



ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ciencias Naturales y Matemáticas

“Imputación basada en modelos de machine learning y
análisis espectral del nivel del mar de La Libertad”

PROYECTO INTEGRADOR

Previo la obtención del Título de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

Presentado por:

Cristhian Marcelo Figueroa Quezada

GUAYAQUIL - ECUADOR

Año: 2021

DEDICATORIA

A mis mamás: FLOR MARÍA
EMPERATRIZ CASTRO, ROSARIO AMELIA
QUEZADA CASTRO, y GINA ELIZABETH
FIGUEROA QUEZADA.

A mi ñaño: JAIME MAURICIO
FIGUEROA QUEZADA.

A mi papi: JAIME ALFONSO
FIGUEROA LÓPEZ.

AGRADECIMIENTOS

A mi mami y a mi ñaño por todo el trabajo y sacrificio a lo largo de todos estos años.

A mi mami Florcita y mi mami Charito porque sé que rezan e interceden por mí desde el cielo.

A mi papi Jaime por brindarnos su apoyo y compañía a mi mami y a mí, mientras estuvo junto a nosotros.

A la familia Figueroa y a la familia Quezada por estar siempre presentes en las buenas y en las malas para mi mami, mi ñaño, y para mí.

A la familia Ramírez-Ramos por abrirnos las puertas de su hogar y brindarnos su hospitalidad y ayuda desde el primer día.

A mis amigos por darme su apoyo incondicional en los momentos más complicados.

A mis profesores de ESPOL porque aprendí mucho de ellos en varios ámbitos.

A mi tutora por su compromiso y guía en la realización de esta tesis de grado.


DECLARACIÓN EXPRESA

"Los derechos de titularidad y explotación, me corresponde conforme al reglamento de propiedad intelectual de la institución; Yo, Cristhian Marcelo Figueroa Quezada doy mi consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

A handwritten signature in blue ink, consisting of a vertical line with several horizontal strokes crossing it, forming a stylized, abstract mark.

Cristhian Marcelo Figueroa Quezada

EVALUADORES



Ph.D. Sandra Lorena García Bustos

PROFESOR DE LA MATERIA



Ph.D. Mariela Alexandra González Narváez

PROFESOR TUTOR

RESUMEN

El aumento en el nivel del mar supone efectos negativos a nivel socio-económico alrededor del mundo, por ello este estudio tiene como objetivo identificar la frecuencia dominante del nivel del mar en la estación fija de La Libertad-Ecuador, a través del análisis espectral singular (SSA), para determinar el componente periódico y su fenómeno generador de fluctuaciones más influyente en la variabilidad del nivel del mar. Además, es usual encontrarse con series de tiempo incompletas en estaciones mareográficas como La Libertad, por lo cual se propone un método de imputación basado en bosques aleatorios para series de tiempo univariantes. Para ello, se empleó los datos horarios del nivel del mar de La Libertad, obtenidos de la UHSLC, en el período de Septiembre/1949 a Enero/2021. Se partió de la imputación para la serie horaria, luego se empleó el filtro de Doodson para determinar la serie diaria y, consecuentemente, la serie de anomalías mensuales. A esta última se le aplicó el SSA para estimar sus componentes. El método de imputación propuesto capturó la estructura subyacente de la serie sin alterar su comportamiento natural. Por otro lado, la tendencia lineal estimada correspondiente al período 1993-2020, permitió identificar un incremento aproximado de $3.2 \pm 0.12 \text{ mm/año}$ en el nivel del mar esperado. Más aún, se espera que el nivel del mar haya aumentado a una tasa más rápida en la última década ($4.05 \pm 0.35 \text{ mm/año}$). Finalmente, se determinó que el nivel del mar está predominado por el componente periódico con periodicidad 3.6 años asociado al fenómeno interanual ENOS.

Palabras claves: Nivel del Mar, Análisis Espectral, Análisis Espectral Singular, Valores Perdidos, Imputación, Machine Learning.

ABSTRACT

Sea level rise brings with it negative effects at socioeconomic level worldwide, as a result the objective of this study is to identify the dominant frequency of sea level at La Libertad-Ecuador gauge-station, through the singular spectrum analysis (SSA), in order to determine the periodic component and its fluctuations phenomenon generator with more influence in sea level variability. Additionally, it's pretty common to deal with incomplete time series at gauge-stations like La Libertad, in consequence an imputation method for univariate time series based on random forests is proposed. Therefore, hourly sea level data from La Libertad over the period of September/1949 to January/2021 were obtained from the UHSLC. The hourly time series imputation was first taken placed, then the Doodson filter was applied to determine the daily time series and, consequently, the monthly time series of mean sea level anomalies was obtained. The latter was used to estimate its components through SSA. The imputation method proposed captured the underlying structure of the time series without changing its natural form. Moreover, the linear trend estimated over the period 1993-2020 allowed to identify an expected sea level rise at a rate of approximately $3.2 \pm 0.12 \text{ mm/year}$. Furthermore, sea level is expected to have risen at a faster rate in the last decade ($4.05 \pm 0.35 \text{ mm/year}$). Finally, sea level was found dominated by the periodic component with periodicity of 3.6 years associate to the interannual ENSO phenomenon.

Keywords: Sea level, Spectral Analysis, Singular Spectrum Analysis, Missing Data, Imputation, Machine Learning.

ÍNDICE GENERAL

RESUMEN	6
ABSTRACT	7
ÍNDICE GENERAL	8
ABREVIATURAS.....	11
SIMBOLOGÍA.....	13
ÍNDICE DE FIGURAS	14
ÍNDICE DE TABLAS.....	16
CAPÍTULO 1	17
1. Introducción	17
1.1 Descripción del problema	18
1.2 Justificación del problema	20
1.3 Datos.....	23
1.4 Objetivos	23
1.4.1 Objetivo General	23
1.4.2 Objetivos Específicos.....	24
1.5 Marco teórico	24
1.5.1 Nivel del mar	24
1.5.2 Series de tiempo univariantes	25
1.5.3 Machine Learning	28

1.5.4	Filtro de Doodson.....	30
1.5.5	Análisis Espectral Singular (SSA)	31
1.5.6	Frecuencia dominante.....	32
1.6	Estado del arte	32
1.6.1	Imputación para series de tiempo univariantes	32
1.6.2	Análisis espectral del nivel del mar	35
CAPÍTULO 2		39
2.	Metodología.....	39
2.1	Bosques Aleatorios (Random Forests).....	39
2.2	Imputación para series de tiempo univariantes basada en bosques aleatorios.....	40
2.3	Filtro de Doodson	42
2.4	Métricas de validación.....	42
2.5	Análisis Espectral Singular	44
2.6	Regresión lineal simple	45
2.7	Periodograma.....	46
CAPÍTULO 3		49
3.	RESULTADOS Y ANÁLISIS	49
3.1	Análisis exploratorio	49
3.2	Imputación basada en bosques aleatorios	54
3.3	Filtro de Doodson	59
3.4	Análisis Espectral Singular	63

CAPÍTULO 4	77
4. CONCLUSIONES Y RECOMENDACIONES	77
4.1 Conclusiones.....	77
4.2 Recomendaciones.....	78
BIBLIOGRAFÍA	80
APÉNDICES.....	84

ABREVIATURAS

UHSLC	University of Hawaii Sea Level Center
FD	Fast Delivery data
RQD	Research Quality Data
MCAR	Missing Completely At Random
MAR	Missing At Random
NMAR	Not Missing At Random
PSMSL	Permanent Service for Mean Sea Level
UNESCO	United Nations Educational, Scientific and Cultural Organization
IOC	Intergovernmental Oceanographic Commission
SSA	Singular Spectrum Analysis
PSF	Pattern Sequence Forecasting
SVM	Support Vector Machine
LOCF	Last Observation Carried Forward
ARIMA	Autoregressive Integrated Moving Average
RMSE	Root Mean Square Error
MAPE	Mean Absolute Percentage Error
NMAE	Normalized Mean Absolute Error
MAE	Mean Absolute Error
FSD	Fraction of Standard Deviation
SD	Standard Deviation
FB	Fractional Bias
DTWBI	Dynamic Time Warping Based Imputation
eDTWBI	extension Dynamic Time Warping Based Imputation

