \right\rangle &= \lambda_j \alpha_j(t) \end{aligned} \tag{3.20}$$

donde, gracias al Teorema de Mercer[1], se garantiza la existencia de la base ortonormal o autofunciones $\{\alpha_j\}$ perteneciente a $L^2(\mathcal{T})$ y una sucesión de autovalores $\{\lambda_j\}$ no negativos del operador de covarianza $\mathcal{V}$, obteniendo la siguiente relación entre los autovalores y las autofunciones:

$$\lambda_j = \left\langle \mathcal{V}\alpha_j,\, \alpha_j \right\rangle = \int (\mathcal{V}\alpha_j)(t) \cdot \alpha_j(t)\, dt = \text{Var}\left\langle \alpha_j,\, \mathcal{X} \right\rangle$$

Además, se realiza la descomposición ortogonal del núcleo del operador $\mathcal{V}$ como

$$v_{\mathcal{X}}(t,s) = \sum_{j=1}^{p} \lambda_j \alpha_j(t) \alpha_j(s).$$

En consecuencia, se determinan las componentes principales $\alpha_1(t), \alpha_2(t), ..., \alpha_p(t)$, ordenadas respecto a sus correspondientes autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ de

[1]Sea el núcleo continuo $\mathcal{K}$, simétrico y definido no negativo del operador integral $\mathcal{V}$. Entonces, existe una base ortonormal $\{e_k\}_{k\in\mathbb{N}}$ de $L^2(\mathcal{T})$ que representan las autofunciones de $\mathcal{V}$ y una secuencia $\{\lambda_k\}_{k\in\mathbb{N}}$ de autovalores no negativos. Además, el núcleo admite la siguiente representación para todo $s,t \in \mathcal{T}$: $\mathcal{K}(s,t) = \sum_j \lambda_j e_j(s) e_j(t)$ donde la convergencia es absoluta y uniforme.

$\mathcal{V}$, es decir, se resuelve el problema (3.17). Por lo tanto, a los datos funcionales de $\mathcal{X}$ se los puede expresar a través de una expansión con las componentes.

3.5.3. Método computacional

En la práctica, necesitamos convertir el problema de auto-análisis del caso continuo a un problema aproximadamente equivalente de auto-análisis con matrices. Así, como en primer lugar se utiliza una base para ajustar a los datos, se tiene la representación dada en la ecuación (3.3), por lo cual, al centrar los datos, el estimador de la función de covarianza tiene la siguiente forma:

$$v_{\mathcal{X}}(s,t) = \frac{1}{n}\boldsymbol{\phi}(s)^{\intercal}\boldsymbol{C}^{\intercal}\boldsymbol{C}\boldsymbol{\phi}(t)$$

Además, se supone que las autofunciones $\alpha_j(t)$ de la auto-ecuación (3.20) sigue una expansión con las funciones base de la forma:

$$\alpha_j(t) = \sum_{k=1}^{K} d_{jk}\phi_k(t) = \boldsymbol{d}_j^{\intercal}\boldsymbol{\phi}(t)$$

con $\boldsymbol{d}_j$ el vector de coeficientes de la expansión.

Así, la notación matricial de (3.19) es:

$$\begin{aligned}\int v_{\mathcal{X}}(t,s)\alpha_j(s)\,ds &= \int \left(\frac{1}{n}\boldsymbol{\phi}(t)^{\intercal}\boldsymbol{C}^{\intercal}\boldsymbol{C}\boldsymbol{\phi}(s)\right)\left(\boldsymbol{\phi}(s)^{\intercal}\boldsymbol{d}_j\right)\,ds \\ &= \boldsymbol{\phi}(t)^{\intercal}\left(\frac{1}{n}\boldsymbol{C}^{\intercal}\boldsymbol{C}\,\boldsymbol{W}\,\boldsymbol{d}_j\right)\end{aligned}$$

con $\boldsymbol{W} = \int \boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^{\intercal}\,ds$.

Por consiguiente, la auto-ecuación (3.20) se la expresa como:

$$\boldsymbol{\phi}(t)^{\intercal}n^{-1}\boldsymbol{C}^{\intercal}\boldsymbol{C}\,\boldsymbol{W}\,\boldsymbol{d}_j = \lambda_j\boldsymbol{\phi}(t)^{\intercal}\boldsymbol{d}_j$$

donde cabe notar que, como esta igualdad se satisface para todo $t \in \mathcal{T}$, lo anterior se reduce a:

$$\frac{1}{n}\boldsymbol{C}^{\intercal}\boldsymbol{C}\,\boldsymbol{W}\,\boldsymbol{d}_j = \lambda_j\boldsymbol{d}_j \tag{3.21}$$

De lo anterior, surgen los siguientes casos, Deręgowski y Krzyśko [2011] y Ramsay y Silverman [2005], respectivamente:

- Si la base es ortonormal, se tiene que $W = I_K$, por lo que el problema se reduce

a un PCA multivariante con el auto-análisis sobre la matriz $\frac{1}{n}\boldsymbol{C}^{\intercal}\boldsymbol{C}$ y, entonces, se determinarían los autovectores $\boldsymbol{d}_j$ y por ende las autofunciones $\alpha_j(t)$.

Se sigue también que la puntuación del $i-$ésimo dato funcional sobre la $j-$ésima componente principal funcional es:

$$f_{j,i} = \langle \alpha_j,\, X_i \rangle = \boldsymbol{c}_i^{\intercal}\boldsymbol{d}_j, \quad i = 1, ..., n;\; j < K$$

- Si W no es la matriz identidad, notemos que de $\|\alpha_j\|^2 = 1$ da como resultado que $\boldsymbol{d}_j^{\intercal}\boldsymbol{W}\boldsymbol{d}_j = 1$, con lo cual, para determinar las componentes principales deseadas, se define

$$\boldsymbol{u}_j = \boldsymbol{W}^{1/2}\boldsymbol{d}_j$$

y se resuelve el problema análogo de autovalores, es decir, de (3.21) se tiene que:

$$\frac{1}{n}\boldsymbol{W}^{1/2}\boldsymbol{C}^{\intercal}\boldsymbol{C}\boldsymbol{W}^{1/2}\boldsymbol{u}_j = \lambda_j\boldsymbol{u}_j.$$

Y por tanto, las autofunciones α_j se obtienen por medio del vector de coeficientes $\boldsymbol{d}_j = \boldsymbol{W}^{-1/2}\boldsymbol{u}_j$.

3.5.4. Interpretación

Ahora, consideremos algunas herramientas importantes que pueden ayudar en la interpretación de las componentes principales funcionales, Ramsay y Silverman [2005].

Perturbaciones de la media

Uno de los métodos más usados, consiste en examinar las gráficas de la función media general $\widehat{\mathcal{X}}(t) = \widehat{\mu}_{\mathcal{X}}(t)$ y las funciones calculadas, sumando y restando una constante factible de la función de la componente principal.

Este método permite examinar y observar el efecto de las componentes sobre las funciones muestrales, pues recordemos que las componentes son útiles para describir la variación respecto a la curva media. Al construir el gráfico, se debe seleccionar una constante $\mathcal{C}$ como la resta en media cuadrática entre la media y su promedio a lo largo del tiempo, es decir,

$$\mathcal{C}^2 = \frac{1}{T}\left\|\widehat{\mu} - \bar{\mu}\right\|^2, \text{ donde } \bar{\mu} = \frac{1}{T}\int \widehat{\mu}(t)\, dt.$$

Así, se establece que es apropiado graficar $\widehat{\mu}$ y $\widehat{\mu} \pm 0.2\mathcal{C}\alpha_j$, aunque en lugar de 0.2 se puede elegir una constante γ cualquiera, dependiendo del comportamiento de $\widehat{\mu}$ y poder lograr interpretar los modos de variabilidad. Se recomienda mantener fijo γ para las gráficas con las demás componentes.

Puntuaciones de las componentes principales

Otro aspecto importante en la interpretación, es examinar las puntuaciones $f_{j,i}$ de los datos sobre las componentes y, así, observar las variaciones entre una componente y otra. Estas puntuaciones se las pueden representar en el plano cartesiano.

Número de componentes

No hay un criterio fijo para determinar el número de componentes principales funcionales a utilizar; sin embargo, existen varios criterios para elegir un número $p < K$ de componentes. Dos de los métodos más comunes son:

- *Cut-off:* Se elige un corte, comúnmente entre el 70 % a 90 % y se selecciona el entero ℓ más pequeño para el cual $VA_\ell > corte$, donde VA_ℓ representa la proporción de varianza acumulada por las $\ell-$ésimas primeras componentes, es decir,
$$VA_\ell = \sum_{k=1}^{\ell} \lambda_k \Big/ \sum_{k=1}^{K} \lambda_k$$
- *Gráfico de Sedimentación:* se grafican los puntos k frente a los λ_k o $\log \lambda_k$ y punto a punto se unen por medio de una línea y, cuando se obtenga una línea recta casi constante, se toma el menor k.

3.5.5. PCA vs FPCA

Para finalizar con esta sección, en Chávez et al. [2015] se describe una comparación de la técnica de componentes principales tanto del contexto multivariante como del funcional, siendo lo más notable: el cambio de espacio sobre el cual se analizan los datos, las metodologías y las operaciones que se emplean, como por ejemplo, el ajuste de los datos por funciones suaves. Así entonces, en la Tabla 3.1 se presentan las diferencias de notación para ambos casos.

	PCA	FPCA
Variables	$X = [X_1, X_2, ..., X_p]$ $X_i = [x_{1i}, ..., x_{ni}]$	$\mathcal{X} = [X_1(t), ..., X_n(t)]$ $t \in \mathcal{T} = [t_1, t_m]$
Datos	Vectores $\in \mathbb{R}^n$	Curvas $\in L^2(\mathcal{T})$
Covarianza	Matriz: $\Sigma = \text{Cov}(X) \in \mathbb{R}^{p \times p}$	Operador $\mathcal{V}$ en $\mathcal{T}$ $\mathcal{V} : L^2(\mathcal{T}) \to L^2(\mathcal{T})$
Vectores propios	Vectores $\alpha_j \in \mathbb{R}^p$, $\Sigma\alpha_j = \lambda_j\alpha_j$ $1 \leq k \leq \text{mín}(n, p)$	Funciones $\alpha_j(t) \in L^2(\mathcal{T})$ $(\mathcal{V}\alpha_j)(t) = \lambda_j\alpha_j$ $1 \leq j \leq K$
Componentes Principales	Variables aleatorias en $\mathbb{R}^p$	Variables aleatorias en $L^2(\mathcal{T})$

Tabla 3.1: Notaciones del PCA y FPCA

3.6. Modelo de predicción basado en FPCA

Hyndman y Ullah [2007] propusieron un nuevo método para el pronóstico de tasas de mortalidad y fertilidad específicas por edad, involucrando una descomposición de las componentes principales, donde señalan que utilizan más de una componente empleando la formulación de datos funcionales de lo cual conlleva a los métodos de suavizamiento para disminuir el error de los datos observados. Así, bajo este enfoque, se logra pronosticar una función completa para períodos de tiempo futuros, es decir, el propósito de este modelamiento consiste en estimar las tendencias futuras.

A consecuencia de este método, surgieron algunos estudios aplicando el mismo; tal es el caso que Shang [2013] presenta el paquete `ftsa` de `R` para aplicar esta técnica, mostrando un caso de estudio y sus resultados. Con esto, basándonos en este último documento, se presenta la estructura del modelo con el método de Hyndman y Ullah (HU) a utilizar en este proyecto.

3.6.1. Modelamiento

Sea una muestra $X_1(t), X_2(t), ..., X_n(t)$ de una $v.f.$ $\mathcal{X}$ evaluadas en un intervalo $\mathcal{T} = [t_1, t_m]$. El ajuste del modelo o la descomposición por componentes principales funcionales se escribe como:

$$X_i(t) = \widehat{\mathcal{X}}(t) + \sum_{j=1}^{p} f_{j,i}\alpha_j(t) + \varepsilon_i(t) \tag{3.22}$$

donde $\widehat{\mathcal{X}}(t)$ es la estimación de la media funcional, $\alpha_j(t)$ la $j-$ésima componente principal funcional empírica del operador de covarianza, $f_{j,i}$ la puntuación del $i-$ésimo dato funcional en la $j-$ésima componente principal, $\varepsilon_i(t)$ es el componente de error no autocorrelacionado (independiente) y de media cero para cada t y p el número de componentes principales a usar.

3.6.2. Pronósticos

Así, para estimar valores futuros $x_i(t_j)$, se logran predecir funciones $X_i(t)$, con $i = n+1, ..., n+h$ y $t \in \mathcal{T}$, siendo h un horizonte de tiempo dado. Para esto, los coeficientes (puntuaciones) $\{f_{j,i}\}_{i=1}^{n}$ de las componentes principales construidas $j = 1, ..., p$, se pronostican por medio de modelos de series de tiempo univariantes, como los modelos de suavizamiento exponencial o ARIMA. Luego, se multiplican los coeficientes pronosticados con las componentes principales, dando como resultado la predicción de una función.

Observación. Los métodos univariantes serán factibles para el pronóstico de cada serie de los coeficientes, debido a la forma en la que se definen las componentes principales, pues los mismos tienen un comportamiento independiente uno de otro. Por tanto, los coeficientes $f_{k,i}$ y $f_{\ell,i}$ no se correlacionan, $k \neq \ell$; aunque pueden haber correlaciones muy pequeñas, Hyndman y Ullah [2007].

De forma matemática, para el pronóstico de observaciones $x_i(t_j)$ hasta un horizonte de tiempo h, condicionando sobre el conjunto de funciones suaves $X(t)$ y las componente principales construidas $\alpha(t) = [\alpha_1(t), ..., \alpha_p(t)]$, se obtiene:

$$\begin{aligned}\widehat{x}_{n+h|n}(t) &= \mathrm{E}\left[x_{n+h}(t) \mid X(t),\, \alpha(t)\right] \\ &= \widehat{\mathcal{X}}(t) + \sum_{j=1}^{p} f_{j,n+h|n}\, \alpha_j(t).\end{aligned}$$

Se usa la notación $\widehat{x}_{n+h|n}$ para denotar un pronóstico puntual de x_{n+h} usando la información disponible en el tiempo n. Por otra parte, para el caso de estudio a tratar en el presente proyecto, se utilizará un modelo univariante de suavizado exponencial, el cual se lo explica en la siguiente subsección.

Para culminar con esta parte, es interesante mencionar lo descrito en Kokoszka y Reimherr [2017], pues señala que la ventaja de este método frente a series de tiempo funcionales es que no es necesario asumir estacionariedad ni una estructura de au-

tocorrelación, por lo cual se pueden aplicar varios procedimientos para pronosticar series de tiempo univariantes. Aunque también, señala que su uso óptimo no se ha investigado teóricamente.

3.6.3. Suavizado exponencial

La referencia principal para estudiar estos métodos de pronósticos es Hyndman et al. [2008]. Existen varios métodos que se encuentran dentro de la familia de suavizado exponencial, con la propiedad de que los pronósticos son combinaciones ponderadas de las observaciones pasadas y a las observaciones más recientes se les asigna relativamente más peso que las anteriores. A medida que se recorren las observaciones, los pesos disminuyen exponencialmente, de ahí el nombre de *suavizado exponencial*.

Suavizado exponencial simple

Para este tipo de modelos se utiliza el error de predicción, pues se toma la predicción del periodo anterior y se lo ajusta utilizando dicho error, es decir, la predicción para un periodo siguiente es:

$$\widehat{x}_{t+1} = \widehat{x}_t + \alpha(x_t - \widehat{x}_t) \tag{3.23}$$

donde $\alpha \in (0,1)$. Notemos que si $\alpha \to 1$, el nuevo pronóstico incluirá un ajuste sustancial del error del pronóstico previo y si $\alpha \to 0$ el nuevo pronóstico tendrá poco ajuste.

Para relacionar lo descrito al inicio del suavizado exponencial, notemos que (3.23) se puede reescribir como:

$$\widehat{x}_{t+1} = \alpha x_t + (1-\alpha)\widehat{x}_t, \tag{3.24}$$

de donde $\widehat{x}_{t+1}$ se basa en ponderar la observación más actual x_t con un valor de ponderación α y la predicción más actual $\widehat{x}_t$ con un valor de ponderación $(1-\alpha)$; por tanto, el modelo se lo puede comprender como un promedio ponderado del pronóstico y la observación más actual.

Además, para finalizar, si la expresión (3.24) se la expande reemplazando $\widehat{x}_t$, se tiene que:

$$\widehat{x}_{t+1} = \alpha x_t + (1-\alpha)\left[\alpha x_{t-1} + (1-\alpha)\widehat{x}_{t-1}\right]$$

$$= \alpha x_t + \alpha(1-\alpha)x_{t-1} + (1-\alpha)^2 \widehat{x}_{t-1}$$

y, de forma general, se obtendría lo siguiente:

$$\widehat{x}_{t+1} = \sum_{\ell=0}^{t-1} \left[\alpha(1-\alpha)^{\ell} x_{t-\ell}\right] + (1-\alpha)^t \widehat{x}_1$$

lo que quiere decir que $\widehat{x}_{t+1}$ representa una media móvil ponderada de todas las observaciones anteriores, donde los pesos disminuyen de manera exponencial. En la Figura 3.3 se presenta la conducta de la secuencia $\alpha(1-\alpha)^{\ell}$ para $\alpha = 0{,}5$ y $\ell = 0, 1, ..., 15$.

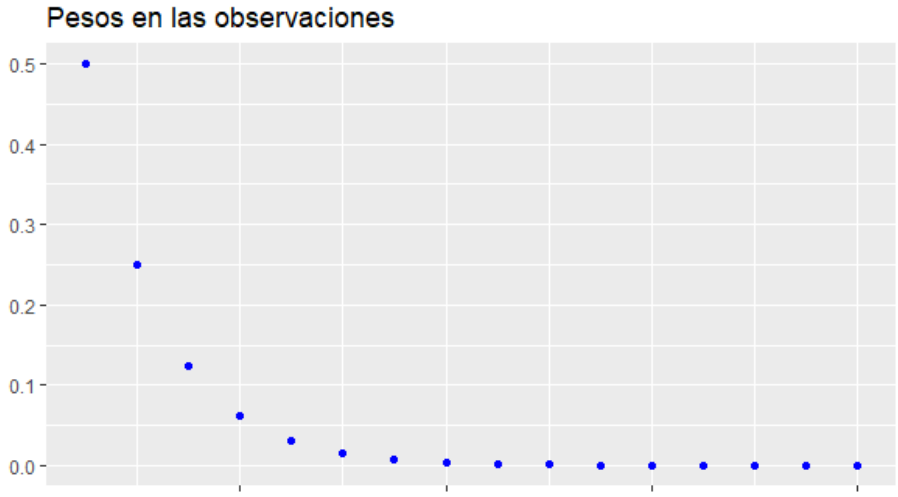


Figura 3.3: Comportamiento exponencial para $\alpha = 0{,}5$
Elaboración: Autor

OBSERVACIÓN. Adaptando la notación que se ha abordado a lo largo de este capítulo, las series de las puntuaciones se estimarían de la siguiente manera (siguiendo la relación (3.24)), para todo $j = 1, ..., p$

$$\widehat{f}_{j,i+1} = \alpha f_{j,i} + (1-\alpha)\widehat{f}_{j,i}.$$

$\widehat{f}_{j,i+1}$ representa la predicción de la puntuación del dato funcional $i+1$ en la componente principal j.

Cabe mencionar que se hace énfasis en este tipo de pronósticos debido a que no se considera una componente de tendencia y de estacionalidad en la serie, siendo necesario por la estructura y contexto que se abarcará en el presente proyecto. Sin embargo, en el libro de Hyndman et al. [2008], se consideran distintos tipos de tendencia y estacionalidad, desarrollando quince métodos de suavizado exponencial,

de los cuales existen dos posibles modelos de espacio de estado: uno correspondiente al modelo con error aditivo ($\varepsilon_t = x_t - \widehat{x}_t$) y otro con error multiplicativo $\left(\varepsilon_t = \dfrac{x_t - \widehat{x}_t}{\widehat{x}_t}\right)$, dando un total de treinta modelos de suavizado exponencial. Para este estudio, y por la estructura del suavizado exponencial simple, se considera el modelo de espacio de estado aditivo, además que es importante aclarar que los modelos multiplicativos son útiles para datos estrictamente positivos.

Por esta razón, a los métodos de suavizado exponencial se los suele abreviar como *ETS*, para especificar el *error*, la *tendencia* y la *estacionalidad*, aunque también se lo considera como abreviación de *ExponenTial Smoothing*.

3.6.4. Bandas de Confianza Bootstrap

Para evaluar la precisión de la curva pronosticada por el modelo, en cuanto a su comportamiento, se calculan las bandas de confianza mediante el procedimiento de bootstrap suavizado, como se lo vio en la subsección 3.4.3. Así, para la construcción de las bandas de confianza (Cuevas et al. [2006]) y con la guía del algoritmo implementado en la función `fdata.bootstrap`, del paquete `fda.usc` (Febrero-Bande y de la Fuente [2012]), se describe el algoritmo para posteriormente implementarlo:

Dada una $v.f.$ $\mathcal{X}$ con una muestra consolidada de N datos funcionales,

1. Obtener Θ muestras bootstrap de N funciones, es decir, para cada $\theta = 1, ..., \Theta$ tomar aleatoriamente $X^{\theta}_{i^*}$ con $i^* = 1, ..., N$, y determinar las muestras de bootstrap suavizado $Y^{\theta}_{i^*} = X^{\theta}_{i^*} + \Gamma^{\theta}_{i^*}$, con $\Gamma^{\theta}_{i^*}$ tales que $\Gamma^{\theta}_{i^*}(t_1), ..., \Gamma^{\theta}_{i^*}(t_m)$ se distribuyan normalmente con media cero y matriz de covarianza $\gamma\Sigma_{\mathcal{X}}$.

2. De la muestra original $\mathcal{S}_N = X_1(t), ..., X_N(t)$ y de las muestras bootstrap $\mathcal{S}^{\theta}_N = Y^{\theta}_1, ..., Y^{\theta}_N$, para $\theta = 1, ..., \Theta$, se aplica el modelo (3.22) para predecir (estimar) las curvas $X_{\mathcal{S}_N}$ y $X_{\mathcal{S}^{\theta}_N}$, respectivamente.

3. Obtener el conjunto de distancias $D = \left\{ d\left(X_{\mathcal{S}_N},\, X_{\mathcal{S}^{\theta}_N}\right) \,:\, \theta = 1, ..., \Theta \right\}$, a partir de la L^2-norma.

4. A través del cuantil $(1-\alpha)$ de D, determinar, a un nivel de confianza $(1-\alpha)$ %, las bandas de confianza bootstrap.

3.6.5. Nueva propuesta para pronosticar dos escenarios

En Hyndman y Shang [2009], se expone que al dar un mayor peso a los datos más actuales (últimos datos), el esquema del modelo puede adaptarse muy bien a otra información. Así, para el caso de estudio a analizar, correspondiente al riesgo país, se desea pronosticar el mejor y peor caso de las puntuaciones diarias para el año 2021; por lo cual, se utilizará todo el modelo descrito anteriormente, utilizando las medidas de profundidad de las curvas, visto en la subsección 3.4.1, pues como se observó, el suavizado exponencial da más peso a las últimas observaciones.

Entonces, supongamos que se usó una medida de profundidad $D \in \{FMD, MD\}$, de donde para cada dato funcional i, con $i = 1, ..., n$, se tienen las profundidades D_i; con esto, recordando que las curvas más centrales tienen mayor profundidad, se proponen los siguientes escenarios para el modelo (3.22):

Mejor Caso

Para este escenario se desea brindar más peso a las curvas más centrales de la muestra considerada, ya que a dichos datos funcionales se los podría relacionar como años normales, es decir, aquellos años donde se han registrado puntuaciones del EMBI comunes en el Ecuador, según el histórico a considerar. A continuación, se describe la metodología para predecir una función que represente un año normal en cuanto a puntuaciones del EMBI en el Ecuador.

1. Ordenar los datos funcionales según la estructura ascendente de los D_i, es decir, si se tiene que:
$$D_{(1)} \leq D_{(2)} \leq \cdots \leq D_{(n)}$$
se utiliza la data funcional "ordenada" como:
$$X_1^+ = X_{(1)}(t),\ X_2^+ = X_{(2)}(t),\ ...., X_n^+ = X_{(n)}(t),$$
es decir, X_n^+ sería el dato más reciente del conjunto de datos.