WMA	Wavelet Multiresolution Analysis
BW	Bandwidth
SVD	Singular Value Decomposition
NCEI	National Centers for Environmental Information
FEO	Funciones Empíricas Ortogonales
ENOS	El Niño-Oscilación del Sur

SIMBOLOGÍA

Mm	Milímetros
Cm	Centímetros
Cpm	Ciclos por mes

ÍNDICE DE FIGURAS

Figura 3.1.1 Serie de tiempo nivel del mar horario de La Libertad	50
Figura 3.1.2 Diagrama de caja del nivel del mar horario	50
Figura 3.1.3 Histograma del nivel del mar horario	51
Figura 3.1.4 Valores perdidos por períodos	52
Figura 3.1.5 Presencia de gaps por períodos	53
Figura 3.1.6 Tamaños de los gaps	54
Figura 3.2.1 Serie de tiempo nivel del mar horario imputada	55
Figura 3.2.2 Serie de tiempo nivel del mar horario-Validación con 100 valores perdidos consecutivos simulados	56
Figura 3.2.3 Serie de tiempo nivel del mar horario-Validación con 500 valores perdidos consecutivos simulados	57
Figura 3.3.1 Serie de tiempo nivel medio del mar diario mediante el filtro de Doodson	59
Figura 3.3.2 Serie de tiempo nivel medio del mar diario-Filtro de Doodson vs UHSLC	60
Figura 3.3.3 (a) Serie de tiempo nivel medio del mar mensual, (b) Serie de tiempo anomalías del nivel medio del mar mensual.....	62
Figura 3.4.1 Valores propios para el SSA con $L=48$	64
Figura 3.4.2 Primer vector propio para el SSA con $L=48$	65
Figura 3.4.3 Tendencia no lineal estimada para la serie anomalías del nivel medio del mar mensual	65
Figura 3.4.4 Tendencias lineales estimadas para la serie anomalías del nivel medio del mar mensual	67
Figura 3.4.5 Serie de tiempo residual	68
Figura 3.4.6 Valores propios para el SSA con $L=432$	69

Figura 3.4.7 Periodograma suavizado	70
Figura 3.4.8 Vectores propios para el SSA con $L=432$	73
Figura 3.4.9 Componentes periódicos 1,2 y 3 estimados para la serie anomalías del nivel medio del mar mensual.....	74
Figura 3.4.10 Componentes periódicos 4 y 5 estimados para la serie anomalías del nivel medio del mar mensual.....	75
Figura 3.4.11 Serie reconstruida a partir de los componentes estimados por SSA	76

ÍNDICE DE TABLAS

Tabla 3.1.1 Estadísticos descriptivos del nivel del mar horario	51
Tabla 3.2.1 Métricas para la validación de la imputación	58
Tabla 3.2.2 Estadísticos descriptivos de la serie de tiempo nivel del mar horario y su imputación	58
Tabla 3.3.1 Métricas para la comparación entre el filtro de Doodson y la UHSLC	61
Tabla 3.4.1 Frecuencias, periodicidades y espectros de potencia de los componentes periódicos ordenados según su importancia	71

CAPÍTULO 1

1. INTRODUCCIÓN

Según la University of Hawaii Sea Level Center (s.f.-a), conocido por sus siglas en inglés UHSLC, el nivel del mar es considerado como, posiblemente, la mejor medida del calentamiento global, y su aumento se asocia tanto a la expansión del océano, que se debe al incremento en su temperatura, como al derretimiento de los glaciares en la tierra. A día de hoy, en el siglo XXI, estamos experimentando esta creciente en el nivel del mar, la cual tendrá un impacto negativo en las infraestructuras, ecosistemas, y vidas de millones de personas. Por lo expuesto, es primordial basarse estudios del nivel del mar, tanto históricamente como en el futuro, para así obtener un panorama mucho más claro sobre los procesos físicos que generan variabilidad en el nivel del mar y en la evolución del clima.

Oceanógrafos y científicos del clima hacen uso de instrumentos especiales para medir el nivel del mar en una localidad de interés, estos toman el nombre de mareógrafos. Dichos instrumentos de medición se los encuentra en puntos estratégicos de las zonas costeras y miden el nivel del mar como la diferencia de alturas entre el mar y una referencia, que en este caso es la tierra (University of Hawaii Sea Level Center, s.f.-a).

A partir de la década de 1830 con el invento del primer mareógrafo automático, se ha evidenciado varios instrumentos de medición hasta el día de hoy, entre estos tenemos: mareógrafos de flotador, acústicos, de presión y de radar. Muchos de los antes mencionados, presentan errores sistemáticos y/o sesgo en sus mediciones; en particular, los medidores de radar, que han venido tomando apogeo gracias al avance tecnológico, se ven afectados en su precisión y/o sistemas electrónicos incorporados por factores como: cambios de temperatura,

humedad, deterioro de su vida útil, interferencias electromagnéticas, entre otros (UNESCO/IOC, 2016).

Junninen et al. (2004), mencionan que los errores en las mediciones, ausencia de mediciones o errores en el registro de mediciones, son algunos de los factores que causan la pérdida de datos, en consecuencia, es muy común encontrarse con bases de datos incompletas en investigaciones ambientales.

Por lo cual, la motivación de este proyecto es proponer una metodología de imputación para series de tiempo univariantes, basada en modelos de machine learning, que permita el tratamiento de valores perdidos presentes en los registros del nivel del mar de la estación fija de la Libertad, para posteriormente realizar un análisis espectral en busca del componente periódico más importante en su variabilidad, que bajo sus características periódicas permita identificar el potencial proceso físico generador de esta oscilación.

1.1 Descripción del problema

Por lo general, siempre que se deba medir y registrar datos, nos enfrentaremos a valores perdidos, ya sea porque estos valores no fueron medidos, fueron medidos, pero se perdieron, o fueron medidos, pero no se los utiliza por alguna incidencia (Moritz et al., 2015).

Estaciones costeras como La Libertad en Ecuador, registran las mediciones realizadas por los mareógrafos en sus bases de datos. Sin embargo, “usualmente están incompletos debido a fallos en los sensores, problemas de comunicación/transmisión, o malas condiciones climáticas para medir o para realizar el mantenimiento manual. Este es el caso particular de muestras marinas” (Rousseeuw et al., 2013; Ceong et al., 2012, citado en Phan et al., 2017).

En la meteorología, las series de tiempo pueden presentar valores perdidos aislados o consecutivos en intervalos de tiempo de tamaño pequeño, mediano, grande, o muy grande (Flores, Tito, & Silva, 2019). Por tanto, la importancia de completar los datos correctamente

se da porque la presencia de valores perdidos afecta negativamente en el desempeño de, por ejemplo, las predicciones (Flores, Tito, & Centty, 2019).

Por otro lado, existe evidencia de cómo ha ido variando el nivel global del mar debido al cambio climático desde hace miles de años (Braker, 1994); entre el siglo XX y XXI, la tasa de cambio del nivel global del mar ha incrementado en 2.4 veces, para el 2015 esta tasa alcanzó los 3.6 mm por año. Para finales de este siglo se estima que el nivel del mar aumente entre 0.61 y 1.10 metros, si no se reduce las emisiones de gas invernadero en el mundo (Oppenheimer et al., 2019).

Oppenheimer et al. (2019) estiman que para el 2050, varias ciudades costeras e islas pequeñas podrían enfrentar graves riesgos por el cambio de periodicidad decadal a periodicidad anual de eventos meteorológicos extremos, debido al aumento en el nivel global del mar. Asimismo, los fenómenos que solían suceder muy rara vez ocurrirán de forma más recurrente para el 2100. Es decir, que el fenómeno El Niño-Oscilación del Sur (ENOS), que ha provocado daños catastróficos en el territorio ecuatoriano, podría ocurrir con más frecuencia.

El Ministerio del Ambiente del Ecuador (2019), a través de la Primera Contribución Determinada a Nivel Nacional para el Acuerdo de París bajo la Convención Marco de Naciones Unidas sobre Cambio Climático, afirma la evidencia en el aumento de temperatura, variabilidad estacional y espacial de la precipitación entre 1960 y 2010 a nivel nacional; estas precipitaciones se ven afectadas por el fenómeno ENOS causando así sequías e inundaciones. Estudios alrededor del mundo sobre el nivel del mar estiman un incremento que, en el territorio ecuatoriano, puede dar lugar a muchas más inundaciones, erosión acelerada en las zonas costeras y salinización de acuíferos y tramos finales de ríos.

1.2 Justificación del problema

La presencia de valores perdidos puede dificultar el análisis y procesamiento de datos, puesto que se basan en datos completos (Moritz et al., 2015), además, ocasiona variabilidad en los resultados y análisis, lo que los hace poco confiables (Hawthorne & Elliott, 2005). Otras consecuencias de lo enunciado son el sesgo asociado a las diferencias sistemáticas entre los datos observados y no observados, y las limitaciones en las posibles técnicas y modelos estadísticos que podemos emplear para el análisis de series de tiempo ya que, gran parte de ellos, necesitan datos completos (Noor et al., 2015).

En series de tiempo, se asume la dependencia de un valor con sus valores pasados. Este supuesto trae problemas en presencia de valores perdidos, pues al no darle un tratamiento adecuado a los datos e ignorarlos, perdemos información, y esto condiciona a las predicciones que queramos realizar (Junninen et al., 2004). Esta pérdida de información se evidencia en la pérdida de eficiencia (Noor et al., 2015). La limitación que nos provee los datos incompletos no solo se refleja en los algoritmos de análisis, sino en los softwares estadísticos que podemos utilizar, pues en su mayoría no soportan valores perdidos (Phan et al., 2017).

A las técnicas que nos ayudan a solucionar estos inconvenientes de manera adecuada se las conoce como imputación de datos (Junninen et al., 2004), las cuales deben garantizar resultados eficientes y confiables, es decir, efectivos y que respeten la forma de la serie de tiempo (Phan et al., 2017).

Cuando no se imputa correctamente, la estructura propia de los datos se ve comprometida y el desempeño del modelo estadístico puede decaer considerablemente (Junninen et al., 2004). De ahí que, los valores perdidos deben ser reemplazados con valores adecuados a través de la imputación de datos (Moritz et al., 2015). Así, es preciso proponer métodos de imputación para series de tiempo univariantes que capturen las características

propias de los datos, sobre todo cuando tratamos con datos que muestran un comportamiento complejo (Phan et al., 2020).

De esta manera, una vez que contemos con los datos completos de la serie de tiempo podemos utilizar cualquier tipo de técnica o modelo estadístico para llevar a cabo análisis o predicciones, puesto que hemos superado las limitaciones que nos imponía la presencia de valores perdidos. Para este proyecto en particular, el análisis espectral del nivel del mar de La Libertad se puede desarrollar sin ningún contratiempo una vez imputados los datos.

Por otra parte, los datos provenientes de la altimetría por satélite, desde su llegada en 1993, nos brindan información altamente precisa sobre el nivel global del mar (Khelifa et al., 2016); de modo que, se tasó su aumento en aproximadamente 3.3 ± 0.4 mm por año desde la era de la altimetría (Cazenave et al., 2014). Esta tasa es distinta entre localidades, es por esto que es fundamental identificar los lugares más vulnerables y tomar una postura preventiva según los estudios sobre el nivel del mar que se realicen (Beşel & Tanır Kayıkçı, 2020).

Para el caso particular de nuestro país Ecuador, el Banco de Desarrollo de América Latina (2017) señala que Guayaquil es una de las ciudades costeras más vulnerables al incremento del nivel del mar debido al cambio climático, pues su ubicación geográfica, características biofísicas, alta densidad poblacional e impermeabilización en sus parroquias, y deforestación asociada al crecimiento urbano, elevan su sensibilidad frente al cambio climático tanto ambiental como socioeconómicamente. Según modelos climáticos, entre 2020 y 2099, se evidenciará un mayor número de sequías, inundaciones, precipitaciones y escorrentía con gran intensidad; además, se prevé un aumento en el nivel del mar. En la mayoría de parroquias de la ciudad hay un riesgo considerable de inundación, solo entre 2012 y 2015, ocurrieron 79 inundaciones en la “Perla del Pacífico” (Guayaquil). Estas inundaciones han causado daños a infraestructuras industriales y ha acabado con la vida de

varias personas. Asimismo, las lluvias provocan desbordamiento de ríos que acaban con las zonas de cultivo y afectan negativamente a industrias, por ejemplo, las camaroneras.

Hallegatte et al.(2013), asumiendo un escenario optimista que envuelve a varios factores, entre ellos un aumento de 20 cm en el nivel del mar, estiman a Guayaquil como la cuarta ciudad costera con más pérdidas económicas promedio en el mundo a causa de las inundaciones para mediados de siglo, tasándose en la increíble cifra de \$3 189' 000 000 por año, superando a ciudades como: Miami, Nueva York y Bangkok.

Las potenciales causas de la variabilidad del nivel del mar son centro de investigación de varios estudios a nivel mundial (Khelifa et al., 2016); de tal manera que el propósito de estudio de series de tiempo es el de buscar evidencia de un posible incremento o decremento en el nivel del mar de un lugar en específico que nos conduzca a un potencial indicio de variación en el nivel global del mar, ya que, es posible adquirir información del nivel global del mar a partir del nivel del mar relativo. Y que, profundizando en su análisis, podamos asociar los componentes oscilatorios de la serie temporal con fenómenos oscilatorios conocidos por la comunidad científica (Braker, 1994).

Dentro de este análisis, es importante el estudio de la frecuencia dominante, ya que nos ayuda a realizar mejores diagnósticos, predicciones, e inferir en posibles consecuencias asociadas a su predominancia en cualquier campo de estudio (Telgársky, 2013).

En consecuencia, es pertinente estudiar al nivel del mar desde el análisis de frecuencias dominantes con el fin de identificar los componentes periódicos asociados a los fenómenos oscilatorios generadores de fluctuaciones en el nivel del mar, de tal forma que podamos tomar decisiones oportunas ante la adaptación y mitigación del cambio climático y sus repercusiones socio-económicas.

1.3 Datos

Los datos fueron tomados de la UHSLC y corresponden a los registros horarios del nivel del mar en la estación fija de La Libertad-Ecuador. El período de datos horarios comprende desde las 5 horas del 1 de septiembre de 1949 hasta las 23 horas del 31 de enero de 2021, con un total de 626059 observaciones y 96.96% de datos válidos. Además, se utilizó los registros diarios de la UHSLC para comparar con los resultados del filtrado de datos horarios a datos diarios mediante el filtro de Doodson. El período de datos diarios comprende desde el 2 de septiembre de 1949 hasta el 31 de enero de 2021, con un total de 26085 observaciones y 96.80% de datos válidos. A estos datos se los conoce como datos de entrega rápida, Fast Delivery (FD) data en inglés, y han pasado por un control de calidad básico. Por otra parte, están los datos de calidad de la investigación, Research Quality Data (RQD) en inglés, y han sido sometidos a un control de calidad exhaustivo, por esta razón su publicación tarda entre 1 a 2 años. Sin embargo, los datos FD cuentan con los datos RQD conforme su disponibilidad. Así, los sets de datos con los que se trabajó para este proyecto cuentan con datos RQD en el período de septiembre de 1949 hasta diciembre de 2018, y datos FD en el período de enero de 2019 hasta enero de 2021 (Caldwell et al., 2015; University of Hawaii Sea Level Center, s.f.-b).

1.4 Objetivos

1.4.1 Objetivo General

Identificar la frecuencia dominante del nivel del mar, imputada bajo modelos de machine learning, de la estación fija de La Libertad-Ecuador a través del análisis espectral singular, para la detección del componente periódico más influyente en la variabilidad del nivel del mar y su potencial fenómeno asociado.

1.4.2 Objetivos Específicos

- Proponer una metodología de imputación para series de tiempo univariantes basada en modelos de machine learning, que capture sus características propias y permita el posterior análisis espectral singular.
- Calcular la serie nivel medio del mar diario, a partir del filtro de Doodson en la serie imputada, que permita la posterior obtención de la serie anomalías del nivel medio del mar mensual.
- Seleccionar adecuadamente las matrices asociadas a los eigentriples mediante sus valores y vectores propios, de tal forma que sus agrupamientos permitan la estimación de los componentes de la serie anomalías del nivel medio del mar mensual.
- Comparar la serie resultante de la adición de los componentes reconstruidos más importantes con la serie anomalías del nivel medio del mar mensual, para la validación de la descomposición realizada.
- Estimar la densidad espectral de la serie anomalías del nivel medio del mensual mediante el periodograma, para la evaluación de los espectros de potencia de los componentes periódicos.