2. Utilizando el ordenamiento, para p componentes principales funcionales, se predicen las puntuaciones $\widehat{f}_{j,i+1}^+$ del dato funcional X_{i+1}^+ en la componente j, por:
$$\widehat{f}_{j,i+1}^+ = \alpha f_{j,i}^+ + (1-\alpha)\widehat{f}_{j,i}^+.$$

3. Como se indicó, a los últimos datos funcionales, por ejemplo, a X_{n-2}^+, X_{n-1}^+, X_n^+

les proporcionarán más peso para predecir X_{n+1}^{+}, de donde tendrá un comportamiento similar al de las curvas centrales de la muestra.

Peor Caso

En este escenario se da más peso a las curvas menos centrales de la muestra considerada, es decir, a aquellas donde se han alejado de las puntuaciones normales del EMBI ecuatoriano, pero no de una manera extrema como en los datos atípicos que se determinen. A continuación, se describe la metodología:

1. Ordenar los datos funcionales según la estructura descendente de los D_i, es decir, si se tiene que:
$$D_{(1)} \geq D_{(2)} \geq \cdots \geq D_{(n)}$$
se utiliza la data funcional "ordenada" como:
$$X_1^{-} = X_{(1)}(t),\ X_2^{-} = X_{(2)}(t),\ ...., X_n^{-} = X_{(n)}(t),$$
es decir, X_n^{-} sería el dato más reciente del conjunto de datos.

2. De dicho orden, para p componentes principales funcionales, se predicen las puntuaciones $\widehat{f}_{j,i+1}^{-}$ del dato funcional X_{i+1}^{-} en la componente j, por:
$$\widehat{f}_{j,i+1}^{-} = \alpha f_{j,i}^{-} + (1-\alpha)\widehat{f}_{j,i}^{-}.$$

3. De igual manera, a los últimos datos funcionales, por ejemplo, a X_{n-2}^{-}, X_{n-1}^{-}, X_n^{-}, les proporcionarán más peso para predecir X_{n+1}^{-}, de donde tendrá un comportamiento distinto a lo normal, ya que dichas curvas tienen las profundidades más bajas, y en este caso, la curva a predecir tomará valores no muy comunes, según la estructura de los datos del EMBI ecuatoriano considerados.

OBSERVACIÓN. Para ambos escenarios, se utilizan las mismas p componentes principales funcionales, puesto que las mismas recogen la mayor variación de los datos funcionales y no se toman en cuenta las medidas de profundidad.

Finalmente, para obtener las bandas de confianza en cada escenario, se ordenarían los datos funcionales de cada muestra de bootstrap suavizado según las medidas de profundidad obtenidas por el algoritmo descrito en 3.4.3. Sin embargo, debido a que en cada muestra bootstrap se aplica el modelo, puede ser el caso que

lo estimado se desoriente totalmente del comportamiento de la función de cierto escenario, por lo cual se recomienda utilizar un nivel de significación de $\alpha = 0{,}1$ (10 %) para tener una mejor confianza de las bandas y evaluar de manera más precisa la función pronosticada, ya que si se trabajaría con una significación más alta, la región que formarían las bandas no explicarían el comportamiento de cierto escenario.

Por otra parte, es importante notar que debido al esquema explicado para el modelamiento, en particular, en la mayor ponderación que se les da a los últimos datos funcionales en el suavizado exponencial; en cada muestra bootstrap, de igual manera, para el pronóstico de las funciones se dará más importancia o peso a los últimos datos funcionales (por ejemplo, a los de mayor medida de profundidad en el mejor caso), y los primeros tendrán un aporte relativamente bajo; por lo cual, de los datos funcionales de cada muestra, no todos aportarán en la construcción de las bandas de confianza.

Capítulo 4

Metodología FDA aplicado a datos históricos del indicador del Riesgo País del Ecuador

Una vez conocido el marco teórico sobre la aplicación de la metodología del FDA, en el presente capítulo se estructurará la misma sobre datos históricos del indicador del riesgo país ecuatoriano, medido a través del EMBI, dando la descripción desde el preprocesamiento de los datos hasta el modelo a crear; además, se presentarán las funciones del software `R` a utilizar. Adicionalmente, en el Apéndice B se encuentra todo el código realizado para este capítulo.

4.1. Descripción de la base de datos

En la página oficial del Banco Central de Ecuador (BCE [2021]), se dispone de las puntuaciones diarias del EMBI desde el año 2004 hasta lo que va del año 2021 (05 de agosto), la información de este último año no se la utilizará para el modelo, sino servirá para efectos de comparación. Sin embargo, en el siguiente enlace[1], de igual manera correspondiente al BCE, se han registrado las puntuaciones del EMBI diario desde el año 2000 hasta el año 2018. Por lo tanto, utilizando ambas bases de datos, se consideran los datos del EMBI diario desde el año 2000 hasta el 2020.

La información recuperada se la presenta en una sola columna y con un formato no conveniente para aplicar la técnica FDA; en la Tabla 4.1 se puede observar cómo

[1] `https://www.bce.fin.ec/index.php/component/weblinks/weblink/18-informacion-economica/521?Itemid=101`

se recuperaron los datos. Así, se debe realizar un preprocesamiento de la información para comenzar con la técnica.

Fecha,Riesgo PaÃs"
2000/01/01 00:00:00,"3353"
2000/01/02 00:00:00,"3353"
2000/01/03 00:00:00,"3293"
2000/01/04 00:00:00,"3379"
2000/01/05 00:00:00,"3413"

Tabla 4.1: Extracto de la información recuperada del BCE

4.1.1. Preprocesamiento

Luego de obtener ambas bases de datos, se deben organizar las mismas en una sola, de manera que cada fila represente el mes-día y cada columna el año, es decir, se debe crear una base de datos con 365 filas (del 01 de enero al 31 de diciembre) y 21 columnas (del año 2000 al 2020). Entonces, tomando la información del 2000 al 2017 y del 2018 al 2020 de las dos fuentes del BCE anteriormente mencionadas, se estructuraron los registros como se muestra en la Tabla 4.2.

	MesDia	2000	2001	2002	2003	2004
1	**1-1**	3353	1415	1233	1801	799
2	**1-2**	3353	1457	1218	1790	770
3	**1-3**	3293	1377	1158	1717	770
4	**1-4**	3379	1370	1135	1717	770
5	**1-5**	3413	1378	1135	1717	770

Tabla 4.2: Extracto de la base de datos ordenada

Ahora, del conjunto de datos construido, se tiene que existen valores perdidos (0.9 % del total) tal como se observa en la Figura 4.1. Cabe mencionar que al haber registro de las puntuaciones en el día 29 de febrero para años bisiestos, esta información no se la considerará ya que no siempre se tiene registro para ese día y un método de imputación no es requerido, pues dicha información, en realidad, no existe para ciertos años.

No obstante, se puede observar que no solo existen valores perdidos el 29 de febrero para los años no bisiestos, sino también faltan puntuaciones en los años 2018, 2019 y 2020; estos valores se los resumen en la Tabla 4.3.

Así, luego de indagar sobre la estructura de los datos, se observó que los días de la Tabla 4.3 corresponden a fines de semana (sábado y domingo), de donde las pun-

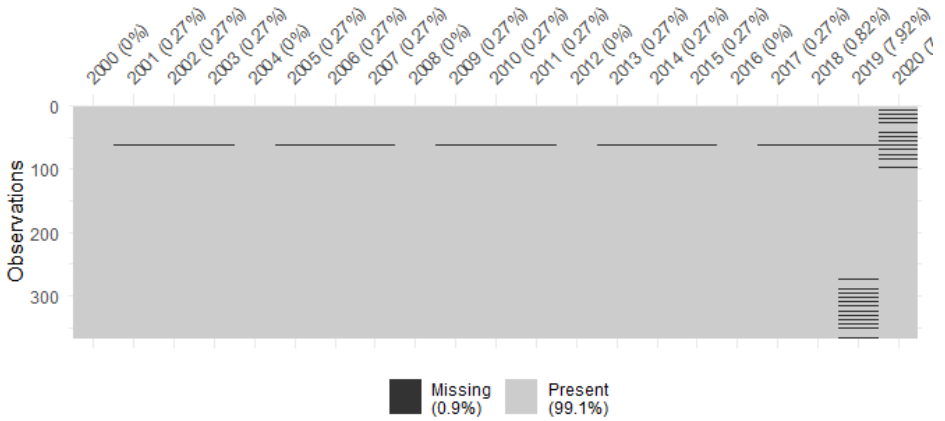


Figura 4.1: Valores perdidos EMBI
Elaboración: Autor

Año	Mes	Días
2018	Febrero	29 (no bisiesto)
	Mayo	19 - 20
2019	Febrero	29 (no bisiesto)
	Septiembre	28 - 29
	Octubre	5 - 6 - 12 - 13 - 19 - 20 - 26 - 27
	Noviembre	2 - 3 - 9 - 10 - 16 - 17 - 23 - 24 - 30
	Diciembre	1 - 7 - 8 - 14 - 15 - 21 - 22 - 28 - 29
2020	Enero	4 - 5 - 11 - 12 - 18 - 19 - 25 - 26
	Febrero	1 - 2 - 8 - 9 - 15 - 16 - 22 - 23 - 29
	Marzo	1 - 7 - 8 - 14 - 15 - 21 - 22
	Abril	4 - 5

Tabla 4.3: Valores perdidos por año, mes y día

tuaciones en estos días se los registra por las puntuaciones de los días viernes de la semana correspondiente. Entonces, para consolidar la base de datos, se retira la fila 60 referente al día 29 de febrero y se imputan los demás datos bajo lo mencionado.

Para finalizar, una vez consolidada la base de datos con las puntuaciones x_{ti} con $t \in \{1, 2, ..., 365\} \equiv \{1-1, 1-2, ..., 12-30, 12-31\}$ e $i \in \{1, ..., 21\} \equiv \{2000, ..., 2020\}$, en la Figura 4.2 se muestran las tendencias diarias de cada año del EMBI ecuatoriano, donde se observa que para ciertos años se tiene un comportamiento muy distinto al de otros; este análisis se lo expondrá al aplicar la técnica FDA sobre los datos, recordando lo mencionado en la sección 2.3.

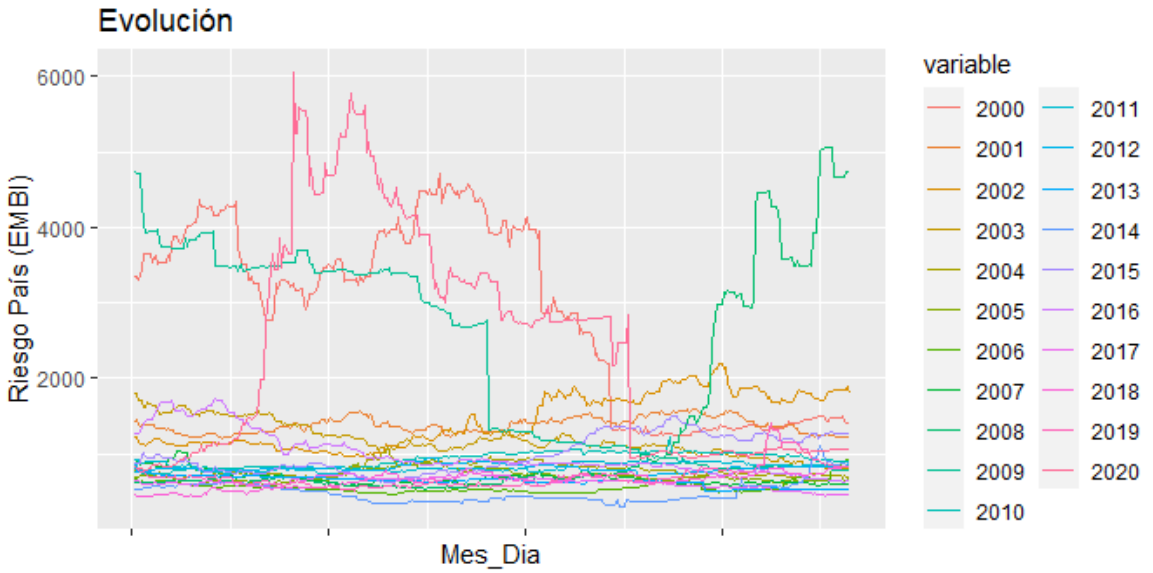


Figura 4.2: Evolución del EMBI anual
Elaboración: Autor

4.2. Representación a datos funcionales

Recordemos que la estructura de las puntuaciones diarias del EMBI por año variaban según eventos característicos que han beneficiado o afectado la economía del Ecuador, siendo el más común el incremento o decremento de los precios del petróleo. Así, para estudiar la forma de los EMBI anuales, es deseable hacerlo mediante una representación funcional para resaltar ciertos patrones y variaciones en los datos, además de analizar su comportamiento mediante la primera y segunda derivada, y poder pronosticar una función completa.

Del registro de los datos filtrados y estructurados en una matriz, en el software `R` trabajaremos con el objeto `fdata`, utilizando los 365 puntos discretos por cada año, y así poder ajustar a los datos a través de funciones base para obtener la muestra funcional de 21 curvas.

Entonces, se hará uso de una base B-spline de orden 4 (splines cúbicos), debido a que, como se ve en la Figura 4.2, los datos no presentan periodicidad, es decir, no hay un patrón fijo en el comportamiento de los EMBI anuales. Luego, para determinar de forma adecuada el número de bases K a utilizar, se aplica el método de minimizar el criterio GCV implementado en la función `optim.basis`, donde además, en este procedimiento se incluye el parámetro de penalización λ, para también determinar el más conveniente. El criterio viene dado por el promedio de los valores GCV de

los datos funcionales, los cuales para cada $i = 1, ..., n$, se definen por:

$$GCV_i(K, \lambda) = \frac{1}{m} \sum_{j=1}^{m} \left(x_{ij} - \hat{X}_i(t_j)\right)^2 \cdot \left(1 - \text{tr}\left(S_\lambda\right)/m\right)^{-2}$$

Así, dentro de la función, se tomaron dos conjuntos de búsqueda: uno que representa el número de bases y otro los parámetros de penalización, para determinar el par más adecuado en el ajuste de los datos. Estos valores se pueden observar en la Tabla 4.4.

Parámetros	**Valores**						
K	7	9	11	13	15	17	19
	21	23	25	27	29	31	33
	35	37	39	41			
λ	0	0.25	0.3252	0.4229	0.5501	0.7156	0.9308
	1.2107	1.5748	2.0484	2.6644	3.4656	4.5078	5.8633
	7.6266	9.9200	12.9032	16.783	21.831	28.395	36.935
	48.041	62.488	81.279	105.72	137.52	178.87	232.66
	302.62	393.63	512				

Tabla 4.4: Conjuntos de valores K y λ

Luego de aplicar el criterio sobre los datos con los distintos valores de K y λ, se elige el par con el menor GCV calculado; dentro del procedimiento se ajustan los datos funcionales utilizando cada par a través de las bases B-spline según K, para luego determinar el menor GCV. En la Figura 4.3, donde horizontalmente se presentan los K y verticalmente los GCV calculados, designados por colores según los valores de λ, se tiene que se alcanza un valor mínimo de $GCV = 4121{,}84$, con valores factibles para el suavizamiento dados por:

$$K = 37 \qquad \text{y} \qquad \lambda = 7{,}6266$$

Del gráfico, se puede notar que para todo $K \leq 15$ del conjunto, los pares (K, λ_i) con $i = 1, ..., 31$, calculan el mismo GCV, por lo cual, el ajuste de los datos no varía demasiado; y, a partir de otros valores de K, con las combinaciones de λ, va mejorando el suavizamiento hasta determinar los valores de K y λ que reduzcan el GCV.

4.2.1. Ajuste con base B-spline

Como vamos a trabajar con funciones base de un conjunto B-spline de orden 4, y se determinaron el número de bases y el parámetro de penalización más adecuados,

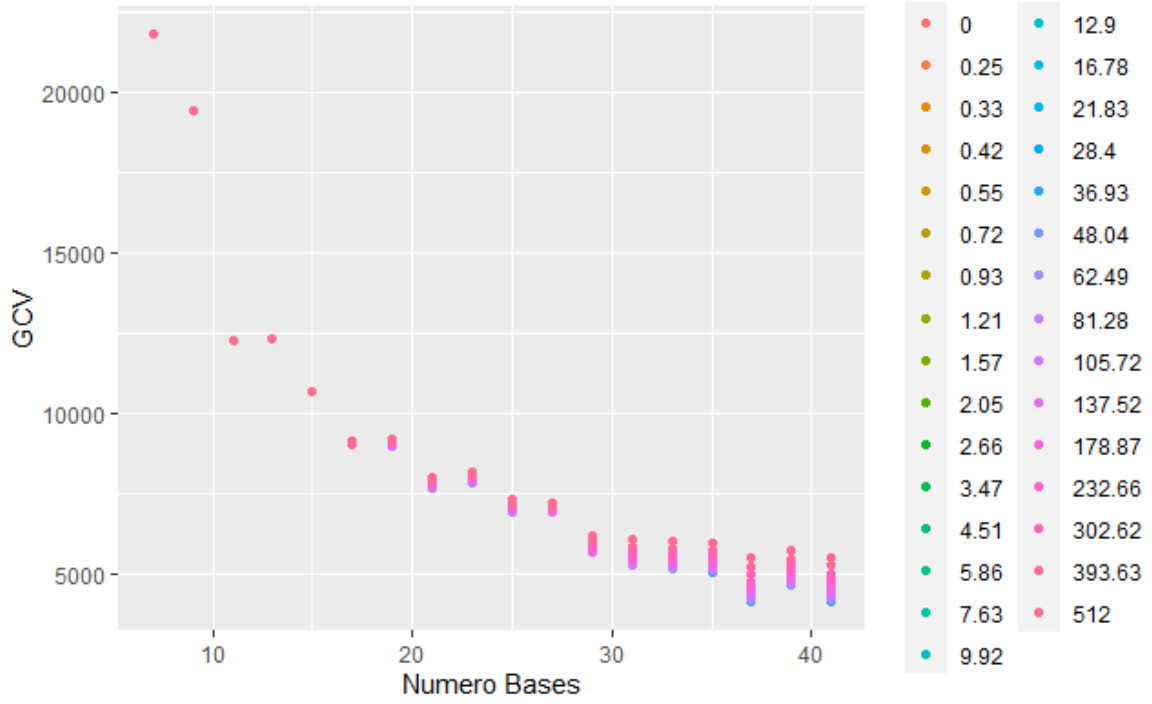


Figura 4.3: Valores GCV según K y λ dados
Elaboración: Autor

se obtienen las $K = 37$ funciones base ϕ_k; estas funciones se las construyen en el intervalo $\mathcal{T} = [1, 365]$ utilizando $L = 33$ puntos de ruptura $\{\tau_\ell\}_{\ell=1}^{L}$, es decir, al intervalo $\mathcal{T}$ se lo divide en 34 subintervalos. Cada punto de ruptura dispondrá de un único nodo, el cual tiene el mismo valor τ_ℓ; con esto, como lo visto en la subsección 3.2.2, se determinan las funciones B-spline $B_{k,m}(t)$ de orden $m = 4$ (grado 3), para todo $k \in \{1, ..., K\}$.

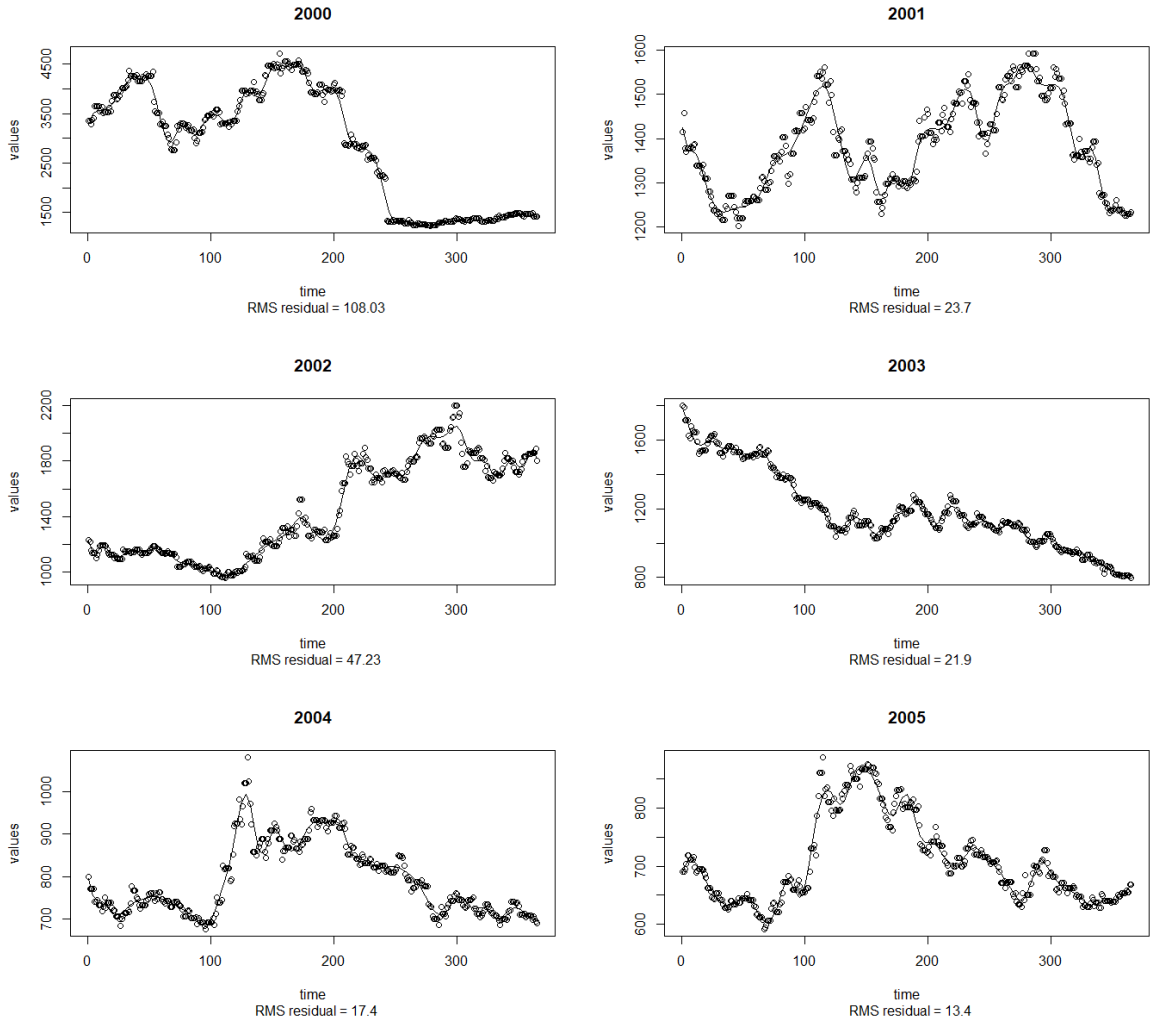
Cabe mencionar que el número de funciones base, el orden y los nodos internos están estrechamente relacionados para la construcción de las funciones, puesto que $K = m + L$ (Ramsay et al. [2009]); así, aunque no se dispongan explícitamente las funciones B-spline, se puede calcular el valor de $B_{k,m}(t)$ para cualquier $t \in \mathcal{T}$ y, en consecuencia, determinar la matriz $\mathbf{\Phi} \in \mathbb{R}^{365\times37}$ para el ajuste de los datos funcionales, utilizando el modelo (3.8). Entonces, en `R`, se hace uso de la función `S.basis` para calcular la matriz de suavizamiento S y, posteriormente, utilizar el criterio (3.10) para obtener las funciones de cada EMBI anual, es decir, $X_1(t) := 2000_t$, $X_2(t) := 2001_t$, $\cdots$, $X_{21}(t) := 2020_t$ y establecer los datos funcionales de la variable funcional $\mathcal{X} =$ EMBI.