1.5 Marco teórico

1.5.1 Nivel del mar

La variabilidad en el nivel del mar tiene un impacto considerable en los habitantes de zonas costeras (e.g. erosión e inundaciones) y en el cambio climático en general; de ahí su relación con la expansión de los océanos, el deshielo de glaciares y los cambios evidenciados en las corrientes oceánicas. Por lo descrito, al nivel del mar se le considera como un indicador importante del cambio climático. Una medición del nivel del mar es, aproximadamente, el resultado de la suma de tres componentes: nivel medio del mar,

marea y residuales meteorológicos. Cada componente está sujeto a un proceso físico distinto y su variabilidad es independiente de los demás. El nivel medio del mar es una medición promedio del nivel del mar que es calculada a partir de datos horarios en un período de tiempo conformado por varios años. En cambio, las mareas son los movimientos periódicos que se producen en los mares y están asociados con fuerzas geofísicas de cierta periodicidad; según la fuerza asociada las mareas pueden ser gravitacionales o meteorológicas. Por otro lado, los residuales meteorológicos son el resultado de extraer las mareas y muestran un comportamiento irregular (UNESCO/IOC, 1985).

1.5.2 Series de tiempo univariantes

Una serie de tiempo univariante es una secuencia de observaciones singulares ordenadas en intervalos de tiempo equiespaciados. En este contexto, al tiempo se le considera como una variable implícita y discreta, más allá que el término “univariante” se refiera a que hay una única variable de interés. Las series de tiempo pueden ser analizadas a través de componentes que describen sus características subyacentes, estas son: componente de tendencia, componente estacional y componente residual. La tendencia describe el incremento o decremento del nivel de la serie a largo plazo. Por otro lado, la estacionalidad refleja el patrón de cambio sistemático subyacente en la serie asociado con períodos fijos, esto es, con fechas calendario (4 meses, 6 meses, 12 meses, 7 días, etc.). Por último, el residual es el resultado de extraer los anteriores componentes en la serie original y muestra las influencias irregulares (Moritz et al., 2015).

1.5.2.1 Valores perdidos

1.5.2.1.1 Mecanismos de pérdida de datos

La distribución de los valores perdidos está asociada con los mecanismos de pérdida de datos, estos están clasificados en: missing completely at random (MCAR), missing at random (MAR) y not missing at random (NMAR). MCAR se refiere a que la pérdida de datos es meramente de forma aleatoria, esto es, que la probabilidad de que cierta observación se pierda es independiente tanto del tiempo asociado a esta observación como de su propio valor. En cambio, MAR establece que la probabilidad de que cierta observación se pierda es únicamente independiente de su propio valor, más no del punto en el tiempo asociado a esta observación. Finalmente, NMAR se refiere a que la probabilidad de que cierta observación se pierda es dependiente de su propio valor, pero esta dependencia se puede extender hasta una dependencia tanto de su propio valor como la del tiempo. La importancia de conocer estos mecanismos se fundamenta en que en base a ellos podemos seleccionar técnicas de imputación apropiadas, aunque categorizarlos puede resultar en una tarea no muy sencilla (Moritz et al., 2015).

1.5.2.1.2 Gaps

Un gap está constituido por uno o más valores perdidos consecutivos. En series de tiempo podemos encontrar gaps de distintos tamaños que se pueden categorizar como: gaps pequeños, medianos, grandes y muy grandes. Los gaps pequeños pueden ser un valor perdido aislado o 2 valores perdidos consecutivos. A su vez, los gaps medianos van de 3 a 10 valores perdidos consecutivos. En cambio, los gaps

grandes son aquellos gaps con más de 10 valores perdidos consecutivos. Por último, los gaps muy grandes van entre 1 y 72 meses con presencia de valores perdidos consecutivos (e.g. 1 978 valores perdidos consecutivos en 72 meses de registros) (Flores, Tito, & Silva, 2019).

1.5.2.2 Imputación

Cuando tratamos con series de tiempo univariantes es importante realizar las imputaciones en base al tiempo, ya que, como se ha mencionado antes, es una variable implícita; sin embargo, no todos los algoritmos de imputación utilizan este criterio. Así, los algoritmos de imputación se pueden clasificar en: algoritmos univariantes, algoritmos para series de tiempo univariantes y algoritmos multivariantes basados en retardos. Los algoritmos univariantes son aquellos que no emplean las características de la serie de tiempo para realizar la imputación; por lo general, no suelen tener un buen desempeño (e.g. media). Por el contrario, los algoritmos para series de tiempo univariantes utilizan las características subyacentes en la serie para efectuar una imputación más adecuada (e.g. interpolación lineal). Por último, los algoritmos multivariantes basados en retardos utilizan la información disponible del tiempo, es decir, se apoyan en la premisa de que el tiempo es una variable implícita en una serie temporal; esto se logra empleando retardos (lags) y avances (leads) (Moritz et al., 2015).

1.5.2.2.1 Imputación basada en modelos de machine learning

Esta técnica de imputación para series de tiempo univariantes basada en modelos de machine learning busca aprovechar las correlaciones entre los valores y sus valores tanto pasados como futuros, como la disponibilidad de las observaciones para imputar valores perdidos. Con esta finalidad, para cada gap, se toman dos subseries

univariantes. Una subserie constituida por las observaciones antes del gap y otra subserie que contiene las observaciones después del gap. Estas subseries son tratadas como series de tiempo distintas y son convertidas en series multivariantes. Por un lado, con la subserie antes del gap, se realiza una conversión hacia adelante (Forward converting) para obtener la serie multivariante; por otro lado, con la subserie después del gap, se emplea una conversión hacia atrás (Backward converting). Posteriormente, cada serie multivariante es utilizada para entrenar los modelos de machine learning y pronosticar hacia adelante y hacia atrás, según el tipo de conversión empleado para la obtención de la serie multivariante. Finalmente, estas dos predicciones son promediadas para imputar los valores perdidos. Hay casos especiales en los que se puede obtener solo una serie multivariante para entrenar un modelo; en tales casos, se realiza predicciones hacia adelante o hacia atrás, según corresponda (Phan, 2020).

1.5.3 Machine Learning

Machine learning es un enfoque del statistical learning, su objetivo es estimar o “entrenar” un modelo estadístico que permita realizar predicciones precisas de una variable de respuesta en función de p -variables predictoras. Estas predicciones se ven afectadas por dos tipos de errores: el error irreducible y el error reducible. El error irreducible tiene que ver con el error externo a la modelización, y el error reducible es el error intrínseco en la estimación de la relación entre la variable de respuesta y las variables predictoras, por lo que se busca aumentar la precisión de las predicciones según se reduzca dicho error. La estimación del modelo estadístico puede ser paramétrica o no paramétrica, y se realiza a través de un conjunto de entrenamiento

conformado por las observaciones del conjunto de datos; por otra parte, la validación del modelo se puede llevar a cabo de dos maneras: 1) por medio del mismo conjunto de entrenamiento (e.g. cross-validation), y 2) mediante un conjunto de prueba. La segunda opción es más factible cuando se dispone de un set de datos grande, ya que se divide los datos en dos grupos, el conjunto de entrenamiento y el conjunto de prueba. Por último, según sea la variable de respuesta, cuantitativa o cualitativa, podemos estar ante problemas de regresión o clasificación, respectivamente (James et al., 2013).

1.5.3.1 Árboles de decisión

Es un modelo de machine learning, construido a partir del conjunto de entrenamiento, que se puede aplicar tanto a problemas de regresión como de clasificación. Se busca segmentar el espacio predictor en m regiones, simples, distintas y que no se traslapan, mediante reglas de decisión. Estas reglas de decisión se conocen como nodos internos; en cambio, las regiones resultantes de estas reglas de decisión se las conoce como hojas o nodos terminales. Particularmente, para árboles de regresión, la predicción del valor de la variable de respuesta de cierta observación, que no ha sido considerada en el conjunto de entrenamiento, será la media de los valores de respuesta de las observaciones de entrenamiento de la región a la que pertenecería dicha observación. Así, para determinar las regiones, se empieza con una partición binaria de una sola región conformada por todas las observaciones, es decir, se evalúan todas las variables predictoras y todos sus posibles puntos de corte que resulten en dos regiones particionadas R_1 y R_2 . Se selecciona la variable predictora j y el punto de corte s que minimiza la expresión matemática 1.5.3.1. Este proceso se repite, hasta alcanzar un criterio de continuidad, para cada una de las regiones resultantes; no obstante, ya no se divide a todo el espacio predictor, por el contrario, únicamente

a la región resultante (e.g. R_1 o R_2). Este enfoque utilizado para la obtención de regiones u hojas se conoce como partición binaria recursiva (James et al., 2013).

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \quad (1.5.3.1)$$

Donde,

$$X = \{x_1, x_2, \dots, x_n\},$$

$X_j = j$ – ésima variable predictora,

$s =$ punto de corte,

$$R_1(j, s) = \{X|X_j < s\} \text{ y } R_2(j, s) = \{X|X_j \geq s\},$$

$y_i =$ respuesta de la i – ésima observación,

$\hat{y}_{R_1} =$ respuesta media de la región R_1 ,

$\hat{y}_{R_2} =$ respuesta media de la región R_2

1.5.4 Filtro de Doodson

El filtro de Doodson es un filtro lineal de paso bajo (low-pass) que se considera como un filtro supresor de mareas, es decir, remueve las energías de la marea diurna y de mareas asociadas a altas frecuencias de las mediciones del nivel del mar en el cálculo del nivel medio del mar. Para este fin, primero se determinan los días promedio, a través del filtro, donde se requiere de 39 horas de datos centrados a las 12:00 horas de cada día. Posteriormente, los meses promedio son calculados como la media aritmética de los niveles del mar de los días promedio que constituyen cada mes; a este resultado final se lo conoce como el nivel medio del mar mensual. Sin embargo, el filtro de Doodson es distinto al filtro de la UHSLC por lo cual, la Permanent Service for Mean Sea Level (PSMSL) los comparó, junto con la media aritmética, en la determinación del nivel medio del mar por medio de varias pruebas en el 2000. Tomaron los datos horarios de dos años de alrededor de 800 estaciones que contaban con los registros del nivel del mar

producidos por la UHSLC y aplicaron el método station-year para realizar las comparaciones. Concluyeron que las desviaciones estándar entre los dos filtros en cuestión fueron de 4.7 *mm* y 1.3 *mm* para los promedios diarios y los promedios mensuales, respectivamente. En definitiva, para futuras investigaciones, recomiendan emplear cualquiera de los dos filtros por encima de la media aritmética, ya que ambos disponen de una base matemática rigurosa y la diferencia de resultados entre los filtros es apenas milimétrica (UNESCO/IOC, 2002).

1.5.5 Análisis Espectral Singular (SSA)

Como lo mencionan Golyandina et al. (2018), SSA es un método no paramétrico para el análisis espectral de una serie de tiempo, donde la serie se descompone en componente de tendencia, componentes periódicos y ruido, y se basa en dos etapas. La primera etapa, descomposición, está conformada por los siguientes pasos: 1) construcción de la matriz Hankel \mathbf{X} de tamaño $(L \times K)$ a partir de las observaciones de la serie de tiempo de tamaño N , donde L se conoce como el parámetro de ventana y K será, tal que, $K = N - L + 1$, y 2) descomposición en valores singulares de $\mathbf{S} = \mathbf{X}\mathbf{X}^T$ en matrices de rango 1. Mientras que, la segunda etapa, reconstrucción, está conformada por: 3) agrupamiento de las matrices de rango 1 según sus valores propios asociados a sus eigentriples, y 4) reconstrucción de los componentes de la serie de tiempo original a través de las matrices agrupadas.

1.5.5.1 Supuesto de separabilidad y tamaño del parámetro

SSA se basa en el supuesto de que una serie de observaciones X de tamaño N puede ser utilizada para extraer sus componentes X_1 y X_2 , donde $X = X_1 + X_2$, y X_1, X_2 son series separables; es decir, que existe un agrupamiento en el método, tal que, $\overline{X}_1 = X_1$ y $\overline{X}_2 = X_2$. Asintóticamente se puede alcanzar la

separabilidad, $\bar{X}_1 \approx X_1$ y $\bar{X}_2 \approx X_2$, si el parámetro L y/o K son lo suficientemente grandes; esto se logra con $L \sim \frac{N}{2}$. Además, dicha separabilidad puede ser mejorada si nuestro objetivo es estimar un componente periódico donde conocemos *a priori* su periodicidad, entonces L deberá ser también divisible para el período (Golyandina et al., 2018).

1.5.6 Frecuencia dominante

Las series de tiempo pueden mostrar comportamientos periódicos asociados a las frecuencias de sus componentes oscilatorios, es decir, se evidencia comportamientos sistemáticos, que se repiten. A su vez, estas frecuencias están asociadas a energías o potencias. Así, la frecuencia que almacena la mayor cantidad de energía de entre todas las frecuencias se la conoce como frecuencia dominante. En cambio, la menor frecuencia de entre todas las frecuencias es llamada frecuencia fundamental, y está en “armonía” con otras frecuencias conocidas como armónicos; esta frecuencia fundamental también puede ser una frecuencia dominante. Cabe resaltar que, no todas las frecuencias son armónicos, por lo que se pueden encontrar distintas frecuencias a las descritas. Por consiguiente, al identificar la frecuencia dominante de una serie de tiempo adquirimos información importante sobre su estructura y esto nos permite estudiar los potenciales efectos que dicha frecuencia dominante puede generar sobre la serie temporal (Telgársky, 2013).

1.6 Estado del arte

1.6.1 Imputación para series de tiempo univariantes

Bokde et al. (2018) presentaron un método de imputación para series de tiempo univariantes basado en Pattern Sequence Forecasting (PSF). El algoritmo de PSF identifica patrones en la serie según sus características periódicas y realiza predicciones

hacia adelante; este nuevo método modifica dicho algoritmo para realizar predicciones hacia adelante y hacia atrás, y así imputar los datos con el promedio de estas predicciones. La precisión de imputación se evaluó a través de simulaciones de distintos porcentajes de valores perdidos (10%, 20%, 30%, 40% y 50%) y tamaños de gaps (entre 25% y 100%). Se comparó con los siguientes métodos de imputación: media, interpolación lineal, last observation carried forward (LOCF), ARIMA con filtro de Kalman y bosques aleatorios; la métrica de validación fue: Root Mean Square Error (RMSE). Concluyeron que el método propuesto, en general, tenía un mejor desempeño respecto a los demás. Asimismo, recomiendan este método para series con componentes periódicos y que la causa para que la observación se pierda sea del tipo MAR, caso contrario, el método puede ser superado por otros. Además, el costo computacional para llevar a cabo este método de imputación es alto, sobre todo cuando se trabaja con sets de datos extensos. Cabe mencionar que el método de imputación por bosques aleatorios que se menciona en este estudio se encuentra en el paquete “imputeMissings” del software estadístico R, y se basa en imputar el conjunto de entrenamiento con la mediana o moda, para luego entrenar el modelo e imputar los datos con las predicciones de dicho modelo. Este método de imputación por bosques aleatorios es totalmente distinto al propuesto por Phan que se detalla más adelante y es empleado en este proyecto.

Moritz et al. (2015) compararon varios métodos de imputación para series de tiempo univariantes, asumiendo que la causa para que la observación se pierda era del tipo MCAR, a través de simulaciones de valores perdidos en el software estadístico R. Los métodos comparados fueron: media, LOCF, interpolación con filtro estacional de Kalman, interpolación lineal basada en la descomposición de la parte estacional de la serie, interpolación lineal sin descomponer la serie, y un modelo iterativo basado en crear un set de datos multivariante con los retardos de la serie. Emplearon las siguientes métricas

para validar la imputación: RMSE y Mean Absolute Percentage Error (MAPE). Obtuvieron que los métodos de imputación más efectivos fueron: interpolación lineal con filtro estacional de Kalman e interpolación lineal basada en la extracción de la estacionalidad.