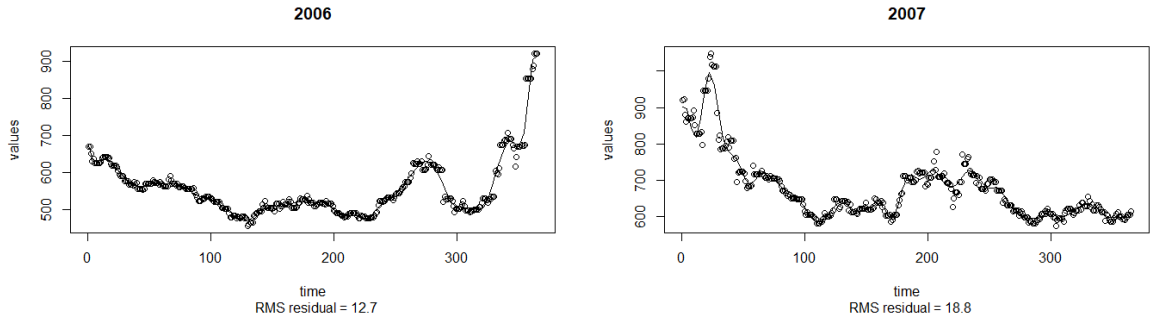
Así, se ajustaron las curvas de las puntuaciones del EMBI para cada año, considerando un periodo de tiempo diario. En la Figura 4.4, se presentan los ajustes individuales, donde para cada $i = 1, 2, ..., 21$, se calcula la medida de la raíz del error

cuadrático medio, dada por:

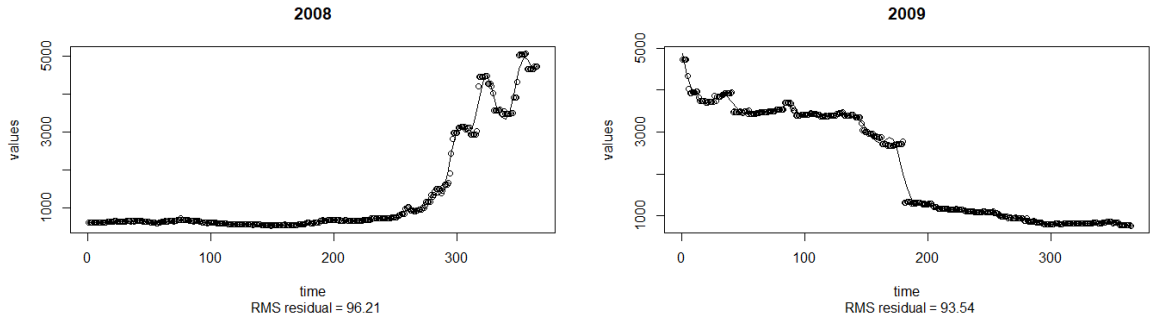
$$RMS_i = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(x_{ij} - X_i(t_j)\right)^2}.$$

Con esto, se tiene que para los años 2000, 2008, 2009 y 2020 existe una mayor dispersión en los residuos, por lo que los puntos de datos se alejan más de la curva ajustada. Los años 2001, 2002, 2003, 2014, 2016, 2019, tienen un *RMS* considerablemente bajo al de los cuatro años anteriores y para el resto de años se tiene un mejor ajuste. Sin embargo, cabe recordar que no se requiere de un ajuste perfecto, sino de una curva suave que explique el comportamiento intrínseco de los datos discretos, reduciendo el ruido existente en los datos originales, para luego realizar el estudio estadístico sobre las funciones.

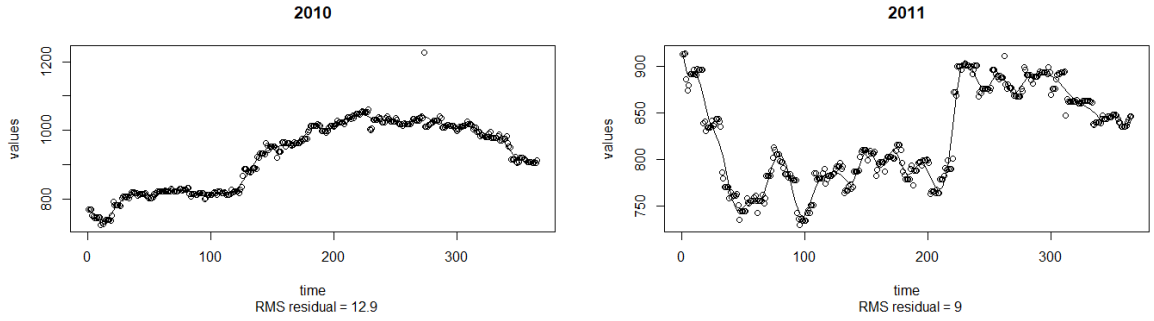



2006
values
time
RMS residual = 12.7

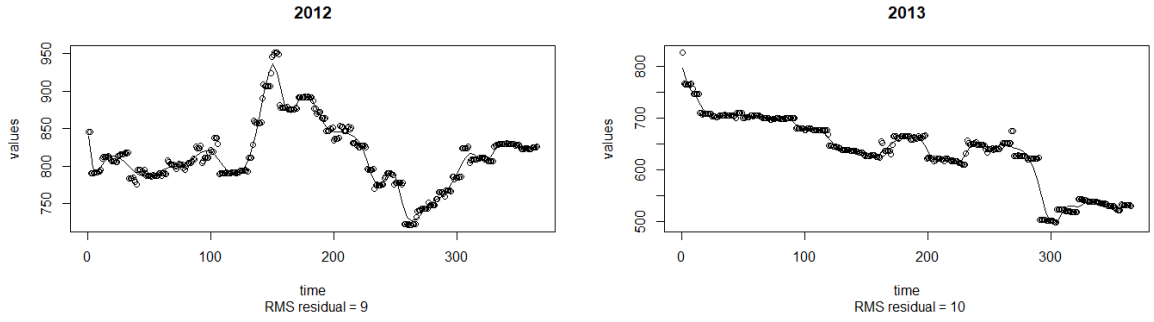
2007
values
time
RMS residual = 18.8


2008
values
time
RMS residual = 96.21

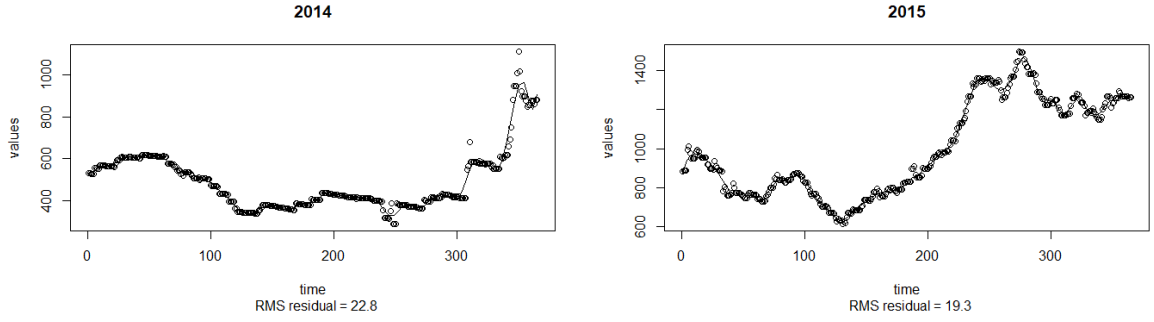
2009
values
time
RMS residual = 93.54


2010
values
time
RMS residual = 12.9

2011
values
time
RMS residual = 9


2012
values
time
RMS residual = 9

2013
values
time
RMS residual = 10


2014
values
time
RMS residual = 22.8

2015
values
time
RMS residual = 19.3

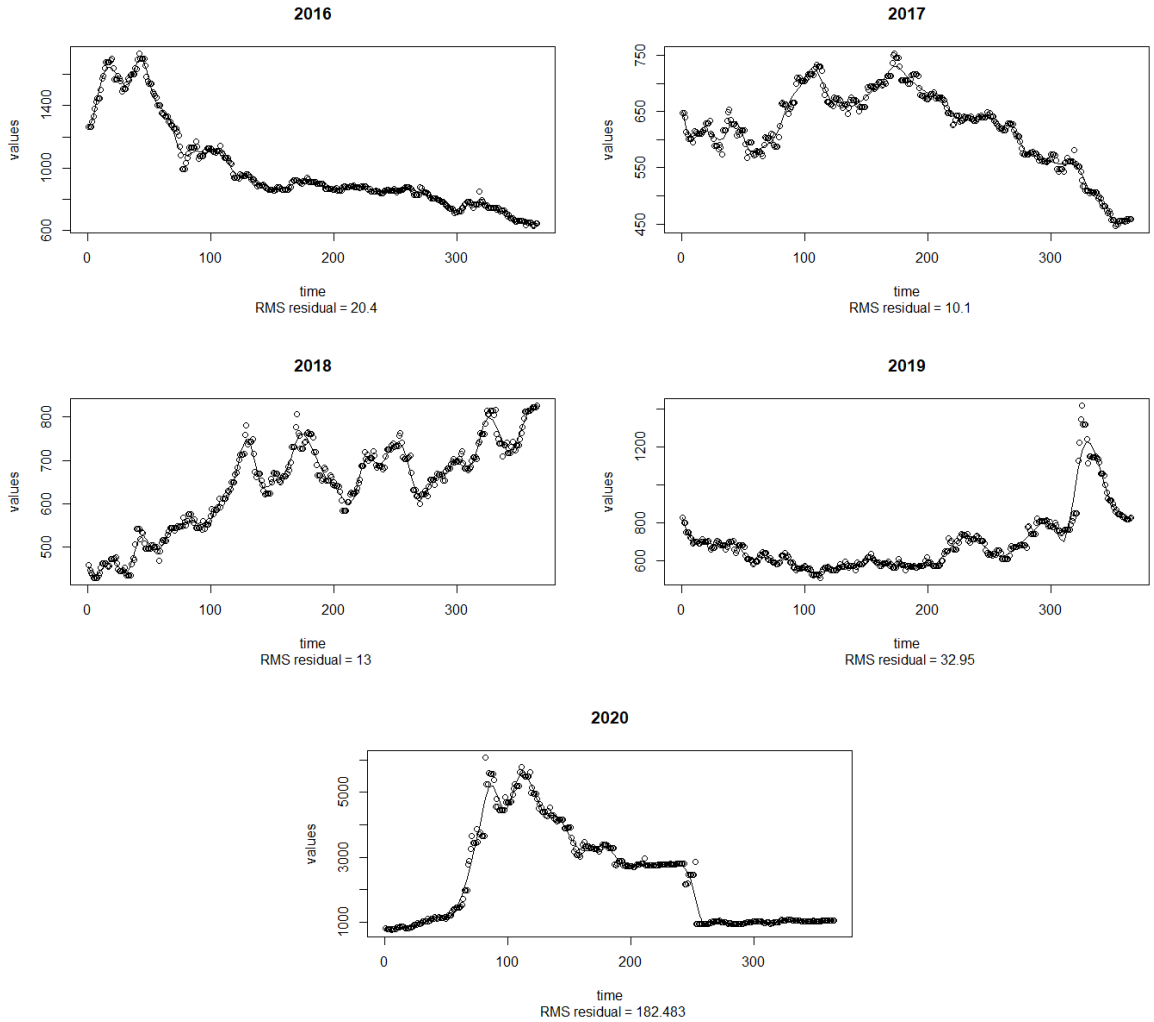


Figura 4.4: Ajuste de datos funcionales
Elaboración: Autor

Para culminar con esta subsección, una vez observado el procedimiento y la facilidad de los cálculos en R para consolidar los datos funcionales, en la Figura 4.5 se tiene el conjunto de datos de la $v.f.\mathcal{X}$, donde se observa notablemente que no todos los datos tienen un mismo comportamiento, con lo cual en la siguiente sección comenzaremos a realizar las distintas técnicas estadísticas en su versión funcional.

4.3. Análisis exploratorio funcional

Como se ha ido abordando a lo largo del proyecto, con la estructura funcional de los datos se pueden encontrar otras características en los mismos y analizar de mejor manera la variabilidad y patrones existentes, dado que en un contexto multivariante no se lo puede ver o hacer fácilmente. Así, a continuación se presenta un análisis

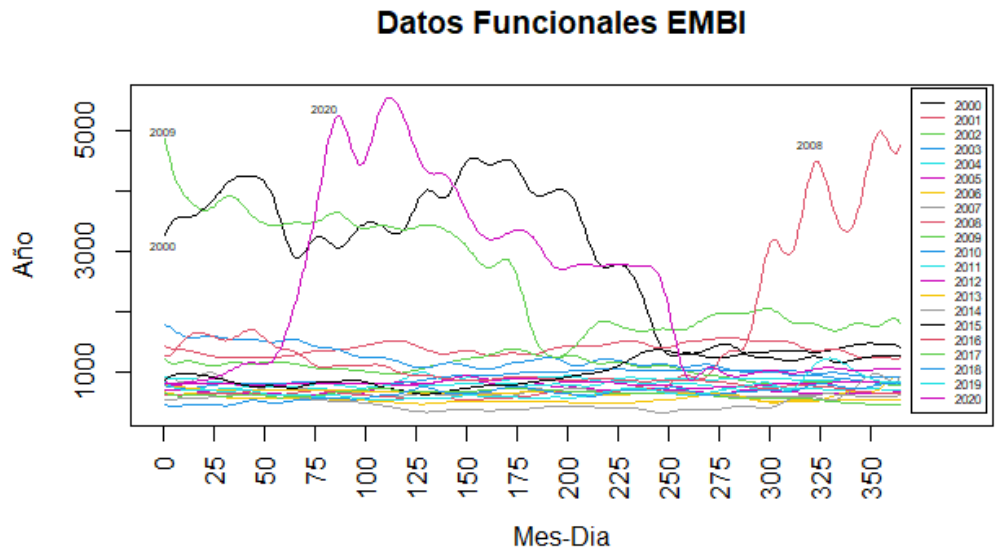


Figura 4.5: Datos Funcionales del EMBI por año
Elaboración: Autor

previo sobre la data funcional para luego utilizarla en el modelamiento y obtener buenos resultados en los pronósticos.

4.3.1. Área de las curvas

Una forma inicial para comprender de igual manera cuan distintos son los datos funcionales es a través del cálculo de sus áreas, ya que representan un valor o magnitud de las curvas; aunque no se trate explícitamente su estructura. Entonces, puesto que no se ha determinado de forma explícita la función $X_i(t)$ de cada dato, pero se disponen de puntos t que se evalúan en la función, en el presente proyecto se utilizó el *Algoritmo de la regla de Simpson* para aproximar dichas áreas (integración numérica), involucrando la función `eval.fd` del paquete `fda`. Este algoritmo se basa en la siguiente fórmula recursiva:

$$\frac{h}{6}\sum_{k=1}^{m}\left[f(x_{k-1}) + 4f\left(\widetilde{x}\right) + f(x_k)\right] \qquad \text{con } \widetilde{x} = \frac{1}{2}(x_k + x_{k-1})$$

con h el salto de x_{k-1} a x_k donde, en nuestro caso, los x_k representan los tiempos t_j, el salto $h = 1$ y la función $f = X_i$, para cada $i = 1, ..., 21$.

Así entonces, en la Figura 4.6 se puede colegir que en el 2000, 2009 y 2020 se han tenido puntuaciones muy elevadas del EMBI, debido a que el área de sus curvas ajustadas tienen un valor muy grande. Por otra parte, es interesante observar que

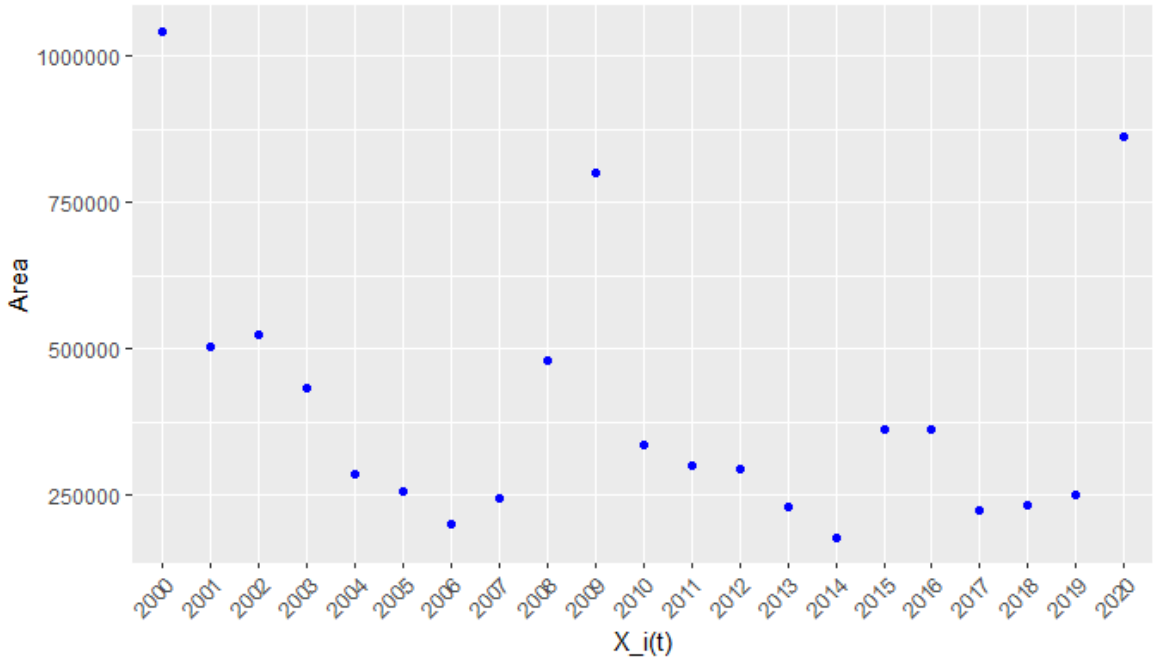


Figura 4.6: Áreas aproximadas de los datos funcionales
Elaboración: Autor

para el 2008, año en el que EE.UU. sufrió una crisis financiera grave, se tiene un área como en 2001 y 2002, esto se lo puede relacionar a que para el 2008 las puntuaciones crecieron abruptamente en el último trimestre, mientras que para el 2001 y 2002 las puntuaciones diarias eran altas pero no superaban los 2000 pb en todo el año. También, se observa que 2006 y 2014 son los años con menor área, de donde se intuye que en dichos años se han registrado las puntuaciones más bajas. Así, es importante aclarar que el área no nos proporciona una idea fija de la estructura de los datos funcionales, sino una medida explicativa de cuan altos o bajos han sido los valores de las funciones, en este caso, las puntuaciones del EMBI.

4.3.2. Primera y Segunda derivada

Al trabajar con funciones suaves, que fue lo que se trató en la sección anterior, se tiene la posibilidad de utilizar la información respecto al cambio o variación que ocurren en las mismas, y esto se lo puede ver a través de las derivadas, esencialmente, en la primera y segunda, pues recordando resultados fundamentales del cálculo, se tiene que:

- A la primera derivada, que representa una expresión de la pendiente en cada punto de la función, se la relaciona con el comportamiento de la función

original, es decir, si es creciente, decreciente o constante en ciertos intervalos.

- A la segunda derivada se la relaciona con la curvatura de la función original, lo que en cierto modo se lo puede ver como un ritmo del crecimiento, es decir, cuando la función pasa de ser convexa a cóncava o viceversa.

Por lo tanto, como ya se construyeron las funciones que representan a los EMBI anuales, en `R` se pueden graficar las derivadas de estas funciones, haciendo uso de la función `fdata.deriv` y, así, poder explicar de mejor manera el comportamiento de las funciones $X_i(t)$.

En la Figura 4.7, se presentan las primeras derivadas de cada función ajustada, de donde al observar cuándo son positivas o negativas, se colige que las funciones originales crecen o decrecen, respectivamente. Así, se puede ver que hay más fluctuación en los años más críticos ya mencionados; en el 2000 (negro) se puede deducir que hay un ascenso y descenso muy variable de las puntuaciones entre los meses de febrero y agosto (día 32 a 245), luego se aprecia que esa variación desaparece notablemente. Para el 2008 (rojo), desde finales de agosto las puntuaciones del EMBI crecen demasiado, aunque en un corto periodo del mes de noviembre (día 325) se logra un descenso, y para inicios del año 2009 (verde) hasta mediados de julio (día 200), estas puntuaciones disminuyen (su derivada se mantiene negativa), para luego tener un leve aumento y mantenerse en ese nivel, ya que en la primera derivada se distingue un comportamiento casi constante. En el 2020 (morado), es notorio ver que entre el día 50 y 80, correspondiente a febrero y marzo, se tiene un aumento considerable del EMBI, y posterior a ello se tiene una variación como en el 2000, hasta que por el mes de octubre se estabiliza este comportamiento. Para los demás años, se distingue que hay un comportamiento leve de ascenso y descenso de las funciones; los periodos de los años más notables que salen de ese conjunto son: los últimos bimestres del 2006 (amarillo), 2014 (gris) y 2019 (celeste), pero aún así se alejan de lo que ocurre en los años críticos.

Ahora, de la Figura 4.8, correspondiente a las segundas derivadas de cada dato, se puede analizar de mejor forma los cambios bruscos de las funciones, pues si la segunda derivada es positiva la función original tiene un comportamiento convexo, caso contrario es cóncava; por tal motivo, a esta derivada se la suele relacionar con la curvatura de la función. Entonces, se puede ver que de los períodos de tiempo anteriormente mencionados en los años críticos, se tiene que la estructura de estas funciones son muy cambiantes por aproximadamente cada 10 días, pasando de convexas ($X_i'' \geq 0$) a cóncavas ($X_i'' \leq 0$) y viceversa ($i = 1, 9, 10, 21$) por lo cual,

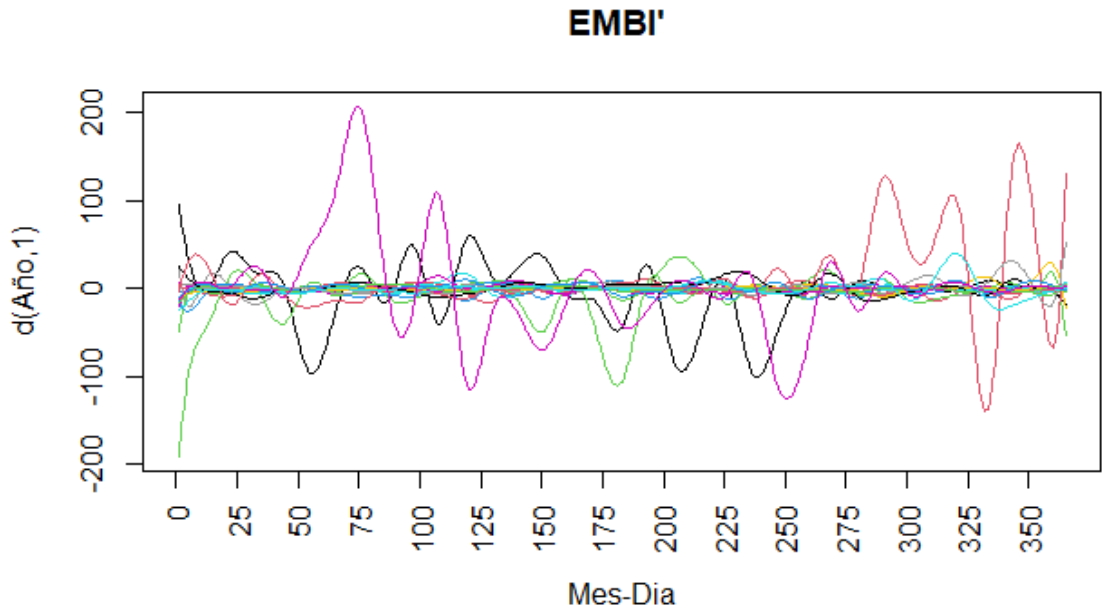


Figura 4.7: Primera derivada de los datos
Elaboración: Autor

ya se logra intuir que dichos datos pueden afectar en nuestro modelo a crear; en la siguiente sección se tratarán a estos datos. Respecto a las demás funciones, aunque sí existen estos cambios, no se presentan en períodos de tiempo cortos.

Una vez comprendido de mejor manera las variaciones que existen en los datos a través de sus derivadas; a continuación, se estudia qué datos funcionales son atípicos dentro de la muestra.