Phan et al. (2017) propusieron un método de imputación basado en Dynamic Time Warping para series de tiempo univariantes (DTWBI) que presentaban gaps grandes. El método identifica patrones en la serie de tiempo para imputar y asume que la causa para que la observación se pierda era del tipo MAR, y lo compararon, mediante simulaciones empleando varias series de tiempo, con los siguientes métodos de imputación: interpolación lineal basada en la descomposición de la parte estacional de la serie, interpolación lineal sin descomponer la serie, LOCF, media, e interpolación polinomial (spline). Para este estudio utilizaron las siguientes métricas de validación: Similaridad, Normalized Mean Absolute Error (NMAE), Fraction of Standard Deviation (FSD) y RMSE. Concluyeron que, con el método propuesto, DTWBI, obtuvieron los mejores resultados en comparación con los demás. Sin embargo, este método está basado en el supuesto de datos recurrentes en una serie de tiempo.

Phan et al. (2020) mejoraron el método de imputación previo que propusieron, DTWBI, y lo llamaron eDTWBI. Se basa en el mismo supuesto de recurrencia en los datos, gaps grandes, causa para que la observación se pierda del tipo MAR e identificación de patrones en la serie. eDTWBI fue comparado, a través de simulaciones, con los siguientes métodos de imputación: Enfoque heurístico (método aleatorizado), interpolación basada en el filtro de Kalman, interpolación lineal basada en la descomposición de la parte estacional de la serie, media, spline, LOCF y DTWBI. Las métricas utilizadas para la validación del desempeño de los métodos fueron: Similaridad, NMAE, RMSE y Fractional Bias (FB). Los resultados demostraron una mejora considerable respecto al método DTWBI y superó ampliamente a los demás métodos.

Phan (2020) propuso un método de imputación para series de tiempo univariantes basado en machine learning, en particular, empleó modelos por bosques aleatorios y support vector machine (SVM) en su versión de regresión. El método utiliza una subserie antes del gap y otra después del gap, para posteriormente convertirlas en series multivariantes y a partir de ellas entrenar los modelos. Luego, con los modelos antes y después del gap se realizan predicciones hacia adelante y hacia atrás, respectivamente. Por último, las predicciones resultantes de ambos modelos son promediados y conforman el vector de imputaciones finales. Este proceso fue realizado tanto para bosques aleatorios como para SVM y fueron comparados, mediante simulaciones, con los siguientes métodos: LOCF, filtro de Kalman, interpolación lineal, y eDTWBI. Por otro lado, las métricas de validación utilizadas fueron: Similaridad, Mean Absolute Error (MAE), RMSE, FB y FSD. Concluyó que los métodos de imputación basados en bosques aleatorios y SVM proveían una mejora considerable en el desempeño respecto a los demás métodos en todas las métricas y con todos los tamaños de gaps simulados. Además, la forma que tomaban los valores imputados por este método propuesto fue aproximadamente idéntico a la forma que tomaban los valores reales.

1.6.2 Análisis espectral del nivel del mar

Bayot & Cornejo Rodríguez (1996) evidenciaron ondas ecuatoriales en Salinas y Galápagos a través de series de tiempo de promedios diarios de Temperatura Superficial del Mar, Nivel Medio del Mar, Presión Atmosférica a Nivel del Mar y Vientos Zonales y Meridionales provenientes de las estaciones de La Libertad y Salinas (arreglo SAL), y de Santa Cruz, San Cristóbal y Baltra (arreglo GAL). El período de datos comprende entre 1985 y 1988. Realizaron un análisis de Funciones Empíricas Ortogonales (FEO) para identificar los modos (componentes principales) que explicaban la mayor variabilidad de las series. Consecuentemente, realizaron análisis espectral tanto de las series de tiempo

originales, como de los modos; encontrando así, periodicidad anual en las series, y presencia de oscilaciones con períodos menores al anual que se asociaban con las ondas ecuatoriales Kelvin y Rossby-gravedad.

Beşel & Tanır Kayıkçı (2020) investigaron la variabilidad media del nivel del mar del Mar Negro a través del SSA. Analizaron 10 estaciones mareográficas a lo largo del mar negro, cada una con períodos de datos distintos, de donde concluyeron que en ciertas estaciones se presenciaba un incremento en las tendencias, y determinaron la presencia de un componente estacional dominante en las series de tiempo estudiadas. Sin embargo, no se asoció a este componente periódico con ninguna posible onda conocida.

Ivanov et al. (2019) realizaron un Análisis Espectral Singular (SSA) del nivel medio del mar mensual del Mar Negro; utilizaron los datos de la estación Varna en el período de 1929-2019. A través del SSA descompusieron la serie de tiempo en tendencia, componentes oscilatorios y ruido, y lo complementaron con un Análisis de Fourier para estimar fases y amplitudes de los componentes oscilatorios más significativos. El SSA mostró periodicidad anual, semi-anual y decadal del nivel medio del mar. Además, estimaron la tendencia mediante regresión lineal, concluyendo que la tasa de cambio se estima en 1.2 mm por año.

Khelifa et al. (2016) evaluaron al Análisis Espectral Singular (SSA) y al Análisis Multiresolución Wavelet (WMA). Trabajaron con la anomalía del nivel global del mar semanal, medida mediante la altimetría satelital entre 1993 y 2013. El estudio estuvo enfocado en la estimación de tendencias no lineales y componentes estacionales. Los resultados mostraron que la serie de tiempo estaba dominada por una tendencia no lineal y componentes anuales y semi-anuales. Además, que el SSA identificó de manera correcta a eventos El Niño y La Niña (ENOS) asociados a la variabilidad interanual.

Ozsoy et al. (2016) estudiaron las variaciones del nivel del mar asociadas a frecuencias altas y su implicación en las inundaciones costeras en el estrecho de Solent, Reino Unido. Utilizaron observaciones del nivel del mar de la estación de Southampton y también de 23 estaciones más ubicadas alrededor de Solent para evaluar las características espaciales de los fenómenos en el período de 2000-2013. Descompusieron la serie de tiempo mediante el análisis armónico para consecuentemente identificar 8 eventos con mayor energía a través del análisis espectral y análisis wavelet, concluyendo así que estos fenómenos tienen un período dominante de alrededor de 4 horas. El evento con la mayor energía fue asociado con inundaciones costeras menores en Yarmouth, isla de Wight.

Wong (2011), en su proyecto de tesis de grado, estudió la variabilidad del nivel medio del mar en Ecuador. Los datos eran horarios y provenían de 4 estaciones mareográficas (entre ellas La Libertad) en el período de diciembre de 1992 a septiembre de 2008. Realizó un análisis armónico de mareas (determinar amplitud y fase de los armónicos) junto con un análisis espectral no paramétrico (método de Welch). Para el análisis espectral trabajó con datos mensuales de anomalías y datos diarios que fueron determinados mediante el promedio de las observaciones horarias. Identificó la frecuencia significativa asociada a la periodicidad de 16 meses para 3 estaciones (incluido a La Libertad) en las anomalías mensuales, y se atribuyó dicha periodicidad al ciclo anual solar. Con los datos diarios se concluyó que la banda de frecuencia significativa para las series analizadas comprendía entre 0.00146-0.00158 ciclos por día, es decir, entre períodos de 21 y 22 meses. Además, encontró un ciclo semestral (186 días) en común en todas las estaciones, tanto en el análisis espectral como armónico. Encontró también bandas de frecuencias que corresponden a períodos entre 77.3 y 91 días que podría relacionarse con ondas kelvin 20-90días y una frecuencia

correspondiente al período de 7.4 días que se podría asociar con una onda Rossby-gravedad o una onda kelvin 8-10 días. Por otra parte, determinó la tendencia a través de la regresión lineal utilizando los datos diarios completos, donde encontró evidencia de un decremento en el nivel del mar en la estación de La Libertad (-0.586 mm por año), y en los datos diarios excluyendo El Niño (-0.772 mm por año). Al comparar dichas tendencias evidenció influencia del fenómeno en la variabilidad del nivel del mar.

CAPÍTULO 2

2. METODOLOGÍA

En el presente capítulo se describen las técnicas y modelos estadísticos empleados en el proceso de imputación, aplicación del filtro de Doodson y análisis espectral para la serie de tiempo nivel del mar de la estación fija de La Libertad-Ecuador.

2.1 Bosques Aleatorios (Random Forests)

Los modelos por bosques aleatorios fueron utilizados para realizar predicciones hacia adelante y/o hacia atrás en el proceso de imputación de datos para la serie de tiempo nivel del mar horario de La Libertad que se detalla en la sección 2.2.