4.3.3. Detección de datos funcionales atípicos

Como se ha ido observando, cuatro datos funcionales son los que se alejan del comportamiento normal de la muestra; por tanto, se puede pensar que para algún estudio posterior, utilizar esta información podría causar algún sesgo, por ejemplo, en la estimación de la media funcional. Entonces, para detectar los datos funcionales atípicos, se utilizarán las dos medidas de profundidad explicadas en la sección 3.4 (Fraiman & Muniz y Modal), para en primer lugar, comparar cuál de las dos medidas detecta más atípicos y luego de elegir una de las dos, retirar de la muestra estas observaciones para reducir la influencia de esta contaminación a estudios posteriores.

Así, se presenta lo que calculan ambas medidas de profundidad al compilar una

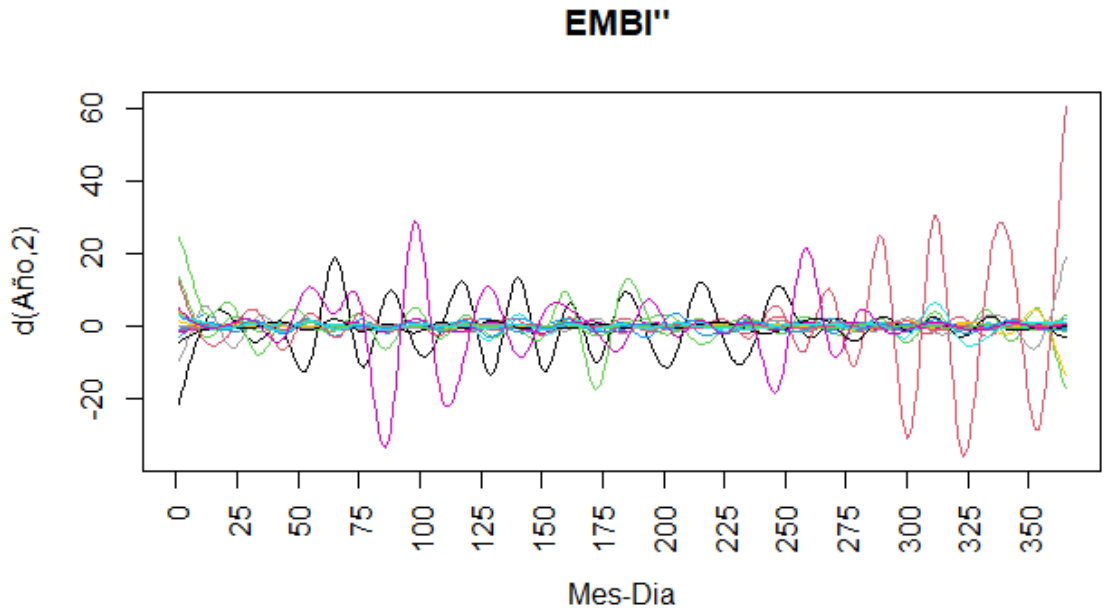


Figura 4.8: Segunda derivada de los datos
Elaboración: Autor

sola vez sus algoritmos y corroborar si los resultados son adecuados.

Profundidad FM

Recordemos que esta medida utiliza la función de distribución empírica de cada $t = 1, ..., 365$ del $i-$ésimo dato funcional para calcular la profundidad FM del dato i (FMD_{X_i}). En `R`, se dispone de la función `depth.FM`, la cual determina la mediana funcional y, de paso, la media funcional $\alpha-$recortada. Así, tomando un $\alpha = 0{,}15$, de la Figura 4.9, se tiene que la mediana funcional ($\hat{\mu}_{FM} = FM.med$) corresponde a la curva del 2012, mientras que la media funcional robusta ($\hat{\mu}_{0{,}15,TM} = FM.tr15\,\%$) se la calculó considerando todos los años excepto el 2000, 2006 y 2014, ya que son las curvas menos centradas por este criterio (profundidad baja) y, en principio, se los considerarían como datos atípicos.

Profundidad MD

Para esta medida, la cual extiende el concepto de moda, recordemos que se utilizan funciones Kernel para que de un dato i se ponderen los demás datos funcionales, y así determinar la profundidad MD_{X_i}. En `R`, se dispone de la función `depth.MD` la cual, similar a lo anterior, calcula la moda (o indistintamente la mediana) funcional

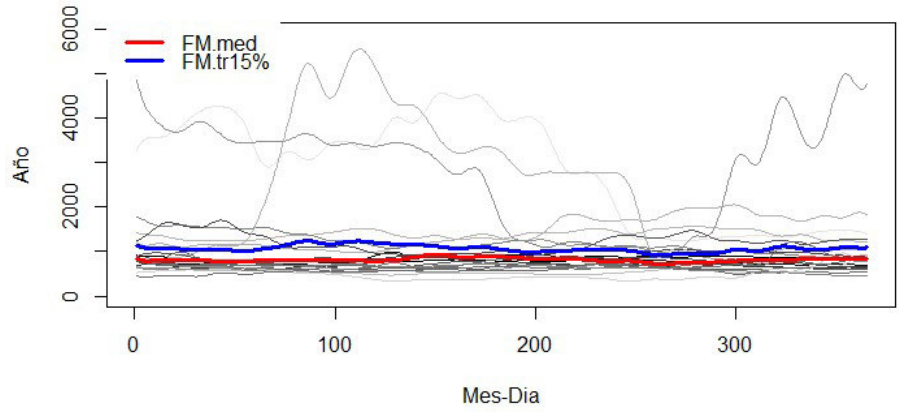


Figura 4.9: Profundidad Fraiman & Muniz
Elaboración: Autor

y la media funcional $\alpha-$recortada. Así, para un $\alpha = 0{,}15$, en la Figura 4.10 se tiene que la moda funcional $(\hat{\mu}_{MD} = mode.med)$ corresponde a la curva del 2005, mientras que la media funcional robusta $(\hat{\mu}_{0{,}15,TM} = mode.tr15\,\%)$ se la calculó considerando todos los años excepto el 2008 y 2020, ya que son las curvas con menor profundidad con lo cual, en principio, se las considerarían como datos atípicos.

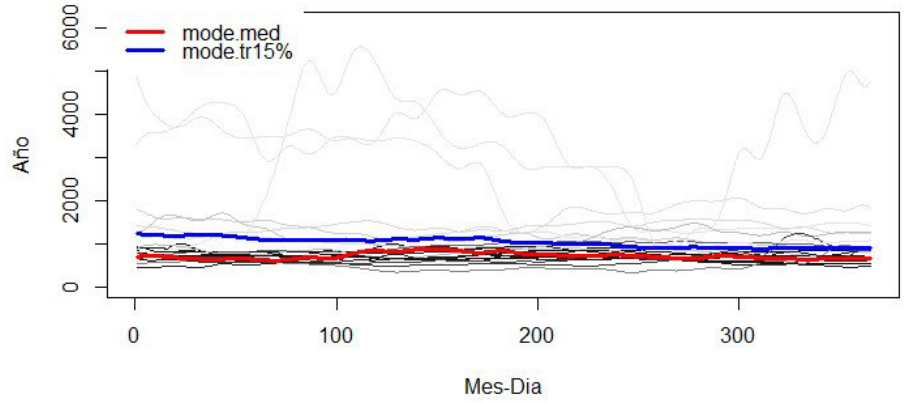
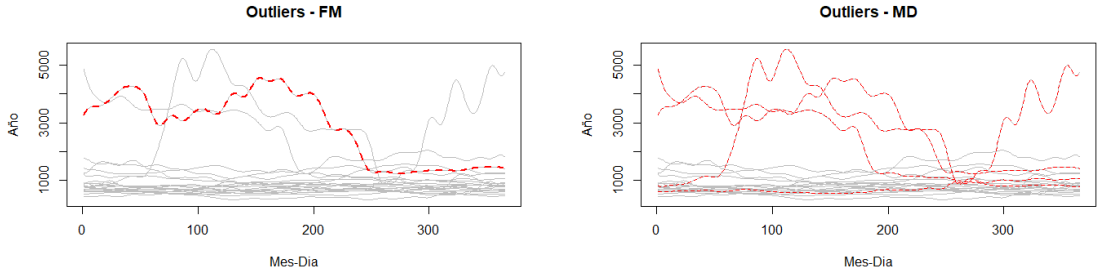


Figura 4.10: Profundidad Modal
Elaboración: Autor

De las dos medidas calculadas, no se logran consolidar a los cuatro datos que se los catalogaron como atípicos o incluso a más, aunque la medida MD determina dos de los candidatos. Así, para precisar en la búsqueda de datos atípicos, se aplica el algoritmo bootstrap descrito en la subsección 3.4.3, basado en las dos medidas de profundidad.

Entonces, utilizando la función `outliers.depth.trim` con 200 muestras, $\alpha = 15\,\%$, $\gamma = 10\,\%$ (parámetro de bootstrap suavizado) y las medidas FM y MD, en la Figura 4.11 se presentan los datos funcionales atípicos que han sido detectados

por cada criterio.



(a) Fraiman & Muniz (b) Modal

Figura 4.11: Detección de datos funcionales atípicos
Elaboración: Autor

Además, en la Tabla 4.5, se muestra una breve comparación de los resultados obtenidos por el algoritmo usando ambas profundidades, de donde se puede afirmar que los años más críticos de las puntuaciones del EMBI, en efecto, son los datos funcionales atípicos de la muestra, siendo la profundidad Modal la más adecuada a considerar.

Característica	FM	MD
No. Atípicos detectados	1	4
Atípicos detectados	2000	2000, 2008, 2009, 2020
Tiempo ejecución (secs)	36.94	28.12

Tabla 4.5: Comparación de las profundidades *FM* y *MD*

En consecuencia, una vez detectados los datos atípicos (2000, 2008, 2009 y 2020), se los retiran de la muestra, para no sesgar las estimaciones de lo posterior a realizar en el proyecto. Finalmente, en la Tabla 4.6, se tienen las medidas de profundidad calculadas por el algoritmo, de donde además, se observa que el dato más profundo es el del año 2005, y se lo consideraría como la moda/mediana funcional.

Año	MD_{X_i}
2000	0.3989
2001	0.5226
2002	0.4597
2003	0.6589
2004	2.8521
2005	3.1327
2006	2.4447

Año	MD_{X_i}
2007	2.9449
2008	0.3989
2009	0.3989
2010	1.8955
2011	2.8058
2012	2.8787
2013	2.9496

Año	MD_{X_i}
2014	1.5774
2015	0.8963
2016	0.7719
2017	2.8483
2018	2.7616
2019	2.3548
2020	0.3989

Tabla 4.6: Profundidad *MD* de cada dato funcional

4.3.4. Medidas estadísticas funcionales

Para culminar con esta sección, se exhiben las funciones descriptivas vistas en la sección 3.3, las cuales proporcionan información respecto a la centralidad y variabilidad de los datos funcionales.

Respecto a la media, Figura 4.12, se puede concluir que en promedio, las puntuaciones diarias del EMBI varían entre 786 y 906 pb, de donde aún representan puntuaciones altas para que el Ecuador sea una atracción inversionista; sin embargo, al tratarse de un estimador, los valores de esta función media se ven reflejados por las puntuaciones del histórico considerado, siendo muy pocos los años en los que se han registrado puntuaciones bajas.

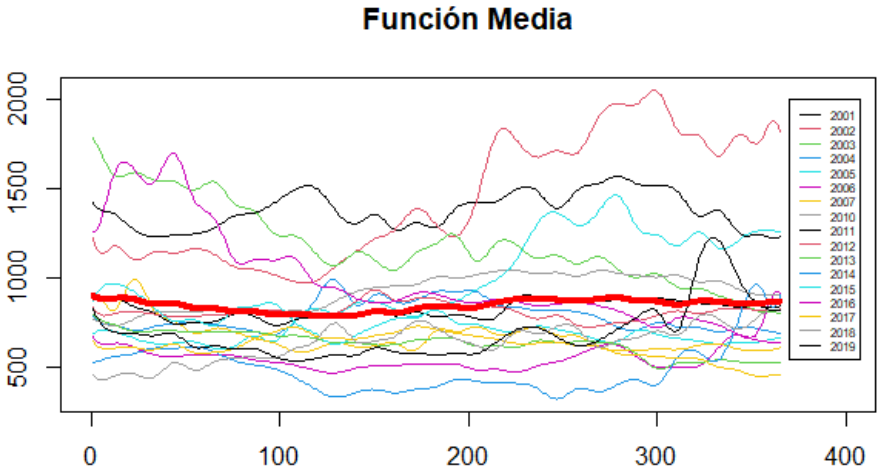


Figura 4.12: Media funcional, $\mu_{\mathcal{X}}(t)$
Elaboración: Autor

De la desviación estándar funcional, Figura 4.13, se puede ver que los datos no se encuentran tan dispersos de la media funcional. De hecho, el valor mínimo de esta función es de 237 pb y el valor máximo es de 405 pb; aunque los valores altos de dispersión se deben a que hay datos funcionales que se alejan en ciertos periodos de tiempo de la media, como por ejemplo en el segundo semestre de los años 2001 (negro), 2002 (rojo) y 2015 (celeste) o en los primeros meses del 2003 (verde) y 2016 (morado).

Otras medidas estadísticas interesantes de calcularlas son la covarianza y la correlación, las cuales brindan información sobre la relación entre los valores de las funciones en t_j y t_k. En la Figura 4.14, se presenta el gráfico de calor de estos dos

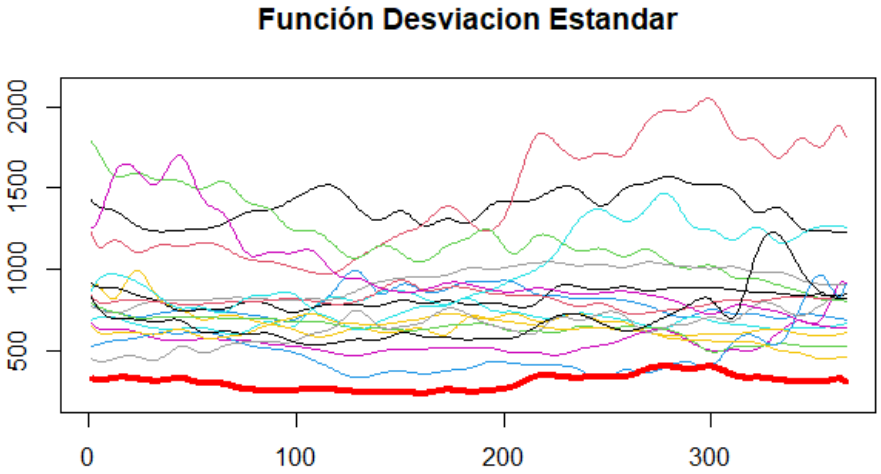
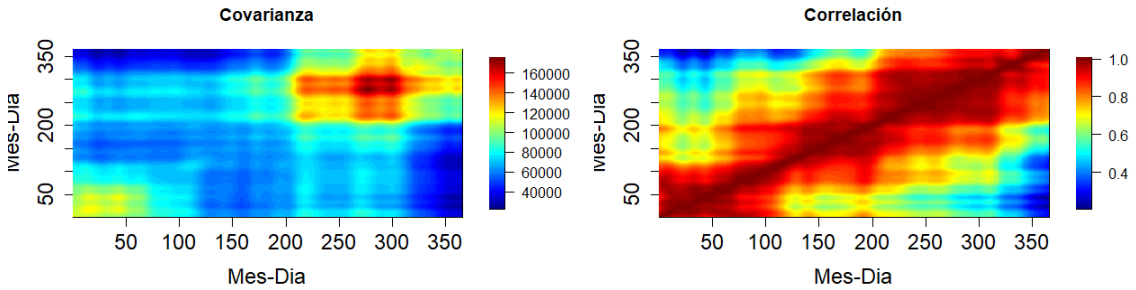


Figura 4.13: Desviación Estándar funcional, $s_{\mathcal{X}}(t)$
Elaboración: Autor

estadísticos; en la covarianza se observa que entre cada día no existe una variación fuerte, a excepción de los días de los últimos meses. Una forma más útil de observar esta relación es por medio de la correlación, la cual en la gráfica, se puede ver que existe una fuerte correlación entre cada trimestre, debido a que entre esos días no existen cambios bruscos en las curvas. En cambio, se puede ver que hay muy baja correlación entre los días del mes de diciembre y enero, esto se lo puede relacionar porque la mayoría de curvas en diciembre tienen puntuaciones altas, mientras que en enero esas puntuaciones son reducidas.



(a) $\nu_{\mathcal{X}}(t_1, t_2)$ (b) $\rho_X(t_1, t_2)$

Figura 4.14: Covarianza - Correlación
Elaboración: Autor

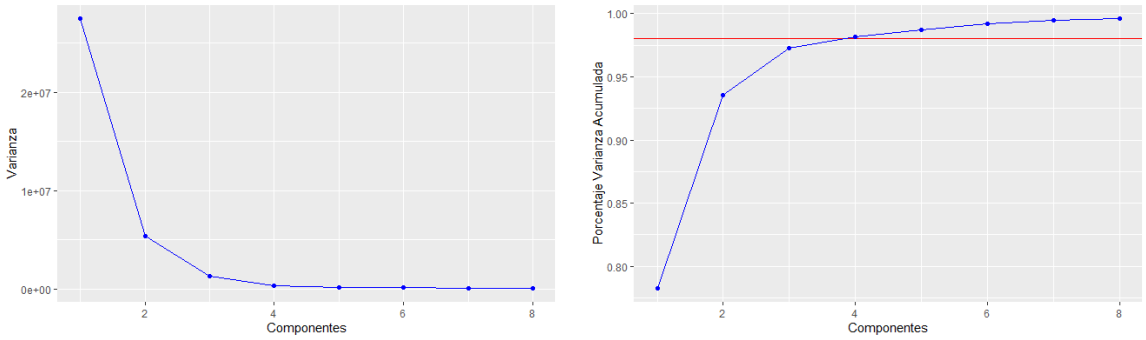
Sin embargo, como en el caso multivariante, es complicado observar toda la información presente en estos estadísticos por la gran cantidad de datos, y no se puede establecer un perfil adecuado sobre la estructura de la variabilidad en los mismos,

por ello para el caso funcional se puede utilizar la técnica de componentes principales y así comprender de mejor manera las variaciones y patrones existentes en los datos.

4.4. Aplicación de la técnica FPCA

Ahora, de la base de datos funcional consolidada, para analizar los modos de variación que se encuentran en los datos, se aplica el Análisis de Componentes Principales Funcionales, el cual se lo abordó en la sección 3.5. Para esto, se utiliza la función `pca.fd` del paquete `fda`, donde la misma calcula todo lo esencial para este análisis mediante los métodos matriciales, utilizando el análisis de autofunciones sobre el operador de covarianza $\mathcal{V}$ para determinar las funciones de peso $\alpha_j(t)$ definidas sobre el mismo dominio de los datos funcionales.

Entonces, en primer lugar, para determinar el número de componentes principales funcionales a utilizar, en la Figura 4.15(a) se presenta el gráfico de sedimentación según la variabilidad explicada por cada $\alpha_j(t)$ con $j = 1, ..., 8$, de donde se puede ver que a partir de la cuarta, el gráfico tiene un comportamiento casi constante, por lo que el uso de cuatro componentes resultaría adecuado para el análisis. De hecho, en el gráfico 4.15(b), correspondiente al criterio *cut-off* con un corte del 98 %, se observa que hasta la tercera componente se acumula dicha proporción de variabilidad; sin embargo, no hay demasiada diferencia de acumulación hasta una cuarta componente. Por lo tanto, se realizará el análisis para las cuatro primeras componentes principales funcionales.



(a) Varianza de cada FPC (b) % Varianza Acumulada

Figura 4.15: Varianza de las FPC
Elaboración: Autor

Luego, de la Tabla 4.7, se observa que en la primera FPC se explica un buen

porcentaje de la variabilidad de los datos funcionales, es decir, es la componente que más explica la tendencia o patrones subyacentes en los datos; sin embargo, el análisis del comportamiento de las tres componentes siguientes es interesante observarlo, ya que pueden contener ciertas fluctuaciones que en la primera FPC no se explican, pese a que recogen menos variabilidad.

Componentes Principales	**Varianza explicada %**
FPC1	78,2277
FPC2	15,3500
FPC3	3,7122
FPC4	0,8778
Total	98,1671

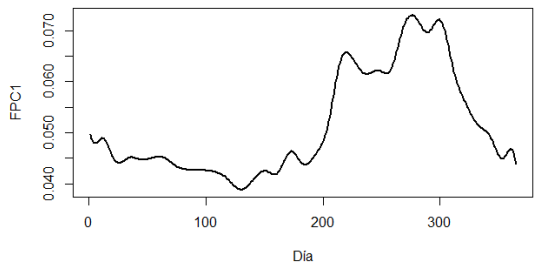
Tabla 4.7: Variabilidad explicada por cada FPC

Así, se muestran las funciones que representan las componentes principales $\alpha_j(t)$, con $j = 1, ..., 4$ (Figura 4.16). Además, en la Tabla 4.8 se encuentran los valores de las puntuaciones $f_{j,i}$ para cada dato, los cuales se los puede catalogar como el aporte del dato en la formación de la componente principal y, mediante esas puntuaciones se determina la varianza recogida por cada componente, como lo mostrado en la gráfica 4.15(a).

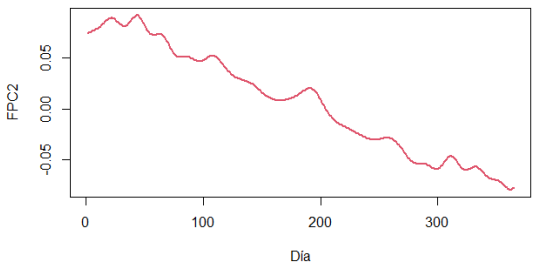
	$\alpha_1(t)$	$\alpha_2(t)$	$\alpha_3(t)$	$\alpha_4(t)$
2001	10.209,38	844,78	-1.311,17	351,69
2002	12.208,19	-4.237,82	763,48	198,89
2003	6.054,68	4.954,10	512,17	-32,09
2004	-1.172,00	203,92	-1.642,79	119,80
2005	-2.753,81	31,57	-1.483,97	91,14
2006	-5.515,34	-744,88	780,11	-254,01
2007	-3.329,89	876,81	408,66	-524,12
2010	1.662,19	-1.025,93	-1.040,71	-5,53
2011	-349,97	-368,99	-76,04	-171,49
2012	-759,08	248,06	-731,45	610,40
2013	-4.114,86	835,11	-286,53	-598,47
2014	-6.889,83	-760,38	1.839,96	747,83
2015	3.468,48	-3.031,63	1.416,18	-1.377,22
2016	2.246,86	5.394,92	1.609,74	36,48
2017	-4.357,93	364,60	-1.221,79	-535,62
2018	-3.816,16	-1.812,63	-909,53	289,87
2019	-2.790,90	-1.771,62	1.373,70	1.052,46
Varianza	27’451.791,18	5’409.474,98	1’302.129,93	307.191,32

Tabla 4.8: Puntuaciones de los datos funcionales en cada FPC

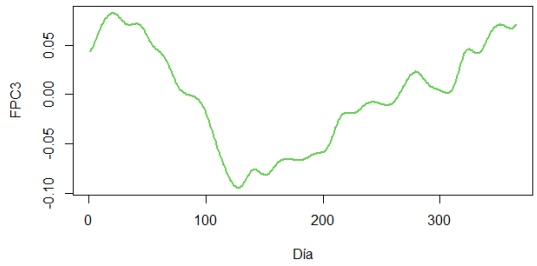
OBSERVACIÓN. La última fila de la Tabla 4.8 representa la varianza explicada por



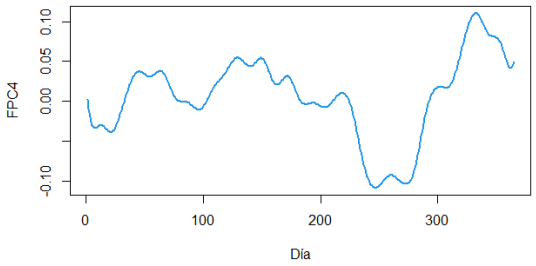
(a) Primera componente principal funcional



(b) Segunda componente principal funcional



(c) Tercera componente principal funcional



(d) Cuarta componente principal funcional

Figura 4.16: Componentes Principales Funcionales
Elaboración: Autor

cada componente, $\text{Var}\langle \alpha_j, \mathcal{X} \rangle$, es decir, es la varianza calculada por la función objetivo del planteamiento (3.17).

Ahora, como se había abordado, una forma de dar una interpretación a las componentes es a través de perturbaciones respecto a la media de los datos funcionales. Además, para facilitar esta interpretación, se grafican las contribuciones o aportes de los datos funcionales en la formación de las componentes; para esto, en lugar de utilizar directamente las puntuaciones dadas en la Tabla 4.8, se calcula la proporción de la contribución de cada dato i sobre la componente principal j, es decir, se calculan:

$$\frac{f_{j,i}^2}{n \cdot \text{Var}\left\langle \alpha_j, \mathcal{X} \right\rangle} = \frac{f_{j,i}^2}{\sum_{i=1}^{n} f_{j,i}^2}.$$

Con esto, a continuación vamos a analizar las cuatro FPC de los datos funcionales correspondientes a los EMBI anuales, tomando como referencia la Tabla 2.1 y lo visto en la subsección 2.3. Este análisis se basará en el efecto de las perturbaciones de la media y el aporte de los datos funcionales respecto a su variabilidad en la definición

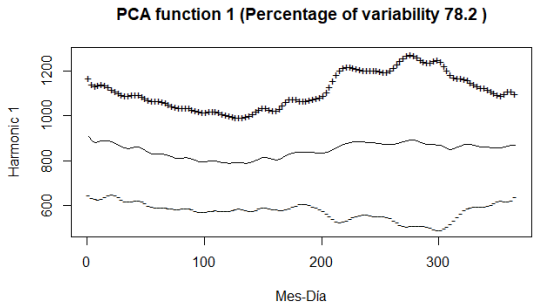
de las componentes:

- **FPC1.** De la estructura de la primera FPC (Figura 4.16(a)) se puede ver que es positiva en todo el periodo de tiempo considerado, la cual comúnmente se la interpreta como un promedio ponderado de los datos funcionales (EMBI anuales), vista también como el *tamaño* de los datos. Por ello, de la Figura 4.17(a), se observa que hay una separación notable entre la media (línea negra sólida), lo que significa que presenta la mayor variabilidad de los datos, como ya se vio. Sin embargo, a partir de mediados del mes de julio (día 200) hasta los últimos días de octubre (día 300) hay una mayor separación entre la media, es decir, la componente brinda más peso a este lapso de tiempo, y esto puede deberse a algún fenómeno más grave que alteró el comportamiento de los meses anteriores a julio.

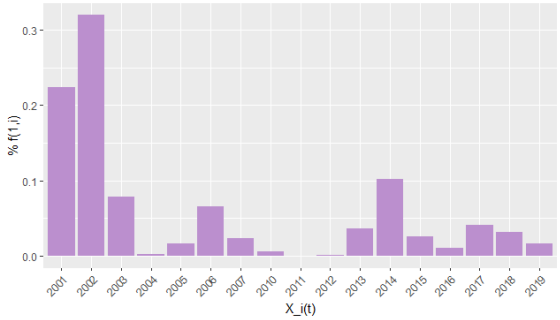
  De hecho, observando la Figura 4.17(b), se tiene que los datos funcionales correspondientes a los años 2002, 2001, 2014 y 2003, respectivamente, son los datos que más se han involucrado en la construcción de esta FPC, pues en el 2001, 2002 y 2003 se tenían puntuaciones altas de este indicador, por estragos de la dolarización, y en ciertos periodos de tiempo, sobre todo entre junio y octubre, aumentaron más por situaciones de inestabilidad políticas y económicas. En 2014, por otro lado, fue un año en donde se registraron las puntuaciones más bajas del EMBI, llegando a un mínimo de 289 pb en septiembre; por lo cual, con esta información también existe una variación considerable. Un caso similar al 2014, ocurre con la aportación del año 2006.

- **FPC2.** Del gráfico 4.16(b), se puede distinguir que hay un cambio entre el primer y segundo semestre. En la Figura 4.18(a), claramente se observa este viraje a partir de mediados de julio (día 200), producto del incremento en las puntuaciones del EMBI, por alguna razón como la falta de acuerdos con el FMI; es importante notar también que esta componente recoge una mayor variabilidad entre el primer y último trimestre, es decir, se tiene más explicación en la tendencia de esos intervalos de tiempo.

  De la Figura 4.18(b), se tiene que los años 2016, 2003 y 2002, respectivamente, tienen un alto porcentaje de contribución en esta componente, en particular, el 2016 y 2003 estarían aportando en los primeros meses, ya que estas curvas presentan valores más distantes a la media en dicho periodo y el 2002, tendría un efecto mayor entre noviembre y diciembre, al igual que el 2015.



(a) Perturbación de la media



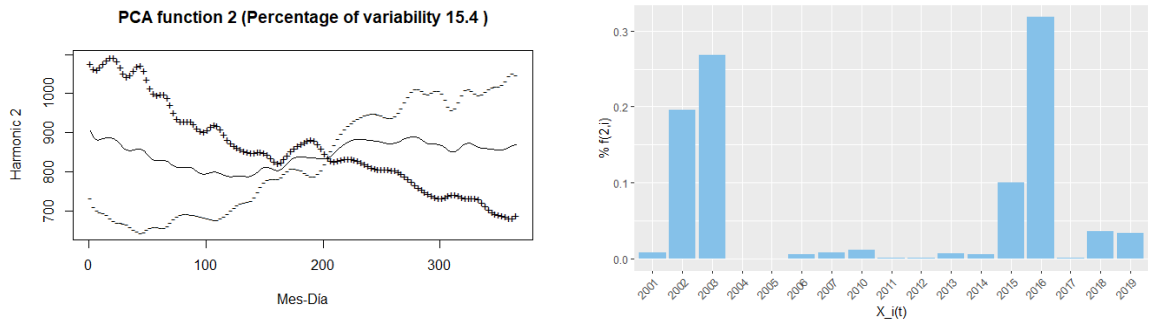
(b) Contribución de los datos funcionales

Figura 4.17: Interpretación de la primera FPC

- **FPC3.** En el gráfico 4.19(a), se observan dos efectos importantes, a finales del primer trimestre y a comienzos del último trimestre, siendo los años 2014, 2004, 2016, 2005 y 2015 los que más contribuyen en la formación de la tercera FPC, respectivamente, Figura 4.19(b).

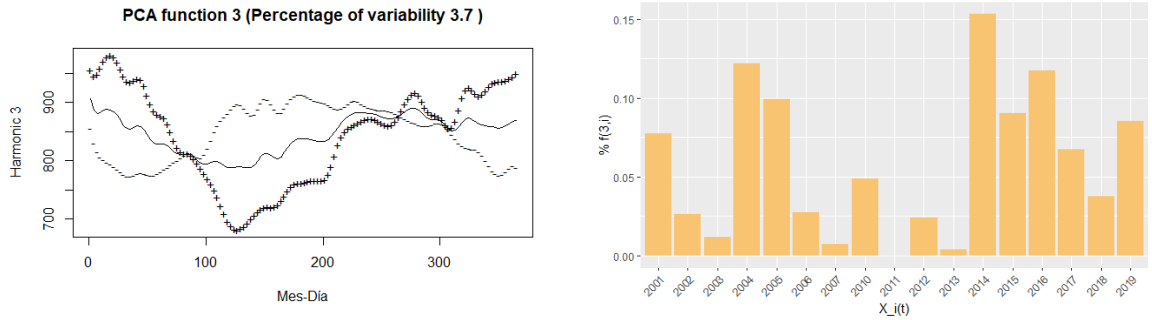
  Como se puede observar, se involucran años donde en ciertos periodos de tiempo no varían demasiado las puntuaciones respecto a la media, pero que aún así recogen cierta variabilidad. También es curioso ver que entre los meses de septiembre y octubre (días 240 a 305) no se explica una gran variación; una razón puede ser, por ejemplo que el mes de septiembre del 2014 ya se consideró en la primera FPC, pues en ese tiempo se registraron las puntuaciones más bajas.

- **FPC4.** En la Figura 4.20(a), se aprecia que existe una marcada sensibilidad en las puntuaciones del EMBI, debido a que los datos funcionales que conforman esta componente ya no presentan una alta varianza respecto a la media, pues la mayor dispersión de estos datos (2015, 2019, 2014 - Figura 4.20(b)) ya han sido expuestos en las componentes precedentes; aunque cabe notar que en esta componente se recoge mayor variabilidad desde el mes de agosto (día 230).



(a) Perturbación de la media (b) Contribución de los datos funcionales

Figura 4.18: Interpretación de la segunda FPC
Elaboración: Autor

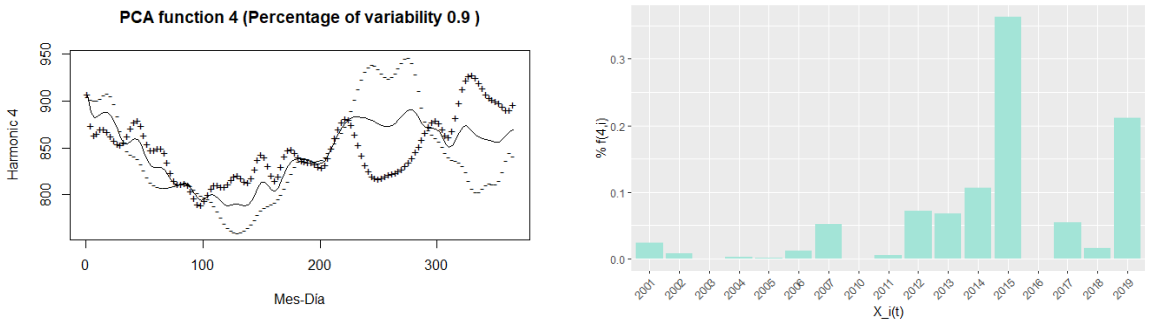


(a) Perturbación de la media (b) Contribución de los datos funcionales

Figura 4.19: Interpretación de la tercera FPC
Elaboración: Autor

Como se pudo observar, a partir de la segunda FPC se pueden interpretar a las mismas como funciones de *forma*, pues presentan valores positivos y negativos que se los relacionan como cambios, rupturas o fluctuaciones entre los datos funcionales debido a algún fenómeno económico que ha hecho que las puntuaciones tengan una tendencia creciente o decreciente, durante ciertos periodos de tiempo. Además, es esencial aclarar que dentro de la explicación dada, solo se han mencionado los datos funcionales que más contribuyen en las componentes; sin embargo, hay otros datos que tienen su efecto, por ejemplo, el 2019 tiene su aporte en α_1, la cual puede relacionarse más con el aumento del EMBI a raíz de las protestas de octubre o el 2015 en α_2, que desde mediados del tercer trimestre presentó aumento en sus puntuaciones por el desplome en los precios de crudo.

Para culminar, una vez expuesto el tratamiento y análisis exploratorio del FDA, utilizando las tres primeras FPC, se desarrolla el modelo con el método HU bajo dos enfoques propuestos, para pronosticar la función completa que describa las puntuaciones diarias del EMBI para el año 2021.



(a) Perturbación de la media (b) Contribución de los datos funcionales

Figura 4.20: Interpretación de la cuarta FPC
Elaboración: Autor

4.5. Desarrollo del modelo basado en FPCA bajo dos enfoques

Ahora, utilizando el modelo propuesto por Hyndman y Ullah [2007], y que se lo estudió en la sección 3.6, vamos a pronosticar la función que describa las puntuaciones diarias del EMBI para el año 2021, basándonos en dos escenarios: el primero, descrito como el mejor caso o el caso normal de las puntuaciones que pueden presentarse en el año; el segundo, descrito como el peor caso, aquel donde pueden haber puntajes muy altos del indicador. Obviamente, aquellos años que tuvieron fuertes alzas del EMBI y que en este capítulo se los identificaron como datos atípicos, no se los considerarán.

Así entonces, para los dos enfoques del modelo (3.22), se utilizará el parámetro $\alpha = 0{,}6$ para el suavizado exponencial sobre las series de las puntuaciones de los datos funcionales en cada componente principal. La razón de elegir dicho parámetro es debido a que si se toma un valor muy alto, se dará más peso a los $X_i(t)$ más actuales, sobre todo al primero. Por otro lado, si se toma un valor muy pequeño, se dará importancia a casi todos los $X_i(t)$, situación que tampoco se desea, debido al contexto de los dos enfoques.

Además, como ya se explicó en el capítulo anterior, se utilizará una medida de profundidad para "ordenar" a los datos funcionales, de donde por lo visto en la subsección 4.3.3, se estableció a la profundidad modal MD como la más adecuada a utilizar, por su mayor aporte en la detección de atípicos; por lo tanto, nos valdremos de las profundidades dadas en la Tabla 4.6. De otro lado, por lo tratado en la sección 4.4, se harán uso de cuatro componentes principales funcionales para los modelos.

4.5.1. Ajuste del modelo y verificación de supuestos

Gracias al paquete `ftsa`, se hace uso de la función `ftsm` para ajustar el modelo (3.22), el cual realiza la descomposición de las componentes principales vistas en la sección anterior. Ahora, como se trata de un modelo poco conocido y utilizado, para verificar los supuestos en los errores del modelo, se realizó lo siguiente:

1. *No autocorrelación:* para cada t, se verifica la no existencia de correlación entre los errores; para ello, utilizamos el test de rachas (Fernández y Cao [2021]), el cual contrasta lo siguiente:

$$H_0 : \text{La muestra es aleatoria;}$$
$$H_1 : \text{La muestra no es aleatoria.}$$

   Es decir, se contrasta si los residuos sucesivos $\{\varepsilon_i(t)\}_{i=1}^{17}$, son independientes o no; con esto, se verificaría si existe relación entre los residuos. Así, para aplicar este contraste y puesto que se trabaja con valores numéricos, consideramos la media como punto de corte para obtener las rachas.

   Luego, bajo un nivel de significación del 5 %, para casi todos los errores de cada tiempo se acepta H_0, pues se obtuvieron `p-valores` mayores a 0.05, tal como se muestra en la Figura 4.21.

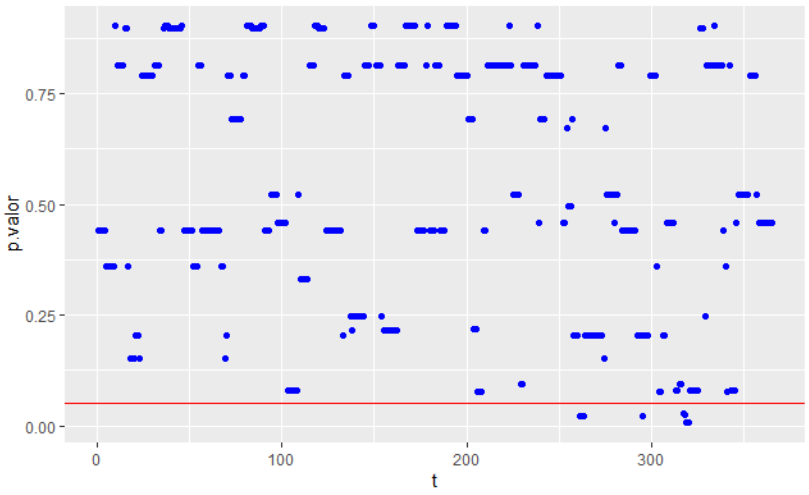


Figura 4.21: `p-valores` del test de rachas
Elaboración: Autor

   Sin embargo, puesto que esta prueba se basa en el número de rachas que se obtienen de la muestra, para verificar de mejor manera la autocorrelación en los residuos se considera la prueba de *Ljung-Box* (Hyndman y Athanasopoulos

[2018]), la cual contrasta una aleatoriedad "general" de los residuos en base a una serie de retrasos, visto de otra forma, se prueba que las k primeras autocorrelaciones son significativamente diferentes.

Con esto, en la Figura 4.22, se muestran los `p-valores` obtenidos para cada t, de donde se tiene que se acepta la hipótesis nula de no autocorrelación en los residuos bajo un nivel de significación del 5 %. Por tanto, para el pronóstico de las funciones bajo los dos escenarios, en cada tiempo no existe una influencia de la información de los residuales.

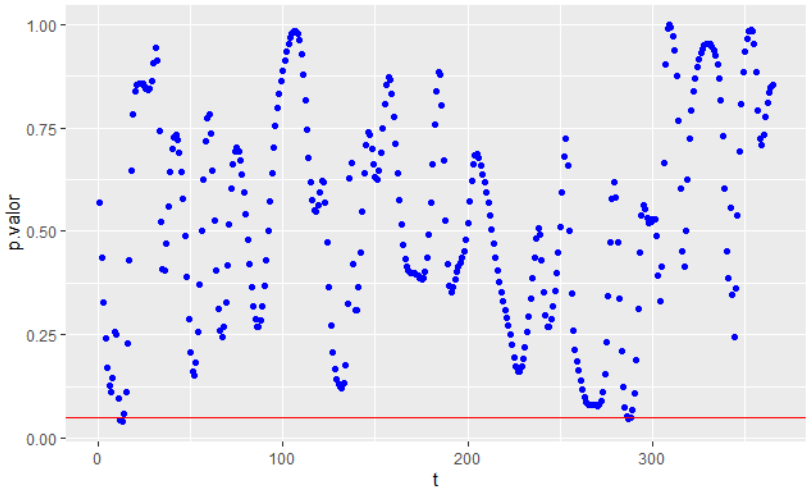


Figura 4.22: `p-valores` del test de Ljung-Box
Elaboración: Autor

2. *Media cero:* en la Figura 4.23, se muestran las medias calculadas de la componente de error para cada t, de donde se puede apreciar que se aproximan a cero (línea roja). Por tanto, se verificaría el supuesto.

   Cabe mencionar que este supuesto o incluso el supuesto de normalidad, podrían considerarse no restrictivos para el modelo, pues en base a varias fuentes bibliográficas, esto es útil para obtener intervalos de predicción en las estimaciones. No obstante, se utilizará el método bootstrap explicado en la subsección 3.6.4, para obtener las bandas de confianza.

En consecuencia, puesto que la componente de error del modelo cumple con los supuestos (en esencia la de no autocorrelación) se tiene que para la predicción de la función, los errores no contienen información útil en cada tiempo t. Esto aplica para el enfoque del mejor caso y para el enfoque del peor caso, pues recordemos que el "ordenamiento" de los datos funcionales es para el pronóstico de las puntuaciones de los datos en las componentes.

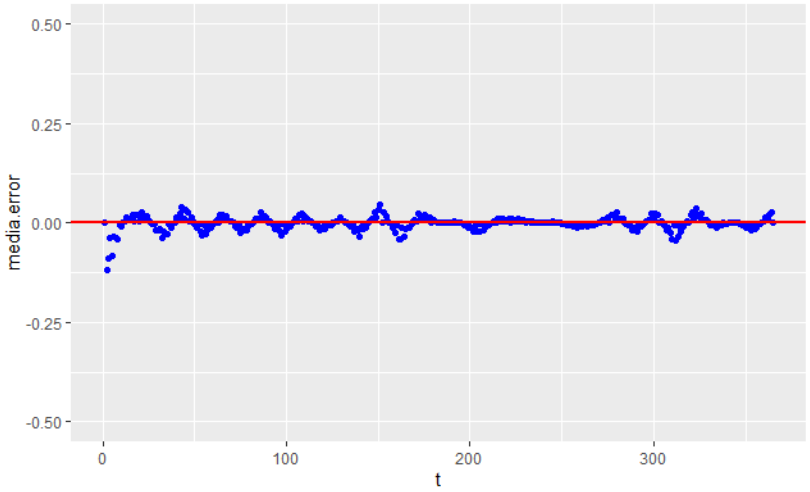


Figura 4.23: Media del error $\varepsilon_i(t)$, para cada t
Elaboración: Autor

Finalmente, en la Tabla 4.9 se muestran algunas medidas de ajuste, de donde se puede observar que son valores que describen un buen modelo, es decir, las componentes principales ajustan adecuadamente los datos funcionales; además que la información atípica fue necesaria retirarla para obtener dichos valores de los estadísticos.

Error Cuadrático Medio	$MSE = 1782,71$
$R-$cuadrado	$R^2 = 98{,}17\,\%$
Criterio de Información de Akaike	$AIC = 10,49$
Criterio de Información Bayesiana	$BIC = 37{,}43$

Tabla 4.9: Medidas de bondad de ajuste del modelo

Por consiguiente, del modelo ajustado, a continuación, se presentan los pronósticos de las puntuaciones EMBI en base a los dos escenarios explicados en la sección 3.6.5.

4.5.2. Estudio del mejor caso para el EMBI 2021

Para este escenario, se tiene que el orden de la muestra funcional según el crecimiento de las MD_i es: 2002 (menos profunda), 2001, 2003, 2016, 2015, 2014, 2010, 2019, 2006, 2018, 2011, 2017, 2004, 2012, 2007, 2013 y 2005 (más profunda); por tanto, en base a este orden, se realizan los pronósticos de las puntuaciones $f_{j,i}$ para predecir la función que contenga los EMBI diarios del 2021 en el mejor caso, o visto de otro modo, predecir una función que contenga registros del EMBI que comúnmente ha obtenido el Ecuador.

Pronóstico

Del modelo ajustado, se determina el comportamiento de las puntuaciones $f_{j,(i)}$ para todo $(i) = (1), ..., (17)$, con $j = 1, 2, 3, 4$, según el orden de las MD_i. Esto con el fin de predecir una nueva puntuación en cada componente, utilizando el suavizamiento exponencial. Entonces, de la función `forecast`, en la Figura 4.24, se presentan las series de los $f_{j,(i)}$ para cada j (línea negra) junto con lo ajustado por el modelo ETS de espacio de estados aditivo, sin tendencia ni estacionalidad (línea azul); además de la predicción de una puntuación en cada componente.

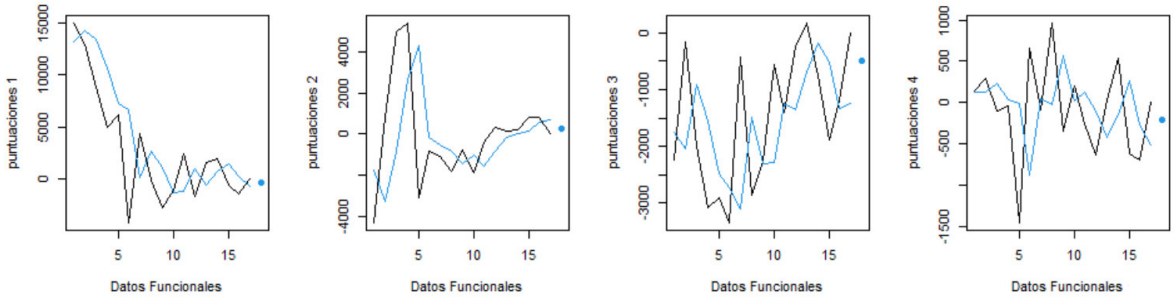


Figura 4.24: Serie de las puntuaciones de los datos $X_i(t)$ para el mejor caso
Elaboración: Autor

Como se puede ver, la nueva puntuación se relaciona bastante con la última puntuación $f_{j,(17)}$, lo que nos lleva a pensar que las primeras puntuaciones no brindan una buena ponderación al pronóstico, lo cual es lo que se deseaba, pues ahora, al multiplicar estas nuevas puntuaciones con sus respectivas $\alpha_j(t)$ y sumar la media funcional, se estima la función que representa las puntuaciones diarias del indicador del riesgo país en el 2021, en un caso normal (Figura 4.25). Es decir, se ha pronosticado una curva que pertenece al conjunto de curvas más centrales; aún así, se observa que se requieren tomar medidas más importantes para seguir reduciendo este indicador y lograr ser una mejor atracción en inversión. A esta curva estimada la denotaremos como 2021_t^+.