Estos modelos de machine learning se basan en Bootstrap y árboles de decisión. Mediante Bootstrap se generan k muestras aleatorias a partir del conjunto de entrenamiento original para, posteriormente, estimar k árboles de decisión por medio de dichas muestras. Retomando lo enunciado en el apartado 1.5.3.1, las regiones de un árbol de decisión se construyen a través de la partición binaria recursiva. Sin embargo, para bosques aleatorios, en cada partición binaria se toma en cuenta únicamente una muestra aleatoria de variables predictoras; esta muestra es generada cada vez que se realiza una partición. Así, se evalúan los predictores pertenecientes a dicha muestra y todos sus posibles puntos de corte para seleccionar el predictor j y el punto de corte s que minimiza la expresión (1.5.3.1). Este criterio es empleado en la construcción de los k árboles y permite reducir las posibles correlaciones entre los mismos. Consecuentemente, cada árbol de decisión estimado proporciona una predicción; todas estas k predicciones resultantes son promediadas para obtener una predicción final con una varianza considerablemente reducida (James et al., 2013).

2.2 Imputación para series de tiempo univariantes basada en bosques aleatorios

Se realizó la imputación para la serie de tiempo nivel del mar horario de La Libertad según lo descrito por Phan (2020), donde si $X = \{x_1, x_2, \dots, x_N\}$ es la serie de tiempo de interés de tamaño N , T el tamaño del gap y t el índice del inicio del gap, entonces:

- Si el gap se encuentra dentro de las primeras $3xT$ observaciones de la serie, entonces se utiliza las observaciones después del gap para realizar la imputación. Por el contrario, si el gap se ubica en las últimas $3xT$ observaciones de la serie, entonces se emplea las observaciones antes del gap para la imputación. Mientras que, si el gap se localiza entre las primeras $3xT$ y las últimas $3xT$ observaciones de la serie, entonces se utiliza tanto los datos antes del gap como después del gap para la imputación final.
- Se toman dos subseries de la serie de tiempo original (o una subserie según lo descrito en el apartado a). Por un lado, la subserie antes del gap denotada por, $Da = X[1:t - 1]$; y, por otro lado, la subserie después del gap denotada por, $Dd = \text{invertido}(X[t + T:N])$.
- A partir de las subseries identificadas se obtienen las series multivariantes:

$$MDa = \begin{bmatrix} x_1 & x_2 & \cdots & x_T & x_{T+1} \\ x_2 & x_3 & \cdots & x_{T+1} & x_{T+2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{t-T-1} & x_{t-T} & \cdots & x_{t-2} & x_{t-1} \end{bmatrix}$$

$$MDd = \begin{bmatrix} x_N & x_{N-1} & \cdots & x_{N-T+1} & x_{N-T} \\ x_{N-1} & x_{N-2} & \cdots & x_{N-T} & x_{N-T-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{t+2T} & x_{t+2T-1} & \cdots & x_{t+T+1} & x_{t+T} \end{bmatrix}$$

Donde, MDa y MDd son matrices con $T+1$ columnas, y el número de filas depende de la cantidad de observaciones disponibles antes del gap o después del gap, según corresponda.

- d) Los conjuntos de datos multivariantes del apartado anterior se emplean para entrenar los modelos por bosques aleatorios. La respuesta será la columna $T+1$ y los T -predictores el resto de columnas, tal que,

$$\hat{f}_a = f(MDa)$$

$$\hat{f}_d = f(MDd)$$

De esta manera, \hat{f}_a y \hat{f}_d son los modelos estimados por bosques aleatorios a partir de la subserie antes del gap y después del gap, respectivamente.

- e) Una vez estimados los modelos se procede a realizar las T predicciones. Para los datos antes del gap, tenemos:

$$\begin{aligned} x_t &= \hat{f}_a(x_{t-T}, x_{t-T+1}, \dots, x_{t-1}) \\ x_{t+1} &= \hat{f}_a(x_{t-T+1}, x_{t-T+2}, \dots, x_t) \\ &\vdots \\ x_{t+T-1} &= \hat{f}_a(x_{t-1}, x_t, \dots, x_{t+T-2}) \end{aligned}$$

Así, el vector de imputaciones por parte del modelo antes del gap es:

$$\widehat{x}_a = (x_t, x_{t+1}, \dots, x_{t+T-1})$$

Para los datos después del gap, tenemos:

$$\begin{aligned} x_{t+T-1} &= \hat{f}_d(x_{t+2T-1}, x_{t+2T-2}, \dots, x_{t+T}) \\ x_{t+T-2} &= \hat{f}_d(x_{t+2T-2}, x_{t+2T-3}, \dots, x_{t+T-1}) \\ &\vdots \\ x_t &= \hat{f}_d(x_{t+T}, x_{t+T-1}, \dots, x_{t+1}) \end{aligned}$$

Por lo que, el vector de imputaciones por parte del modelo después del gap es:

$$\widehat{x}_d = (x_t, x_{t+1}, \dots, x_{t+T-1})$$

- f) Finalmente, se promedian los vectores de imputaciones por parte de los modelos antes y después del gap, \widehat{x}_a y \widehat{x}_d , respectivamente. El vector promedio resultante contiene los valores imputados finales.

2.3 Filtro de Doodson

Para determinar la serie de tiempo nivel medio del mar diario se aplicó el filtro de Doodson en la serie nivel del mar horario, de tal manera que se utilizaron 39 horas de datos centrados a las 12:00 horas para cada día calculado como lo detalló UNESCO/IOC (1985), donde si $N(t)$ es el nivel del mar para $-19 \leq t \leq 19$, siendo t los índices para las 39 horas, y c es el índice correspondiente a las 12:00 horas del día de interés d , entonces:

$$F(t) = \begin{cases} 0, & \text{si } t \in \{-18, -16, -15, -13, -10, -8, -5, 0, 5, 8, 10, 13, 15, 16, 18\} \\ 1, & \text{si } t \in \{-19, -17, -14, -12, -11, -7, -6, -3, -2, 2, 3, 6, 7, 11, 12, 14, 17, 19\} \\ 2, & \text{si } t \in \{-9, -4, -1, 1, 4, 9\} \end{cases}$$

$$NMM_d(c) = \frac{1}{30} \sum_{j=-19}^{19} F(j)N(c+j)$$

$F(t)$ es el filtro de Doodson y $NMM_d(c)$ es el nivel medio del mar del día d centrado en c .

Posteriormente, sea x_{ijk} una observación de la serie nivel medio del mar diario para el año i , mes j y día k , donde n_j es el total de observaciones en el mes j , para $j = 1, \dots, 12$, entonces $w_{ij} = \frac{1}{n_j} \sum_{k=1}^{n_j} x_{ijk}$. Los w_{ij} resultantes conformaron la serie de tiempo nivel medio del mar mensual.

Por último, sea w_{ij} una observación de la serie nivel medio del mar mensual para el año i y mes j , y N el total de años de observaciones en la serie, entonces $\bar{w}_j = \frac{1}{N} \sum_{i=1}^N w_{ij}$, para $j = 1, \dots, 12$. Los \bar{w}_j resultantes son los promedios de los niveles del mar de cada mes j a lo largo de los N años. Así, la serie de tiempo anomalías del nivel medio del mar mensual se obtuvo como $y_{ij} = w_{ij} - \bar{w}_j$.