Precisión

De la curva estimada, se alcanza una puntuación mínima de 600 pb y una máxima de 818 pb, el cual se lo puede describir como un buen caso para el Ecuador, según el histórico considerado. Ahora, Andreozzi [2017] utilizan el Error Medio Absoluto Porcentual ($MAPE$) como una medida de bondad de pronóstico para evaluar el desempeño del modelo en la estimación de valores futuros. Entonces, para vali-

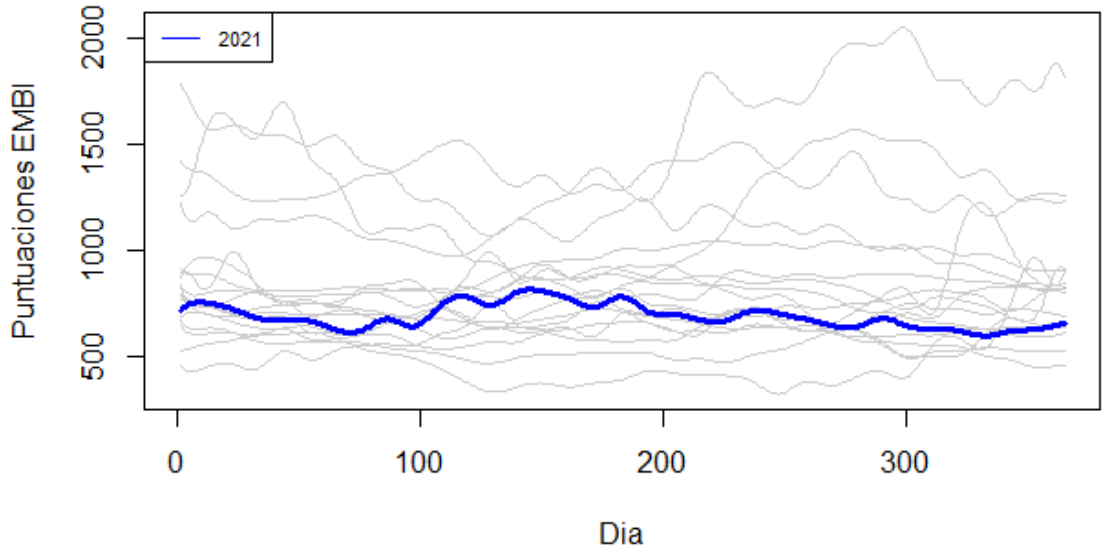


Figura 4.25: Pronóstico de la función de las puntuaciones EMBI 2021 del mejor caso
Elaboración: Autor

dar las predicciones obtenidas de la función 2021_t^+, se han considerado los datos registrados hasta el 05 de agosto de 2021 (día 217).

En la Figura 4.26, se observan las puntuaciones registradas (hasta el momento) del 2021, de donde los puntajes altos en los primeros tres meses, se debe por la incertidumbre de los mercados internacionales frente al proceso electoral que se llevó a cabo; luego hay un notable cambio en el mes de abril, que según El Universo [2021], se debieron gracias a renegociaciones con tenedores de bonos, la Ley de la Defensa de la Dolarización y, el más importante, el repunte en bonos internacionales y lograr acceder a líneas de crédito más baratas.

Luego, bajo esta breve explicación y en lo que se observa del gráfico, se tiene que a partir de mediados de abril, las puntuaciones son inferiores a 950 pb; por lo cual, de la función pronosticada, resulta interesante compararla con dichas puntuaciones del 2021. En el gráfico 4.27, se muestra esta comparación, donde claramente se ve que en la primera parte del gráfico hay una adecuada relación; sin embargo, para el resto de puntuaciones ya no es conveniente relacionarlas, debido a la gran diferencia, ya que para este *mejor caso* lo predicho no superan los 900 pb.

Con esto, para evaluar la precisión del modelo, y pese a no tener una relación directa, utilizamos los datos que mejor se comparan entre lo dado por el modelo

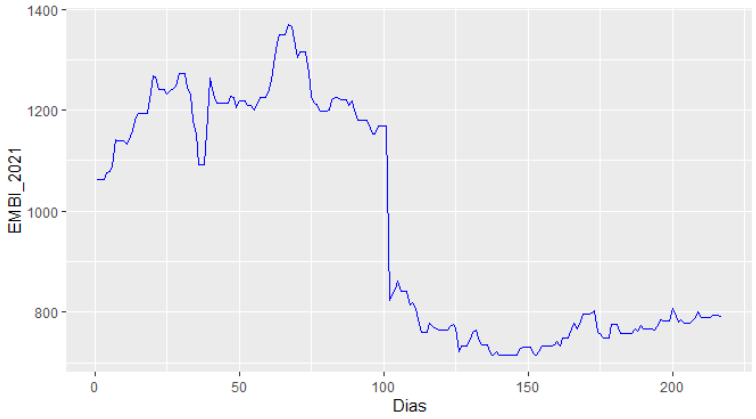


Figura 4.26: Evolución de las puntuaciones EMBI en el 2021 hasta el 05 de agosto
Elaboración: Autor

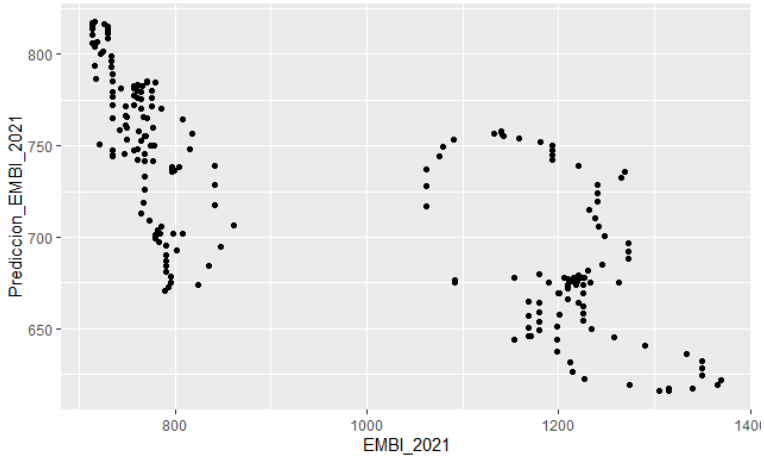


Figura 4.27: Comparación de los pronósticos del mejor caso con los datos actuales del 2021
Elaboración: Autor

y lo registrado en 2021. Así, utilizamos la métrica de rendimiento $MAPE$ definida como:

$$MAPE = \frac{100}{T} \sum_{t=1}^{T} \left| \frac{2021_t - 2021_t^+}{2021_t} \right|$$

con 2021_t las puntuaciones EMBI del 2021 en el tiempo t y 2021_t^+ el valor en t de la curva estimada por el modelo.

Bajo lo anterior, se obtuvo el siguiente valor:

$$MAPE = 7{,}3109\,\%$$

así, la métrica $MAPE$ nos indica que el modelo falla en un 7,31 % de pronóstico con respecto a los datos reales del 2021, lo cual se lo puede perfilar como un buen modelo, tomando en cuenta que el modelo fue adaptado a datos financieros.

En consecuencia, es conveniente considerar a la curva pronosticada para el año 2021 como un caso normal (mejor caso) para el registro de las puntuaciones del EMBI, es decir, un año en el que la economía no se ha visto afectada gravemente por algún fenómeno alarmante como protestas, bajas abruptas en los precios de petróleo o incumplimiento de pagos.

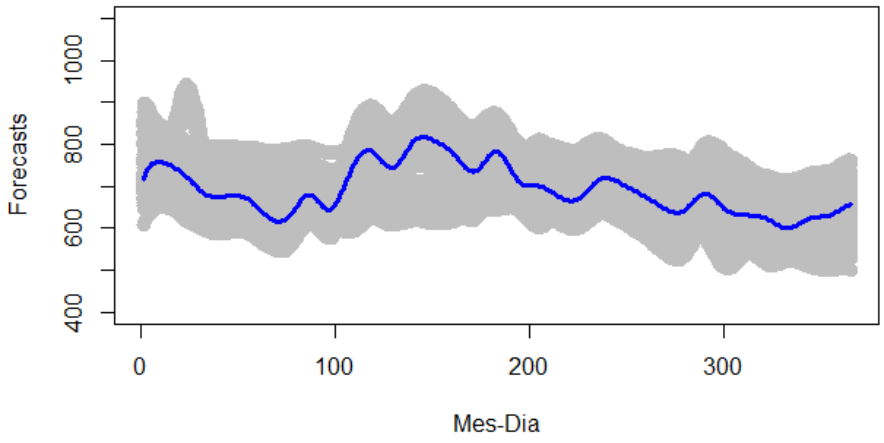


Figura 4.28: Banda de confianza Bootstrap para la función del mejor caso 2021_t^+, con un nivel de confianza del 90 %

Elaboración: Autor

Para concluir con este escenario del modelo, el cual ha sido una manera interesante de pronosticar un caso normal, gracias al algoritmo descrito en la subsección 3.6.4, se presenta la Figura 4.28, en donde se puede observar las bandas de confianza para este pronóstico con un nivel del 90 %. La gráfica nos proporciona una idea de qué valores o puntuaciones podrían darse a lo largo del tiempo para este escenario, teniendo en cuenta que para ciertos periodos de tiempo puede alcanzar un valor mínimo de alrededor 500 pb y para otros periodos, alcanzar a casi los 1000 pb, pero no superarlo. Así, aunque se dispongan de 17 datos funcionales de la v.f. $\mathcal{X}$, gracias al ordenamiento que se especifica para este escenario, las curvas más centrales (años normales) de cada muestra bootstrap tendrán más peso en la construcción de las bandas y, en esencia, no se utilizarían todos los datos funcionales, ya que por ejemplo, si en las muestras intervienen los datos funcionales correspondientes a los años 2002, 2001 o 2003, estos no tendrán un fuerte aporte en la formación de las bandas, ya que son datos de profundidad baja, es decir, se alejan de ser curvas centrales.

4.5.3. Estudio del peor caso para el EMBI 2021

Como se vio, se pudo pronosticar una función que describe un comportamiento normal de las puntuaciones diarias del EMBI para el año 2021; ahora, veamos qué ocurre en el pronóstico de una función que contiene puntuaciones diarias del EMBI altas, y que podrían ocurrir en el 2021. Como primera instancia, se tiene que el orden de la muestra funcional según el descenso de las MD_i es: 2005, 2013, 2007, 2012, 2004, 2017, 2011, 2018, 2006, 2019, 2010, 2014, 2015, 2016, 2003, 2001, 2002, es decir, para este enfoque se dará más ponderación a los datos funcionales que casi no se consideraron en el modelo anterior y son aquellas funciones que han presentado puntuaciones altas. Entonces, similar a lo visto anteriormente, se observa la Figura 4.29 con las series de cada $f_{j,(i)}$ para cada j (línea negra), añadiendo lo ajustado por el modelo ETS (línea azul), donde se aprecia que la puntuación pronosticada en cada FPC, tiene un leve alejamiento de la última puntuación (2002).

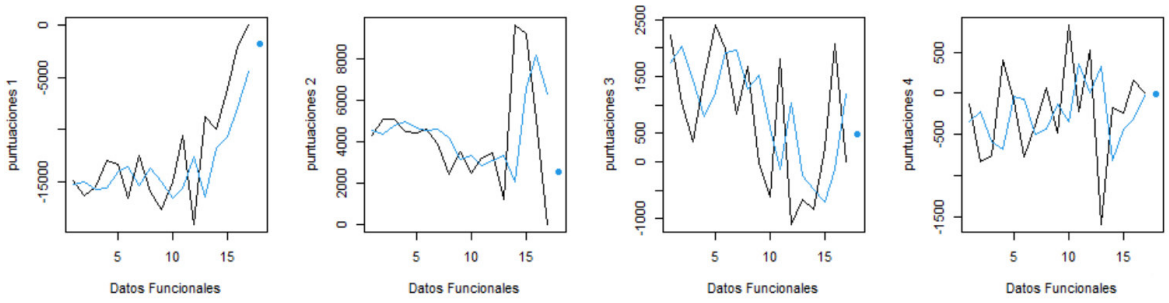


Figura 4.29: Serie de las puntuaciones de los datos $X_i(t)$ para el peor caso
Elaboración: Autor

Luego, siguiendo el mismo procedimiento del modelo, se grafica en 4.30 la función pronosticada bajo el contexto del peor caso (2021_t^-) y, como se observa, esa función tiene una tendencia similar al de los datos funcionales con EMBI altos. Se tiene como mínimo una puntuación de 1050 pb y como máximo 1776 pb, los cuales representan una situación preocupante debido a una situación económica fuerte, como por ejemplo, la falta de acuerdos con el FMI o el desplome en los precios de crudo.

Para validar de alguna forma el modelo, se considera la misma métrica utilizada en el enfoque anterior, con la diferencia de que se toman los datos que presentan una mejor comparación entre lo pronosticado y lo real (Figura 4.31), es decir, se toman en cuenta los datos en donde las puntuaciones EMBI registradas del 2021 superan los 1000 pb, pues menores a ese puntaje, el modelo no presenta buenas comparaciones, debido a que estamos abordando el peor caso.

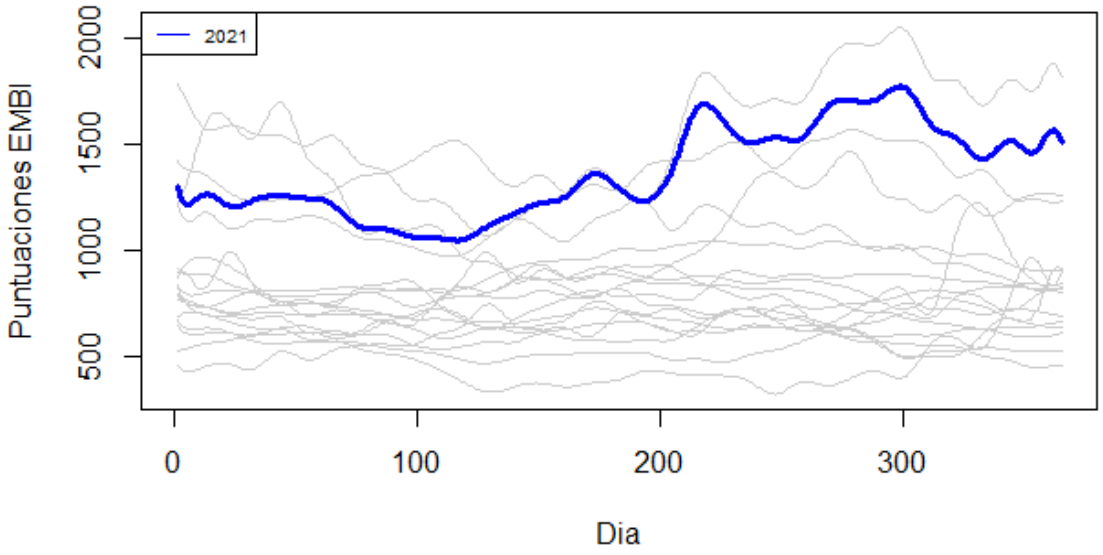


Figura 4.30: Pronóstico de la función de las puntuaciones EMBI 2021 del peor caso
Elaboración: Autor

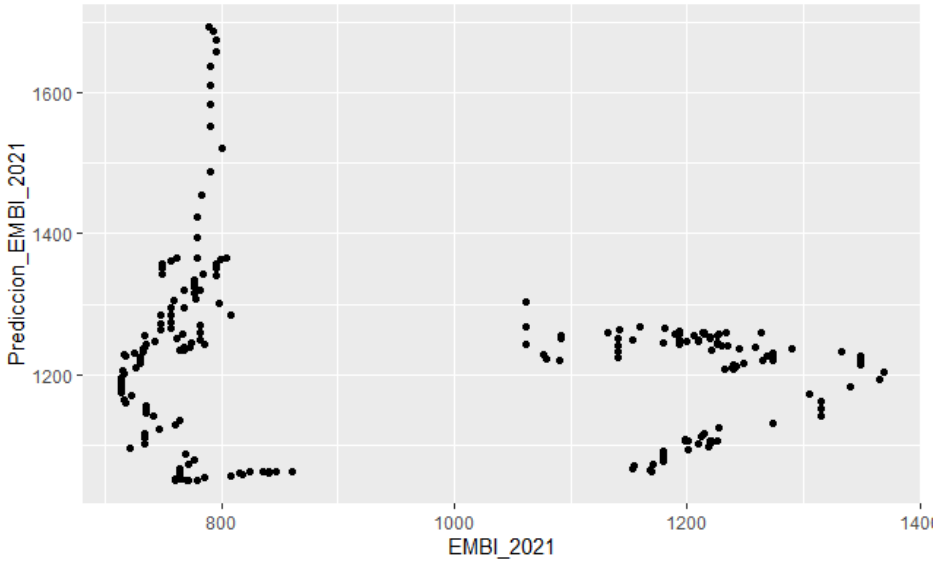


Figura 4.31: Comparación de los pronósticos del peor caso con los datos del 2021
Elaboración: Autor

Así entonces, se tiene que el Error Medio Absoluto Porcentual es:

$$MAPE = 7{,}103\,\%,$$

el cual también tiene un porcentaje bajo y poco distante a lo calculado en el escenario anterior; por lo tanto, al modelo se lo puede relacionar con un rendimiento

adecuado.

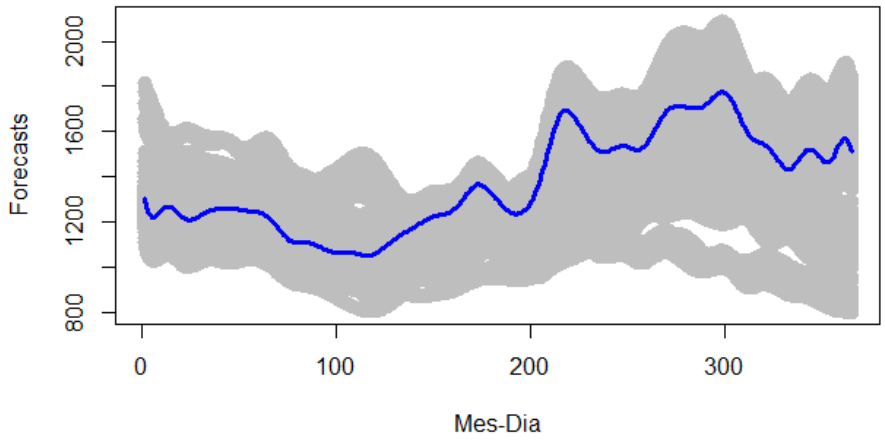


Figura 4.32: Banda de confianza Bootstrap para la función del peor caso 2021_t^-, con un nivel de confianza del 90 %
Elaboración: Autor

Finalmente, se presentan las bandas de confianza con un nivel del 90 % (Figura 4.32) para la curva pronosticada bajo el enfoque del peor caso. En la misma se observan las puntuaciones que podrían alcanzar para este escenario, tomando en cuenta que en la banda inferior se involucran valores que superan a los de la banda superior del gráfico 4.28, es decir, se observa que esta región representaría los valores más altos que podrían darse en el 2021, notando que podría alcanzarse hasta alrededor de 2000 pb.

En resumen, el modelo basado en las componentes principales funcionales bajo los dos escenarios estudiados, nos presentan una idea clara del comportamiento que podrían tener las puntuaciones a través de la función pronosticada y las bandas de confianza; de donde, resultó adecuado retirar los datos atípicos, ya que para el estudio del peor caso, se hubiese pronosticado una función con puntuaciones demasiado altas, situación que ha ocurrido en cuatro de los veintiún años considerados en este histórico.

Capítulo 5

Conclusiones y Recomendaciones

La economía del Ecuador depende fuertemente de los ingresos de las exportaciones de petróleo y de su relación comercial internacional, sobre todo con países como China y Estados Unidos, siendo dos factores esenciales para que la economía sea sensible ante cambios repentinos en los mismos, generando una variación notable en su desarrollo económico, ya sea en los incumplimientos de sus obligaciones financieras, como en los pagos de deuda pública o incluso en las escasas oportunidades de inversión dentro del país. Todos estos factores negativos y también los positivos, como el alza en los precios de barril de petróleo, se los resumen en lo que se denomina el riesgo país, medido a través del Indicador de Bonos de Mercados Emergentes (EMBI), el cual se lo relaciona como la capacidad de un Estado por cumplir con sus obligaciones a tiempo. Lo deseable es tener puntuaciones bajas de este índice para que los inversionistas no exijan un rendimiento muy alto para invertir en el Ecuador.

Es así que se consideraron las puntuaciones diarias del EMBI desde el año 2000 hasta el 2020 para comprender las fluctuaciones dadas en ese histórico y poder predecir las puntuaciones para el 2021. No obstante, el uso de por ejemplo, series de tiempo, no han presentado una buena explicación de este indicador, además que existen pocos estudios dedicados en la predicción del mismo. Por tanto, se hizo uso del Análisis de Datos Funcionales (FDA) para representar a cada año como una función que explique la forma intrínseca de las puntuaciones diarias y lograr tratarlos como un solo dato. Esta representación se la obtuvo mediante el suavizamiento de los datos discretos a través de Bases B-splines de orden 4, incluyendo la penalización por rugosidad λ para ajustar funciones suaves.

Bajo esta construcción de los veintiún datos funcionales definidos en el intervalo

$[1, 365]$, en primer lugar, se presentó un análisis exploratorio, detallando los cambios más notables que hubieron dentro de la muestra funcional. El cálculo de la primera y segunda derivada de las funciones, ayudaron a identificar ciertos patrones o cambios que ocurrieron en los datos; además de evidenciar por medio de un algoritmo basado en bootstrap suavizado con la profundidad modal, que dentro de la muestra, existieron cuatro datos funcionales atípicos: 2000, 2008, 2009 y 2020, pues presentaban un comportamiento muy diferente que el de los demás. Los efectos de la dolarización, la crisis financiera en Estados Unidos y la crisis sanitaria a nivel mundial, afectaron a la economía del Ecuador en los años detectados como atípicos, pues las puntuaciones incrementaron abruptamente en ciertos periodos de tiempo. En consecuencia, se sustrajeron los mencionados datos para calcular las medidas estadísticas funcionales y aplicar el análisis de componentes principales funcionales; se utilizaron cuatro componentes, de donde, respectivamente, se logró explicar el 78,23 %; 15,35 %; 3,71 % y 0,88 % de la variabilidad presentada en la muestra de los ahora 17 datos funcionales, siendo los años del 2001 y 2002 los que más varianza explicaban, debido a los estragos de la dolarización y otros factores como la falta de acuerdos con el FMI.

Con esto, se desarrolló el modelo a través de las cuatro primeras FPC, considerando dos enfoques interesantes que brindan una idea clara en el comportamiento futuro de este indicador, propuestos en este proyecto. Para ambos enfoques, se hizo uso de las medidas de profundidad de cada dato; así, el primer escenario se lo catalogó como el mejor caso, pues se pronosticó toda una función que tiene un comportamiento normal en las puntuaciones y que aún así no se tienen puntuaciones más bajas como para ser una mayor atracción de inversión. No obstante, bajo el segundo escenario, que se lo caracterizó como el peor caso, se pronosticaron las puntuaciones más altas que pueden haber en un año normal y, esto puede darse principalmente por el desplome del precio del petróleo o por alguna decisión ilegítima del Gobierno, como por ejemplo, en no pagar sus préstamos. Además, se implementó un algoritmo de bootstrap suavizado considerando la distancia L^2, para crear bandas de confianza con un nivel del 90 % en las funciones pronosticadas de ambos casos y poder establecer una mejor idea de qué puntuaciones se esperarían que se registren en los dos escenarios.

Así, según los modelos ajustados, es preocupante las puntuaciones que llega a tener el indicador del riesgo país ecuatoriano, pues pese a abordar un mejor caso, no se obtuvieron puntuaciones bajas como para que el Ecuador sea un referente en inversión, y esto se debe a que el país no ha tenido un buen histórico en las puntua-

ciones. Por tanto, aunque las puntuaciones normales del EMBI ecuatoriano oscilan entre 600 y 800 pb, como lo pronosticado, el Estado debería prestar más atención a este indicador y lograr reducir lo más posible el mismo, ya que se tendría un mejor acceso a financiamientos; para ello, se podrían tomar medidas como: el cumplimiento con los pagos de deuda, tener un crecimiento de exportaciones de otras materias primas, evaluar con cautela las propuestas de los candidatos en tiempos electorales (pues esto ha generado cierta incertidumbre en los inversionistas), invertir en proyectos que ayuden en realidad a la economía y como se ha visto, buscar mejores acuerdos con el FMI, el cual ha sido un factor que ha reducido notablemente las puntuaciones. Para el peor caso, las mismas medidas deberían llevarse a cabo y, en especial, la decisión más coherente para que no ocurran estos cambios, es en no depender solamente o mayoritariamente de un solo ingreso como lo es el petróleo, sino hallar otras fuentes que generen buenos o incluso mejores ingresos al país y lograr fortalecer su producción. Además de proyectar mayores niveles de apertura comercial, pero a un mediano plazo.

Finalmente, para futuros proyectos de investigación y con el fin de darle un mayor interés a este campo de la Estadística, se presentan algunas recomendaciones.

- Lo estudiado aquí ha sido un análisis estadístico común sobre una muestra de funciones, de donde se pudo apreciar que todas las técnicas tienen su forma de adaptarse a este contexto, como las medidas de profundidad que adaptan el concepto de la curva más central. Así, utilizando otros conjuntos de datos, se podrían presentar estudios donde se utilicen otras técnicas como: suavizamiento mediante Kernels, uso de otro tipo de bases, uso de otras medidas de profundidad como la de proyecciones aleatorias, mínimos cuadrados parciales funcionales (FPLS), modelos de regresión funcionales (FRM), clustering funcional y series de tiempo funcionales.

- De la misma base de datos sobre el indicador del riesgo país, se podría realizar un modelo basado en el análisis de series de tiempo funcionales, utilizando por ejemplo, un proceso autorregresivo funcional FAR y hacer una comparación de pronósticos con lo expuesto en el presente proyecto.

- Utilizando una variable de interés que se desee relacionar con el riesgo país, se propone realizar algún tipo de modelo de regresión funcional tal que se encuentre alguna relación y poder concluir, si en cierto modo, esta variable afecta en las puntuaciones del riesgo país.

- En el presente proyecto, se ha abordado la técnica FDA, haciendo más énfasis al contexto metodológico para la aplicación de varias técnicas multivariantes en su versión funcional; sin embargo, todas estas técnicas, o a menos las más conocidas, tienen su fundamento teórico utilizando herramientas del Análisis Funcional y la Teoría de Probabilidades; así por ejemplo, al suponer que una función es cuadrado integrable, se utilizan resultados teóricos de los espacios de Hilbert para estructurar a los datos funcionales o, para el FPCA que hace uso de la teoría espectral. Por lo tanto, se recomienda investigar bajo un contexto más teórico lo tratado en este proyecto e incluso otras técnicas que no han sido abordadas.

Referencias

Andreozzi, L. (2017). Pronósticos probabilísticos de la esperanza de vida para la ciudad de rosario, argentina. *SaberEs*, 9:47–67.

BCE (2021). Indicadores económicos - riesgo país, diaria. `https://www.bce.fin.ec/index.php/informacioneconomica`.

Calahorrano, L., Tigse, S., y Caicedo, F. (2020). Variacióm del indicador riesgo país en el flujo de inversión extranjera del ecuador. *Universidad Ciencia y Tecnología*, 24(107):70–78.

Carrillo, P., Díaz, J., y Flores, M. (2021). What are the main variables that influence the dynamics of Ecuador's sovereign risk? Technical Report EPEL-WP-2021-03, Ecuadorian Political Economy Lab.

Chávez, C., Sánchez, J., y DelaCerda, J. (2015). Análisis de componentes principales funcionales en series de tiempo económicas. *GECONTEC*, 3(2).

Clark, E. (2017). Terminology and history of country risk. In *Evaluating Country Risks for International Investments*, pages 3–20. WORLD SCIENTIFIC.

Cuevas, A., Febrero, M., y Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics & Data Analysis*, 51(2):1063–1074.

Cuevas, A., Febrero, M., y Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496.

de Boor, C. (2001). *A Practical Guide to Splines*. Springer, Estados Unidos.

Deręgowski, K. y Krzyśko, M. (2011). Principal components analysis for functional data. *Colloquium Biometricum*, 41:5–17.

El Universo (2002). Riesgo país de ecuador registra su más bajo nivel histórico. https://www.eluniverso.com/2002/04/03/0001/9/7AA821D6D21943CA9FC03598FF1652C0.html/.

El Universo (2003). Rompimiento de gobierno con pachakutik subió el riesgo país. https://www.eluniverso.com/2003/08/22/0001/9/91CAE05224D34061BB601A3468BC1BFF.html/.

El Universo (2006). Riesgo país de ecuador bajó a 503 puntos. https://www.eluniverso.com/2006/04/22/0001/9/6127B54FDB924CFE90DC95F99DC194E6.html/.

El Universo (2008). El riesgo país llega a nivel de crisis bancaria - riesgo país supera los 5.000 puntos. https://www.eluniverso.com/2008/10/31/0001/9/F49504D0666546E19BD9C09299F603E4.html y https://www.eluniverso.com/2008/12/23/1/1356/F12DB77EDEF5452EB8B6B0A3E6EEC3B7.html.

El Universo (2021). Riesgo país a la baja y recuperación de bonos, asociados a la aprobación de la ley de defensa de la dolarización. https://www.eluniverso.com/noticias/economia/ley-defensa-dolarizacion-riesgo-pais-precios-bonos-presidente-lenin-moreno-asamblea-nacional-abril-2021-nota/.

Febrero, M., Galeano, P., y González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics*, 19(4):331–345.

Febrero-Bande, M. y de la Fuente, M. O. (2012). Statistical computing in functional data analysis: TheRPackagefda.usc. *Journal of Statistical Software*, 51(4).

Fernández y Cao (2021). Simulación estadística. https://rubenfcasal.github.io/simbook/.

Ferraty, J. y Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer, Estados Unidos.

Fraiman, R. y Muniz, G. (2001). Trimmed means for functional data. 10(2):419–440.

Gonzalez, M. y Minguez, R. (2005). A Study of Country-Risk for Non-Developed Countries in 1980-2000. *Applied Econometrics and International Development*, 5(1).

Horváth, L. y Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer New York.

Hyndman, R., Koehler, A., Ord, J., y Snyder, R. (2008). *Forecasting with Exponential Smoothing: the State Space Approach*. Springer-Verlag Berlin Heidelberg.

Hyndman, R. J. y Shang, H. L. (2009). Forecasting functional time series. *Journal of the Korean Statistical Society*, 38(3):199–211.

Hyndman, R. J. y Shang, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19(1):29–45.

Hyndman, R. J. y Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956.

Hyndman y Athanasopoulos (2018). Forecasting: Principal and practice. `https://otexts.com/fpp2/index.html`.

Kokoszka, P. y Reimherr, M. (2017). *Introduction to Functional Data Analysis*. Chapman and Hall/CRC.

Kreyszig, E. (1989). *Introductory Functional Analysis with Applications*. Wiley, Estados Unidos.

Lapitz, R., Gorfinkiel, D., Acosta, A., Flórez, M., y Gudynas, E. (2005). *El Otros Riesgo País: indicadores y desarrollo en la economía global*. Coscoroba ediciones.

Lazo, L. (2021). ¿por qué el riesgo país debería considerarse en la evaluación financiera de un proyecto o una empresa? *Estudios de la Gestión: revista internacional de administración*, pages 255–268.

Matsumoto (2019). Do basic arima model forecasts have trouble with crises? evidence from argentina's emerging market bond index. `https://medium.com/@rikimatsumoto/do-basic-arima-model-forecasts-have-trouble-with-crises-evidence-from-argentina-working-draft-d2f8a4f30197`.

Ramsay, J., Hooker, G., y Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer New York.

Ramsay, J. y Silverman, B. (1997). *Functional Data Analysis*. Springer, Estados Unidos.

Ramsay, J. y Silverman, B. (2005). *Functional Data Analysis*. Springer, Estados Unidos.

Shang, H. L. (2013). ftsa: An r package for analyzing functional time series. *The R Journal*, 5(1):64.

Shang, H. L., Hyndman, R., y Booth, H. (2010). A comparison of ten principal component methods for forecasting mortality rates. *Monash University, Department of Econometrics and Business Statistics, Monash Econometrics and Business Statistics Working Papers*.

Ullah, S. y Finch, C. (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology*, 13(1).

UTA (2021). Riesgo país ecuador. `https://fca.uta.edu.ec/v4.0/images/OBSERVATORIO/dipticos/Diptico_N58.pdf`.

Valle, A. y Aguirre, I. (2020). Panorama económico del ecuador en tiempos de coronavirus. *Banco Central del Ecuador: Apuntes de Economía No. 64*.

Apéndice A

Teoría complementaria

A.1. Producto interno

Definición A.1. *Sea E un espacio vectorial sobre $\mathbb{R}$. La función*

$$\begin{aligned} \langle \cdot, \cdot \rangle : & E \times E \longrightarrow \mathbb{R} \\ & (x, y) \longmapsto \langle x, y \rangle \end{aligned}$$

se conoce como producto interno o producto escalar, tal que para todo $x, y, z \in E$ y para todo $\alpha, \beta \in \mathbb{R}$ se satisface:

$(E1)$ *(Simetría)* $\langle x, y \rangle = \langle y, x \rangle$.

$(E2)$ *(Positividad)* $\langle x, x \rangle \geq 0$, *con* $\langle x, x \rangle = 0$ *si y solo si* $x = 0_E$, *donde* 0_E *es el neutro aditivo de* E.

$(E3)$ *(Bilinealidad)* $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$.

Al par $(E, \langle \cdot, \cdot \rangle)$ *se le llama espacio de producto interno.*

Se puede pensar que el producto interno define una medida escalar de asociación entre x e y, pues la simetría de la medida significa que no importa el orden de las cantidades; la bilinealidad se refiere a que al cambiar la escala de alguno de los argumentos tendrá el mismo resultado de escala en la medida de asociación, además que, la medida de asociación de un argumento con la suma de dos argumentos distintos es la suma de las medidas individuales de asociación.

Finalmente, la positividad concierne a una medida de tamaño de un elemento x; así, surge la siguiente definición de magnitud de un elemento de un espacio vectorial:

DEFINICIÓN A.2. *La norma de un elemento x de un espacio vectorial E, se define como:*

$$\|x\| = \sqrt{\langle x, x \rangle},$$

pues cumple con las propiedades de norma:

1. $\|x\| \geq 0$ *y* $\|x\| = 0$ *si y solo si* $x = 0_E$.
2. $\|\alpha x\| = |\alpha| \, \|x\|$, *para todo* $\alpha \in \mathbb{R}$.
3. $\|x + y\| \leq \|x\| + \|y\|$.

A $\|\cdot\|$ *se la conoce como la norma asociada a* $\langle \cdot, \cdot \rangle$.

Un resultado importante es la desigualdad de *Cauchy-Schwarz-Buñakovski*, la cual, para todo $x, y \in E$, se cumple que:

$$| \langle x, y \rangle | \leq \|x\| \cdot \|y\| \tag{A.1}$$

Por otra parte, el ángulo entre x e y, el cual se lo denota por θ, se define como:

$$\cos\theta = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}.$$

Además, un elemento $x \in E$ es ortogonal a un elemento $y \in E$, denotado por $x \perp y$, si:

$$\langle x, y \rangle = 0.$$

Así, una buena interpretación de (A.1) es la siguiente: si $| \langle x, y \rangle |$ se acerca a la cota $\|x\| \cdot \|y\|$, los elementos se definen en direcciones semejantes según el espacio vectorial, de donde, se tendría que la medida de asociación entre los elementos es alto; por otro lado, si el producto interno tiende a cero, las direcciones de los elementos serán aproximadamente ortogonales y su medida de asociación es baja. Además, se puede vincular el producto interno con el concepto de ángulo a través de la siguiente propiedad, conocida como la desigualdad del coseno:

$$-1 \leq \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|} \leq 1.$$

Adicionalmente, se puede establecer una medida de distancia entre x e y, definida por:

$$d_{xy} = \|x - y\| = \sqrt{\langle x - y, x - y \rangle},$$

de donde, a d se la conoce como la métrica asociada a $\langle \cdot, \cdot \rangle$. Con esto, se tiene la siguiente relación: si las direcciones y la norma definidas en x e y son similares, se tendrá que $x - y$ es bajo, por lo cual, la distancia entre dichos elementos será corta.

Para concluir con esta parte, las propiedades algebraicas de simetría, positividad y bilinealidad que definen a un producto interno producen definiciones relevantes del tamaño de un elemento y, del ángulo y distancia entre dos elementos; por ende, independientemente de cómo se definan x e y, las caracterizaciones de estas tres propiedades permanecen intactas. Así por ejemplo, si consideramos el espacio vectorial $E = \mathbb{R}^n$ con $n > 0$, dotado del siguiente producto escalar:

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i \cdot y_i, \qquad \forall x, y \in \mathbb{R}^n$$

se tiene que:

- la norma de un $x \in \mathbb{R}^n$ denota la longitud del vector; y,
- θ es el ángulo entre los vectores x e y, particularmente en $\mathbb{R}^2$, es fácil observarlo bajo el sentido geométrico usual, al igual que para d_{xy}.

OBSERVACIÓN. Se dice que un espacio es de Hilbert, si es un espacio de producto interno completo (bajo la métrica definida por el producto interno), donde un espacio métrico es completo si toda sucesión de Cauchy del espacio, converge a un elemento de dicho espacio.

A.2. Análisis de Componentes Principales (PCA)

En el análisis y procedimiento de información, se suelen presentar dos problemas: la dimensionalidad de la data e información redundante, entonces, la idea general del PCA es en reducir esta dimensión de las variables iniciales y obtener un número menor de variables no correlacionadas llamadas *componentes principales* (PC) las cuales son combinaciones lineales de las originales.

Regularmente, la variación de los datos se lo resume a través de la matriz de covarianza o de correlación, siendo la idea fundamental del PCA, el hallar una sucesión de vectores perpendiculares que expliquen eficientemente la varianza de las observaciones. Así pues, para abordar la técnica y como usualmente se utiliza, denotaremos por p el número total de variables (columnas) del conjunto de datos $X = (X_1, ..., X_p)$ y por n al número de individuos u observaciones (filas) de cada va-

riable, es decir, asumimos que se dispone de una matriz de datos x_{ij} con $i \in \{1, ..., n\}$ y $j \in \{1, ..., p\}$.

OBSERVACIÓN. Para evitar la influencia de variables con grandes variaciones, se consideran que las mismas estén centradas y reducidas (media 0 y varianza 1), por lo cual se trabaja sobre la matriz de correlación R.

Notando que cada variable X_j constituye un eje en $\mathbb{R}^m$, se tiene que el dato x_{ij} es la coordenada del individuo i sobre el eje X_j. Con esta idea, se construyen nuevas variables Z_k (PC) que serán p nuevos ejes sobre $\mathbb{R}$.

A.2.1. Cálculo de las componentes

Primera componente

Supongamos que $\alpha_1 = (\alpha_{11}, ..., \alpha_{1p})$ es el vector unitario (también llamado *vector de pesos*) de la componente Z_1 de donde, puesto que se desea recoger la mayor cantidad de información posible de los individuos sobre este nuevo eje, se necesita que en promedio, las proyecciones sean lo máximo posible. La proyección del individuo i sobre el eje Z_1 viene dada por:

$$Z_{i1} = X_{(i)} \cdot \alpha_1 = \sum_{j=1}^{p} \alpha_{1j} \cdot x_{ij}$$

con $X_{(i)}$ los datos de la fila i de la matriz.

Así, si $Z_1 = (z_{11}, ..., z_{i1}, ..., z_{n1}) \in \mathbb{R}^n$ que contiene las coordenadas de los n individuos sobre Z_1, se tiene que:

$$Z_1 = X \cdot \alpha_1 = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix}.$$

Luego, para que Z_1 recoja la mayor información de los individuos, se requiere que la varianza Z_1 sea máxima. Entonces, como Z_1 es centrada, pues X_j lo son, se sigue que:

$$\text{Var}[Z_1] = \frac{1}{n} \sum_{i=1}^{n} z_{i1}^2 = \frac{1}{n} Z_1^\intercal Z_1 = \alpha_1^\intercal \underbrace{\frac{X^\intercal X}{n}}_{R} \alpha_1,$$

que equivalentemente, bajo el criterio algebraico (promedio de las proyecciones al

cuadrado), se lo puede ver como:

$$\text{Var}[Z_1] = \frac{1}{n}\sum_{i=1}^{n}(X_{(i)}\alpha_1)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\alpha_{1j}\cdot x_{ij}\right)^2.$$

Con esto, como R es la matriz de correlación (y de covarianza), a Z_1 se la obtiene mediante el siguiente problema:

$$\begin{cases} \text{Max Var}[Z_1] = \alpha_1^\intercal R\,\alpha_1 \\ \alpha_1^\intercal\alpha_1 = 1 \end{cases}$$

de donde, por la técnica de multiplicadores de Lagrange, se obtiene que $R\alpha_1 = \lambda_1\alpha_1$, es decir, α_1 es el autovector de R asociado al autovalor λ_1; además, es importante notar que

$$\text{Var}[Z_1] = \alpha_1^\intercal(\lambda\alpha_1) = \lambda_1,$$

por lo cual, se tendría que λ_1 tiene que representar el valor propio más grande de R.

Segunda componente

Para calcular Z_2, se procede análogamente, considerando en recoger la mayor información de los individuos sin considerar los ya recogidos en Z_1. Para esto, se añade una restricción que denote que los vectores unitarios que forman los ejes sean ortogonales (componentes no correlacionadas); así, sea α_2 el vector unitario de Z_2, entonces se resuelve el siguiente problema:

$$\begin{cases} \text{Max Var}[Z_2] = \alpha_2^\intercal R\,\alpha_2 \\ \alpha_2^\intercal\alpha_1 = 0 \\ \alpha_2^\intercal\alpha_2 = 1 \end{cases}$$

de donde, se tiene que $R\alpha_1 = \lambda_1\alpha_1$ y $R\alpha_2 = \lambda_2\alpha_2$, siendo λ_2 el segundo autovalor más grande de R y, $\text{Var}[Z_2] = \lambda_2$.

Generalización

Para $k = 3, ..., p$, se tiene que la $k-$ésima PC viene dada por el vector unitario α_k, el cual se lo obtiene al resolver

$$R\alpha_k = \lambda_k\alpha_k$$

siendo λ_k el $k-$ésimo autovalor más grande de R y α_k su autovector asociado, tal que $\alpha_k^\intercal \alpha_k = 1$ y $\alpha_k^\intercal \alpha_\ell = 0$ si $\ell < k$. Además que,

$$\begin{aligned} &\text{Var}[Z_k] = \lambda_k \\ &\text{Cov}(Z_k, Z_\ell) = 0, \quad \ell < k. \end{aligned}$$

A.2.2. Elección del número de componentes

Dos de los criterios más comunes para determinar el número k de componentes principales son:

- *Criterio de la media:* elegir las componentes principales tales que λ_k sobrepase la media de los autovalores, es decir, elegir aquellas que:

$$\lambda_k > \frac{1}{p} \sum_{j=1}^{p} \lambda_j.$$

  Si las variables están centradas y reducidas, se eligen aquellas tales que $\lambda_k > 1$.

- *Gráfico de sedimentación:* representar los autovalores $\lambda_1, ..., \lambda_p$ en un plano y, como $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, se tendrán cambios bruscos al inicio del mismo y, desde el primer autovalor que se estabilice el gráfico, se selecciona el valor k.

Apéndice B

Implementación en el software R

B.1. Análisis previo

```
##############################################
#### ANALISIS DE DATOS FUNCIONALES (FDA) #####
##############################################
#### Carrera: Ingenieria Matematica
#### Autor: Alexander Constante
##############################################

###########################################
## Lectura y Preprocesamiento de los datos
###########################################

library(readr)       #lectura del archivo
library(lubridate)   #formato de fechas
library(tidyr)       #para spread
library(naniar)      #graficos NA
library(openxlsx)    #para guardar bases finales
library(reshape2)    #para grafico de evolucion
library(ggplot2)
library(tidyverse)

### datos 2000-2018
embi <- as.data.frame(read_csv("01Preprocesamiento/RiesgoPais2000.csv"))
str(embi)

embi[,1] <- as.Date(substring(embi[,1], 1, 10)) # para que no aparezca la hora
embi1 <- as.data.frame(cbind('Anio' = year(embi[,1]), 'Mes' = month(embi[,1]),
                             'Dia' = day(embi[,1]), 'RiesgoPais' = embi[,2]))

embi2 <- spread(data = embi1, key = Anio, value = RiesgoPais)  ## anios como columnas.

embi2 <- unite(embi2, 'MesDia', 1:2,  sep = "-", remove = TRUE)  # en una sola columna el
    mes y dia

data1 <- embi2
```

```r
data1 <- data1[,-20]  # data del 2000 al 2017, incluida fecha 2-29

#### datos 2004-2021
embi <- as.data.frame(read_csv("01Preprocesamiento/RiesgoPais2004.csv"))

embi[,1] <- as.Date(substring(embi[,1], 1, 10))
embi <- embi[-c(1:4904),]  # Considero desde 2018
embi1 <- as.data.frame(cbind('Anio' = year(embi[,1]), 'Mes' = month(embi[,1]),
                             'Dia' = day(embi[,1]), 'RiesgoPais' = embi[,2]))

embi2 <- spread(data = embi1, key = Anio, value = RiesgoPais)  ## anios como columnas.

embi2 <- unite(embi2, 'MesDia', 1:2,  sep = "-", remove = TRUE)

data2 <- embi2

library(dplyr)
base <- left_join(data1, data2, by = 'MesDia')
detach(package:dplyr)

embi2021 <- base[,c('MesDia','2021')] # guardo datos 2021
embi <- base[,-23] # retiro 2021
row.names(embi) <- embi['MesDia']
embi <- embi[,-1] # retiro columna MesDia

vis_miss(embi) # grafico valores perdidos

### Ubicacion de los valores perdidos e imputacion
NA2018 <- which(is.na(embi['2018'])==TRUE)
rownames(embi)[NA2018]
NA2018 <- NA2018[-1]  ### quito 2/29
embi[NA2018,'2018'] <- embi[NA2018[1]-1,'2018']

NA2019 <- which(is.na(embi['2019'])==TRUE)
rownames(embi)[NA2019]
NA2019 <- NA2019[-1]  ### quito 2/29
for(i in seq(1, length(NA2019), by = 2)){
  embi[NA2019[i:(i+1)],'2019'] <- embi[NA2019[i]-1,'2019']
}

NA2020 <- which(is.na(embi['2020'])==TRUE)
rownames(embi)[NA2020]
for(i in seq(1, length(NA2020), by = 2)){
  embi[NA2020[i:(i+1)],'2020'] <- embi[NA2020[i]-1,'2020']
}

## Retiro 29 de febrero
embi <- embi[-60,]

### Guardo la base de datos final y del 2021
write.csv(embi, '02Datos/EMBI.csv')
embi2021 <- embi2021[-60,]
write.csv(embi2021, '02Datos/EMBI2021.csv')
```

```r
ind <- embi
ind$MesDia <- 1:365
ind <- melt(ind, id.vars = 'MesDia')

##### Grafico evolucion por anio
ind %>% ggplot(aes(x = MesDia, y = value, col = variable, group = variable, group = 1)) +
  labs(x = "Dia", y = 'Indicador EMBI') +
  theme(axis.text.x = element_blank()) +
  geom_line() +
  facet_wrap(~variable) +
  theme(legend.position="none")

####### Grafico: Evolucion anual
ind  %>% ggplot(aes(x = MesDia, y = value,  col = variable, group = variable)) +
  labs(x = "Mes_Dia", y = 'Riesgo Pais (EMBI)', title = "Evolucion") +
  theme(axis.text.x = element_blank()) +
  geom_line()
```

B.2. Aplicación FDA

```r
library(readr)
library(reshape2)
library(tidyverse)
library(ggplot2)
library(fda.usc)
library(fda)
library(ftsa)
library(tictoc) # medir el tiempo de ejecucion
library(fields) # para el grafico de calor (covarianza)

## Lectura de la base de datos y datos 2021, guardados
data <- as.data.frame(read_csv("02Datos/EMBI2021.csv"))
embi2021 <- data[,c('MesDia','2021')]
embi <- as.data.frame(read_csv("02Datos/EMBI.csv")) # sin NA's

rownames(embi) <- embi[,1]
embi <- embi[,-1]
embi <- as.data.frame(embi)

## Transformar a objeto fdata (functional data)
embit <- t(embi) # transponemos
argvals <- 1:365
embif <- fdata(embit, argvals = argvals,
               names = list(main='EMBI',
                            xlab='Mes-Dia',ylab='Anio'))

is.fdata(embif)
dim(embif)
rangeval(embif)
```

```r
## Buscar K y lambda, adecuados para el SUAVIZAMIENTO

l <- c(0, 2^seq(-2, 9, length.out = 30))
length(l)
nb <- seq(7, 41, by = 2)
length(nb)
opt <- optim.basis(embif, lambda = l, numbasis = nb,
                   type.CV = GCV.S, type.basis = 'bspline')

min(opt$gcv)
K <- opt$numbasis.opt
lambda <- opt$lambda.opt

## De los pares de (K,l) se observa el menor GCV
gcv <- as.data.frame(opt$gcv)
colnames(gcv) <- round(as.numeric(colnames(gcv)), 2)

gcv$nb <- as.numeric(rownames(gcv)) ## creo nueva columna
ind1 <- melt(gcv, id.vars = 'nb')

ind1  %>% ggplot(aes(x = nb, y = value,  col = variable, group = variable)) +
  labs(x = "Numero Bases", y = 'GCV')  +
  geom_point()

######################
## Ajuste de los datos
######################

## 1era forma: utilizando lo dado por optim.basis
dataf <- embif
dataf$data <- embif$data %*% opt$S.opt  # producto X*S
plot(dataf, main=" Suavizamiento con B-spline")

## 2da forma: utilizando la funcion S.basis
rango <- c(1,365)
bs <- create.bspline.basis(rangeval = rango,
                           norder = 4,
                           nbasis = K)

S  <-  S.basis(embif$argvals, bs, lambda = lambda)
dataf <- embif
dataf$data <- embif$data %*% S  # producto X*S
color <- c(1:8, 10:22)
plot(dataf, main="Datos Funcionales EMBI", col = color , xlim = c(0,395),xaxt="n")
axis(1, at = seq(0, 365, by = 25), las=2)
legend(370,5700, c('2000', '2001', '2002', '2003', '2004', '2005',
                   '2006', '2007', '2008', '2009', '2010', '2011',
                   '2012', '2013', '2014', '2015', '2016', '2017',
                   '2018', '2019', '2020'),
       col = c(1:8, 10:22), cex = 0.47,  lwd = 1)
text(80, 5400, '2020', cex=0.45)
text(320, 4800, '2008', cex=0.45)
text(0, 5000, '2009', cex=0.45)
text(0, 3100, '2000', cex=0.45)
```

```
## Objeto fd, para ver cada curva y coeficientes de la combinacion lineal
embifd <- fdata2fd(embif, type.basis = 'bspline',
                   nbasis = K,
                   lambda = lambda)

plotfit.fd(t(embit)[,'2000'], 1:365, embifd['2000'])
plotfit.fd(t(embit)[,'2001'], 1:365, embifd['2001'])
plotfit.fd(t(embit)[,'2002'], 1:365, embifd['2002'])
plotfit.fd(t(embit)[,'2003'], 1:365, embifd['2003'])
plotfit.fd(t(embit)[,'2004'], 1:365, embifd['2004'])
plotfit.fd(t(embit)[,'2005'], 1:365, embifd['2005'])
plotfit.fd(t(embit)[,'2006'], 1:365, embifd['2006'])
plotfit.fd(t(embit)[,'2007'], 1:365, embifd['2007'])
plotfit.fd(t(embit)[,'2008'], 1:365, embifd['2008'])
plotfit.fd(t(embit)[,'2009'], 1:365, embifd['2009'])
plotfit.fd(t(embit)[,'2010'], 1:365, embifd['2010'])
plotfit.fd(t(embit)[,'2011'], 1:365, embifd['2011'])
plotfit.fd(t(embit)[,'2012'], 1:365, embifd['2012'])
plotfit.fd(t(embit)[,'2013'], 1:365, embifd['2013'])
plotfit.fd(t(embit)[,'2014'], 1:365, embifd['2014'])
plotfit.fd(t(embit)[,'2015'], 1:365, embifd['2015'])
plotfit.fd(t(embit)[,'2016'], 1:365, embifd['2016'])
plotfit.fd(t(embit)[,'2017'], 1:365, embifd['2017'])
plotfit.fd(t(embit)[,'2018'], 1:365, embifd['2018'])
plotfit.fd(t(embit)[,'2019'], 1:365, embifd['2019'])
plotfit.fd(t(embit)[,'2020'], 1:365, embifd['2020'])

View(embifd$coefs)

## Curvas estimadas por fd
t <- 1:365
phi <- eval.basis(t, bs)
View(phi)  #365*37
View(embifd$coefs) #37*21

Xest <- t(phi%*%embifd$coefs)
# comparacion fdata y fd
View(Xest)
View(opt$fdata.est$data)

######################
## Analisis Exploratorio
######################

### Calculo del area de las curvas: Regla de Simpson
Area <- function(a=1, b=365){
  x <- vector()
  h <- 1
  for(j in 1:21){
    Int <- 0
    x0 <- a
    for(i in 1:364){
      x1 <- x0 + h
      x.med <- (x1 + x0)/2
      Int <- Int + (h/6)*(eval.fd(x0, embifd)[j] +
```

```
                              4*eval.fd(x.med, embifd)[j] +
                              eval.fd(x1, embifd)[j])
      x0 <- x1
    }
    x <- c(x, Int)
  }
  return(x)
}

A <- Area()
A
area <- data.frame('Xi' = 2000:2020, 'Area' = A)
area %>% ggplot(aes(x = as.character(Xi), y = Area)) +
  labs(x = 'X_i(t)',  y= 'Area') +
  geom_point(col = 'blue') +
  theme(axis.text.x = element_text(angle = 45, vjust=1, hjust=1))

### Derivadas
# con nderiv = 0, se tienen las curvas ajustadas originales
der1 <- fdata.deriv(embif, nderiv = 1, nbasis = K,
                    lambda = lambda, method='bspline')
plot(der1, main= "EMBI' ", col = color, xaxt="n")
axis(1, at = seq(0, 365, by = 25), las=2)

der2 <- fdata.deriv(embif, nderiv = 2, nbasis = K,
                    lambda = lambda, method='bspline')
plot(der2, main= "EMBI'' ", col = color, xaxt="n")
axis(1, at = seq(0, 365, by = 25), las=2)

### Deteccion de Datos Funcionales Atipicos
### Fraiman y Muniz
FM <- depth.FM(dataf, trim = 0.15)
FM  # tambien se imprimen los Fn para cada t
FM$dep
FM$lmed # mediana
FM$ltrim # son las curvas que utilizan para sacar la media recortada
plot(FM, ylim = c(0,5900))

### Modal
MD <- depth.mode(dataf, trim = 0.15)
MD$dep
plot(MD, ylim = c(0,6100))
MD$lmed
MD$ltrim

### Algoritmo Bootstrap (ambos metodos)

tic("sleeping")
outFM <- outliers.depth.trim(dataf, dfunc = depth.FM, smo = 0.1,
                             nb = 200, trim = 0.15)
toc()
plot(dataf,col="grey", main = 'Outliers - FM')
lines(dataf[outFM$outliers], type = 'l', lwd = 2, lty = 2, col = 'red')

tic("sleeping")
```

```
outMD <- outliers.depth.trim(dataf, dfunc = depth.mode, smo = 0.1,
                             nb = 200,  trim = 0.15)
toc()
plot(dataf,col="grey", main = 'Outliers - MD')
lines(dataf[outMD$outliers], type = 'l', lty = 2, lwd = 1.2, col = 'red')

### Medidas Estadisticas
embiF <- dataf[-c(1,9,10,21)] # retiro atipicos

## Media
medf <- func.mean(embiF)
plot(embiF, xlim = c(0,400), col = 1:17, main = 'Funcion Media', xlab = NA, ylab = '')
legend(370,2000, c('2001', '2002', '2003', '2004', '2005',
                   '2006', '2007', '2010', '2011', '2012',
                   '2013', '2014', '2015', '2016', '2017',
                   '2018', '2019'),
       col = 1:17, cex = 0.47,  lwd = 1)
plot(medf, col ='red', lwd = 4, add = T)
View(medf$data)
max(medf)
min(medf)

## SD
sdf <- func.var(embiF)^0.5
plot(embiF, ylim= c(200, 2100),  col = 1:17, main = 'Funcion Desviacion Estandar',
     xlab = NA, ylab = NA)
plot(sdf, col ='red', lwd = 4, add = T)
min(sdf)
max(sdf)

## Var-Cov
embiFD <- fdata2fd(embiF, nbasis = K, lambda = lambda)
embiCOV <- var.fd(embiFD)
teval <- eval.bifd(1:365, 1:365, embiCOV)
dev.new()
image.plot(1:365, 1:365, teval, xlab = 'Mes-Dia',
           ylab = 'Mes-Dia', main  = 'Covarianza',
           cex.axis = 1.5, cex.lab = 1.5)

## Correlacion
embiCORR <- cor.fd(1:365, embiFD)
dev.new()
image.plot(1:365, 1:365, embiCORR, xlab = 'Mes-Dia',
           ylab = 'Mes-Dia', main = 'Correlacion',
           cex.axis = 1.5, cex.lab = 1.5)

############################
###### Aplicacion FPCA
############################
embifd1 <- embifd[-c(1,9,10,21)] # reserva

### para comparar y usar despues el f1
embif1 <- fdata(embifd1)
dataf <- embif1
S1  <-  S.basis(embif1$argvals, bs, lambda = lambda)
```

```
dataf$data <- embif1$data %*% S1  # producto X*S
plot(dataf, main="Suavizamiento con B-splines")#, ylim = c(0,3500),lty=2)

plotfit.fd(t(embit)[,'2019'], 1:365, embifd1['2019']) # para comparar
###

embiPCA <- pca.fd(embifd1, nharm = 4)
embiPCA$varprop
varPC <- embiPCA$values  ## Varianza total por componente
embiPCA$harmonics$coefs # coefcientes de la combinacion lineal de las base

# Grafica de valores propios = varianza
ggplot(data = data.frame('nb' = 1:8, 'lambda'=embiPCA$values[1:8]),
       aes(x = nb, y = lambda)) +
  labs(x = 'Componentes',  y= 'Varianza') +
  geom_point(col = 'blue') + geom_line(col = 'blue')

# Grafica del porcentaje acumulado explicado de la variacion total
propVAR <- cumsum(embiPCA$values[1:8])/sum(embiPCA$values)
ggplot(data = data.frame('nb' = 1:8, 'Var'= propVAR),
       aes(x = nb, y = Var)) +
  geom_hline(yintercept=0.98, color='red') +
  labs(x = 'Componentes',  y= 'Porcentaje Varianza Acumulada') +
  geom_point(col = 'blue') + geom_line(col = 'blue')

# Grafica de cada componente principal
FPC <- embiPCA$harmonics
View(FPC$basis)
FPCvals <- eval.fd(1:365, FPC)

plot(1:365, FPCvals[,1], xlab = 'Dia', ylab = 'FPC1',
     lwd = 2, lty = 1,  type = 'l')
plot(1:365, FPCvals[,2], xlab = 'Dia', ylab = 'FPC2',
     lwd = 2, lty = 1, col = 2, type = 'l')
plot(1:365, -FPCvals[,3], xlab = 'Dia', ylab = 'FPC3',
     lwd = 2, lty = 1, col = 3, type = 'l')
plot(1:365, FPCvals[,4], xlab = 'Dia', ylab = 'FPC4',
     lwd = 2, lty = 1, col = 4, type = 'l')

# Todas las FPC
matplot(1:365, FPCvals, xlab = 'Dias', ylab = 'FPCs',
        lwd = 2, lty = 1, type = 'l')
legend(0, -0.05, c('PC1', 'PC2', 'PC3', 'PC4'),
       col = 1:4, lty = 1, lwd = 2, cex = 0.5)

## Perturbaciones en la media
plot.pca.fd(embiPCA, xlab = 'Mes-DIa')

## Proporcion de los scores de las FPC
scores <- embiPCA$scores
rownames(scores) <- rownames(embif1$data)
scores <- sapply(1:4, function(i){scores[,i]^2/(17*varPC[i])})
scores <- as.data.frame(scores)

scores %>% ggplot(aes(x=rownames(scores), y = V1)) +
```

```
  geom_bar(stat="identity", fill = '#BB8FCE', position="stack") +
  labs(x = 'X_i(t)',  y= '%f(1,i)') +
  theme(axis.text.x = element_text(angle = 45, vjust=1, hjust=1))

scores %>% ggplot(aes(x=rownames(scores), y = V2)) +
  geom_bar(stat="identity", fill = '#85C1E9', position="stack") +
  labs(x = 'X_i(t)',  y= '%f(2,i)') +
  theme(axis.text.x = element_text(angle = 45, vjust=1, hjust=1))

scores %>% ggplot(aes(x=rownames(scores), y = V3)) +
  geom_bar(stat="identity",  fill = '#F8C471',position="stack") +
  labs(x = 'X_i(t)',  y= '%f(3,i)') +
  theme(axis.text.x = element_text(angle = 45,  vjust=1, hjust=1))

scores %>% ggplot(aes(x=rownames(scores), y = V4)) +
  geom_bar(stat="identity", fill = '#A3E4D7', position="stack") +
  labs(x = 'X_i(t)',  y= '%f(4,i)') +
  theme(axis.text.x = element_text(angle = 45, vjust=1, hjust=1))


############################
###### Modelo
############################

x <- outMD$Dep[-c(1,9,10,21)]
x  <- x  %>% as.matrix() %>% rownames() %>% as.numeric()
matriz <- t(Xest)[,x]  # tomo la matriz de las curvas ajustadas (m x n)
colnames(matriz1) <- 1:17  ## renombro porque en ftsm debe considerarse un orden
    cronologico.

embifts <- fts(1:365, matriz, xname = "Mes-Dia",
               yname = "Embi anual") # objeto fts
plot(embifts)  # grafico de las curvas (lo antes obtenido)

## Modelo por componentes principales
modelo <- ftsm(embifts, order=4, method = 'classical')
modelo$coeff # puntuaciones de los datos en las componentes (lo antes obtenido)
modelo$y$y      # lo original (ajustado por curvas)
modelo$fitted$y # curvas estimadas por el modelo

  ### SUPUESTOS DEL AJUSTE DEL MODELO
err <- modelo$residuals$y

### Independencia
  ## Rachas
ind.err <- vector()
library(tseries)
for(i in 1:nrow(err)){
  test <- tseries::runs.test(as.factor(err[i,] > mean(err[i,])))$p.value
  ind.err <- c(ind.err, test)
}
ind.err < 0.05

ggplot(data = data.frame('t' = 1:365, 'p.valor'= ind.err),
       aes(x = t, y = p.valor)) +
```

```
  geom_point(col = 'blue') +
  geom_hline(yintercept=0.05, color='red', lwd = 0.5)


  ## Ljung-Box
ind.err <- vector()
for(i in 1:nrow(err)){
  test <- Box.test(err[i,], lag=3, type="Ljung-Box")$p.value
  ind.err <- c(ind.err, test)
}
ind.err < 0.05

ggplot(data = data.frame('t' = 1:365, 'p.valor'= ind.err),
       aes(x = t, y = p.valor)) +
  geom_point(col = 'blue') +
  geom_hline(yintercept=0.05, color='red', lwd = 0.5)

###### Media
med.err <- vector()
for(i in 1:nrow(err)){ med.err <- c(med.err, mean(err[i,])) }

ggplot(data = data.frame('t' = 1:365, 'media.error'= med.err),
       aes(x = t, y = media.error)) + ylim(-0.5,0.5) +
  geom_point(col = 'blue') +
  geom_hline(yintercept=0, color='red', lwd = 1)


library(caret)
y <- as.vector(modelo$fitted$y)
y_ <-as.vector(modelo$y$y)
R2(y , y_)
summary(modelo)
mean((y - y_)^2)
-2*log(mean((y - y_)^2)/(17*365)) + 2*4
-2*log(mean((y - y_)^2)/(17*365)) + 4*log(17*365)

##### MEJOR CASO

## Prediccion de la funcion 2021+
deps <- outMD$Dep[-c(1,9,10,21)] # tomo las medidas de profundidad (sin considerar los ya
    atipicos)
orden.asc <- deps %>% sort()  %>% as.matrix() %>% rownames() %>% as.numeric()
# Ascendente: las ultimas tienen mas profundidad,
# por tanto son las curvas con comportamiento mas normal.

matriz1 <- t(Xest)[,orden.asc]
colnames(matriz1) <- 1:17

embifts1 <- fts(1:365, matriz1, xname = "Mes-Dia",
                yname = "Embi anual") # objeto fts
modelo1 <- ftsm(embifts1, order=4, method = 'classical')


a <- 0.6 # para el suavizado exponencial
pred1 <- forecast(modelo1, h=1, method = 'ets', alpha = a, model = 'ANN',
```

```
                  jumpchoice = 'actual', level = 0)

plot(pred1, 'components', xlab2 = 'Datos Funcionales',
     ylab2 = 'puntuaciones') # serie de las puntuaciones
  # dibuja la curva estimada, las FPC y las series

# Datos funcionales
plot(embifts1, col = gray(0.8), xlab = "Dia", ylab = "Puntuaciones EMBI")
# Prediccion de la funcion
plot(pred1,  lwd = 3, col = 'blue',  add = TRUE)
legend("topleft", "2021", col = "blue",  cex = 0.7,  lwd = 1)

min(pred1$mean$y)
max(pred1$mean$y)

# Comparacion con datos orginales 2021
datos2021 <- embi2021[1:217,2] # 217: 05 de agosto
data.frame('Dias' = 1:217, 'EMBI_2021' = datos2021) %%
  ggplot(aes(Dias, EMBI_2021)) +
  geom_line(col = 'blue')

data.frame(EMBI_2021 = datos2021, Prediccion_EMBI_2021 = pred1$mean$y[1:217]) %%
  ggplot(aes(EMBI_2021, Prediccion_EMBI_2021)) +
  geom_point()

### Precision
mejor <- data.frame(embi = datos2021,  pred =  pred1$mean$y[1:217])
# tomo los valores para los cuales se observan
# mejores predicciones
mejorVal <- mejor %% dplyr::filter(embi < 1000)

View(mejorVal)
mejorVal %% ggplot(aes(embi, pred)) + geom_point()

obs <- nrow(mejorVal)
MAPE1 <- (100/obs)*sum(abs((mejorVal[,1])- mejorVal[,2])/abs(mejorVal[,1]))

###### Algoritmo para Bandas de confianza mediante bootstrap suavizado

# atp: vector de las ubicaciones de los datos sin atipicos
# X: la matriz de las funciones estimadas (21x365), sin retirar atipicos
# outD: lo calculado por el algoritmo de outliers segun la medida de profundidad usada.
# orderD: tipo de orden para las profundidades de los datos
  # si FALSE: orden ascendente; si TRUE: orden descendente
# nb: numero de muestras bootstrap
# smo: parametro de bootstrap suavizado
# alp: para el suavizado exponencial

N <- 200
IC.bootstrap <- function(atp = c(2:8,11:20), X, outD, orderD=F,
                         nb = N, smo = 0.02, alp = a){
  n <- length(atp)  # longitud para una muestra
  muestras <- list()
  for(j in (1:nb)){
    muestras[[j]] <- sample(atp, size=n, replace=T)
```

```
  }  # guardo en una lista cada muestra sin atipicos.

  preds <- list()
  basef <- fdata(X, argvals = 1:365)
  varX <- var(basef[atp]$data)*smo

  for(i in (1:nb)){
    prof <- outD$Dep[muestras[[i]]]  # no se consideran los atipicos
    # obtengo las profundidades de los datos de la remuestra
    orden.D <- prof %>% sort(decreasing = orderD) %>% as.matrix() %>%
      rownames() %>% as.numeric()
    # se ordenan las profundidades y obtengo sus ubicaciones

    dataB <- basef[orden.D] # para suavizar las muestras bootstrap
    nr <- nrow(dataB)
    nc <- ncol(dataB)
    if(smo > 0){
      dataB$data <- dataB$data + mvrnorm(n = nr, rep(0, nc), varX)
    }

    matriz.boot <- t(dataB$data)
    # obtengo la matriz de la muestra, ordenada segun las profundidades
    colnames(matriz.boot) <- 1:n

    # Reformulo el modelo
    embifts.boot <- fts(1:365, matriz.boot)

    modelo.boot <- ftsm(embifts.boot, order=3, method = 'classical')

    pred.boot <- forecast(modelo.boot, h=1, method = 'ets', alpha = alp, model = 'ANN',
                          jumpchoice = 'actual')

    preds[[i]] <- pred.boot ## cada elemento de la lista contiene el modelo
  }
  return(preds)
}

boot1 <- IC.bootstrap(X = Xest, outD = outMD, orderD= F) # ascendente -> mejor caso

# Calculo la distancia de la curva pronosticada con lo estimado por las remuestras
est.muestras1 <- matrix(nrow = N, ncol = 365)
for(i in 1:N){ est.muestras1[i, ] <- boot1[[i]]$mean$y }

estimacion1 <- fdata(t(pred1$mean$y) , argvals = 1:365) # prediccion del modelo original
B1 <- fdata(est.muestras1 , argvals = 1:365) # prediccion con las remuestras
distboot1 <- as.vector(metric.lp(estimacion1, B1))

# Nivel de confianza del 90%
q1 <- quantile(distboot1, 0.90)
nivel.confianza1 <- which(distboot1 <  q1)

# Grafico de la Banda de confianza
plot(pred1, ylim=c(400,1100), lwd = 5, col = 'white')
for(i in nivel.confianza1){
  plot(boot1[[i]], add=T, col = 'grey', lwd = 10)
```

```
}
plot(pred1, lwd = 3, col = 'blue', add=T)


##### PEOR CASO

orden.desc <- deps %>% sort(decreasing = T) %>% as.matrix() %>% rownames() %>% as.numeric()
# Descendente: las ultimas tienen menos profundidad,
# por tanto, las que tienen puntuaciones EMBI altas.

matriz2 <- t(Xest)[,orden.desc]
colnames(matriz2) <- 1:17

embifts2 <- fts(1:365, matriz2, xname = "Mes-Dia", yname = "Embi anual")
modelo2 <- ftsm(embifts2, order=4, method = 'classical')
pred2 <- forecast(modelo2, h=1, method = 'ets', alpha = a, model = 'ANN',
                  jumpchoice = 'actual', level=0)

plot(pred2, 'components', xlab2 = 'Datos Funcionales', ylab2 = 'puntuaciones')

plot(embifts2, col = gray(0.8), xlab = "Dia", ylab = "Puntuaciones EMBI")
# Prediccion de la funcion
plot(pred2, lwd = 3, col = 'blue',  add = TRUE)
legend("topleft", "2021", col = "blue",  cex = 0.6,  lwd = 1)

min(pred2$mean$y)
max(pred2$mean$y)

data.frame(EMBI_2021 = datos2021, Prediccion_EMBI_2021 = pred2$mean$y[1:217]) %>%
  ggplot(aes(EMBI_2021, Prediccion_EMBI_2021)) +
  geom_point()

# MSE
mejor <- data.frame(embi = datos2021,  pred =  pred2$mean$y[1:217])
# tomo los valores para los cuales se observan
# mejores predicciones
mejorVal <- mejor %>% dplyr::filter(embi > 1000)
mejorVal %>% ggplot(aes(embi, pred)) + geom_point()

obs <- nrow(mejorVal)
MAPE2 <- (100/obs)*sum(abs((mejorVal[,1])- mejorVal[,2])/abs(mejorVal[,1]))

# Banda de confianza
boot2 <- IC.bootstrap(X = Xest, outD = outMD, orderD= T) # descendente, peor caso

est.muestras2 <- matrix(nrow = N, ncol = 365)
for(i in 1:N){ est.muestras2[i, ] <- boot2[[i]]$mean$y }

estimacion2 <- fdata(t(pred2$mean$y) , argvals = 1:365)
B2 <- fdata(est.muestras2 , argvals = 1:365)
distboot2 <- as.vector(metric.lp(estimacion2, B2))
q2 <- quantile(distboot2, 0.90)
nivel.confianza2 <- which(distboot2 <  q2)

plot(pred2, ylim=c(800,2100), lwd = 5, col = 'white')
```

```
for(i in nivel.confianza2){
  plot(boot2[[i]], add=T, col = 'grey', lwd = 10)
}
plot(pred2, lwd = 3, col = 'blue', add=T)
```