2.4 Métricas de validación

Para validar las imputaciones efectuadas por bosques aleatorios se empleó métricas de validación estadísticas como lo señaló Phan (2020), donde si $X = \{x_1, x_2, \dots, x_T\}$ son los

valores perdidos reales, $Y = \{y_1, y_2, \dots, y_T\}$ los valores imputados propuestos, y T el tamaño del gap, entonces:

- a) Root Mean Square Error (RMSE), mide la precisión de la imputación y entre menor sea el RMSE mejor será el modelo de imputación.

$$RMSE(X, Y) = \sqrt{\frac{1}{T} \sum_{i=1}^T (y_i - x_i)^2}$$

- b) Mean Absolute Error (MAE), cuantifica el desempeño del modelo de imputación y para valores bajos muestra altos desempeños del modelo.

$$MAE(X, Y) = \frac{1}{T} \sum_{i=1}^T |y_i - x_i|$$

- c) Similaridad, indica la similitud entre los valores imputados y los valores reales. Toma valores entre 0 y 1, tal que los valores cercanos a 1 muestran mayor similitud y los valores cercanos a 0 menor similitud.

$$Similaridad(X, Y) = \frac{1}{T} \sum_{i=1}^T \frac{1}{1 + \frac{|y_i - x_i|}{\max(X) - \min(X)}}$$

- d) Fractional Bias (FB), evalúa al modelo de imputación; si la métrica tiende a cero, entonces se considera al modelo de imputación como perfecto.

$$FB(X, Y) = 2 \left| \frac{media(Y) - media(X)}{media(Y) + media(X)} \right|$$

- e) Fraction of Standard Deviation (FSD), mide qué tan cerca están los valores imputados de los valores reales; si la métrica tiende a cero, entonces los valores imputados se acercan a los valores reales. Esta métrica se suele emplear, junto con FB, para evaluar la forma que toman los valores imputados respecto a los valores reales.

$$FSD(X, Y) = 2 \left| \frac{sd(Y) - sd(X)}{sd(Y) + sd(X)} \right|$$

Estas métricas también fueron aplicadas para comparar la serie diaria resultante del filtro de Doodson y la serie diaria de la UHSLC. Para este caso particular, $\mathbf{X} = \{x_1, x_2, \dots, x_T\}$ fue la serie nivel medio del mar diario de la UHSLC, $\mathbf{Y} = \{y_1, y_2, \dots, y_T\}$ la serie nivel medio del mar diario determinado por el filtro de Doodson, y T el tamaño de las series en cuestión.

2.5 Análisis Espectral Singular

Se realizó el análisis espectral para la serie anomalías del nivel medio del mar mensual de La Libertad a través del método conocido por análisis espectral singular (SSA), como Golyandina et al. (2018) lo detallaron en su libro; donde si $\mathbb{X} = \{x_1, x_2, \dots, x_N\}$ es una serie de tiempo de tamaño N , $1 < L < N$, $K = N - L + 1$, entonces:

- 1) Sea $\mathfrak{S}: \mathbb{R}^N \rightarrow \mathcal{M}_{L,K}^{(H)}$, se construye la matriz Hankel $\mathbf{X} = [X_1: X_2: \dots: X_K] = \mathfrak{S}(\mathbb{X})$, tal que, $X_i = (x_i, x_{i+1}, \dots, x_{i+L-1})^T$ para $i = 1, \dots, K$. Así,

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 & \cdots & x_K \\ x_2 & x_3 & \cdots & x_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & \cdots & x_N \end{bmatrix}$$

- 2) Sea $\mathbf{S} = \mathbf{X}\mathbf{X}^T$, donde $\lambda_i \geq \dots \geq \lambda_d > 0$ son sus valores propios asociados, y $d = \min\{L, K\}$ el rango de \mathbf{X} , entonces la descomposición en valores singulares (SVD) de \mathbf{X} es:

$$\mathbf{X} = \sum_{i=1}^d \sqrt{\lambda_i} U_i V_i^T = \mathbf{X}_1 + \dots + \mathbf{X}_d$$

Tal que, U_i es un sistema ortonormal que contiene los vectores propios de \mathbf{S} , $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}$, $\sqrt{\lambda_i}$ son los valores propios de la matriz Hankel \mathbf{X} , y las matrices \mathbf{X}_i son de rango 1. A la terna $(\sqrt{\lambda_i}, U_i, V_i)$ se la conoce como eigentriple de la matriz \mathbf{X}_i .

- 3) Sean I_1, \dots, I_m subconjuntos disjuntos del conjunto de índices $\{1, \dots, d\}$, entonces $\mathbf{X}_{I_1}, \mathbf{X}_{I_2}, \dots, \mathbf{X}_{I_m}$ son matrices resultantes del agrupamiento según los

subconjuntos I_1, \dots, I_m . En particular, si $I_1 = \{i_1, \dots, i_p\}$, entonces $\mathbf{X}_{I_1} = \mathbf{X}_{i_1} + \dots + \mathbf{X}_{i_p}$. Así, de manera general se tiene que,

$$\mathbf{X} = \mathbf{X}_{I_1} + \dots + \mathbf{X}_{I_m}$$

Este proceso es llamado agrupamiento por eigentriples, ya que se evalúan sus valores y vectores propios asociados mediante gráficos para realizar dichos agrupamientos que, consecuentemente, conllevan a identificar los componentes de la serie en análisis.

- 4) Sea $\Pi_{\mathcal{H}}: R^{L \times K} \rightarrow \mathcal{M}_{L,K}^{(H)}$ la proyección ortogonal de una matriz que pertenece al espacio $R^{L \times K}$ hacia $\mathcal{M}_{L,K}^{(H)}$; y, para $1 \leq k \leq m$, $\widetilde{\mathbf{X}}_k = \mathbf{X}_{I_k}$, donde $\widetilde{\mathbf{X}}_k \in R^{L \times K}$, entonces $\widetilde{\mathbf{X}}_k = \mathfrak{S}^{-1} \circ \Pi_{\mathcal{H}}(\widetilde{\mathbf{X}}_k)$. Los $\widetilde{\mathbf{X}}_k \in \mathbb{R}^N$ resultantes son los componentes reconstruidos de la serie de tiempo en análisis, y se cumple que,

$$\mathbb{X} = \widetilde{\mathbf{X}}_1 + \dots + \widetilde{\mathbf{X}}_m$$

Los pasos 1 y 2 conforman la etapa de descomposición de la serie, y los pasos 3 y 4 la etapa de reconstrucción de la serie.

Por otro lado, SSA permite realizar un prefiltrado para la serie \mathbb{X} que consiste en reducir el ruido presente basado en la descomposición de \mathbb{X} que se describió en el método; esto fue empleado para facilitar la identificación de componentes periódicos en el periodograma.

2.6 Regresión lineal simple

Se empleó el modelo de regresión lineal simple para estimar tendencias lineales para la serie de tiempo anomalías del nivel medio del mar mensual, con el objetivo de determinar las tasas de cambio (pendiente de la recta de regresión) en el nivel del mar.

Según lo señalado por Kutner et al. (2004), el modelo regresión lineal simple se expresa matemáticamente como:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Donde,

$Y_i =$ respuesta para la i – ésima observación

$X_i =$ predictor para la i – ésima observación

$\varepsilon_i =$ término del error para la i – ésima observación

$\beta_0 =$ parámetro del intercepto de la recta de regresión

$\beta_1 =$ parámetro de la pendiente de la recta de regresión

2.7 Periodograma

Se estimó la densidad espectral mediante el periodograma suavizado para la serie prefiltrada a partir del SSA, y a través del mismo se identificaron los componentes periódicos, sus frecuencias, espectros de potencia, y la frecuencia dominante. Posteriormente, se procedió a determinar la periodicidad de dichos componentes periódicos.

El periodograma de una serie de tiempo es una estimación de la densidad espectral teórica basada en el cuadrado de la transformada de Fourier Discreta de la serie en análisis, y muestra la distribución de energías o espectros de potencia, según sus frecuencias. Esta distribución de los espectros de potencia es también llamada densidad espectral de potencia y permite encontrar la frecuencia dominante como el máximo local (pico predominante) presente en el periodograma (Telgársky, 2013).

Tal como Cryer & Chan (2008) lo expusieron, sea $S(f)$ la densidad espectral teórica e $I(f)$ el periodograma para todas las frecuencias $-1/2 < f \leq 1/2$, entonces la densidad espectral muestral $\hat{S}(f)$ no es un estimador estadísticamente consistente de $S(f)$.

Donde,

$$\hat{S}(f) = \begin{cases} \frac{1}{2}I(f), & \text{para } -1/2 < f < 1/2 \\ I\left(\frac{1}{2}\right), & \text{para } f = 1/2 \end{cases}$$

La inconsistencia del estimador viene dada por la no reducción de la varianza al incrementar el tamaño de muestra n , es decir, la varianza no depende del tamaño de muestra. La solución a este problema se presenta en la suavización de la densidad espectral muestral a través del intercambio entre sesgo y varianza, y esto se logra mediante la ventana espectral Daniell $W_m(f)$, con m como la cantidad de frecuencias Fourier a extender a los lados de f y frecuencia Fourier $f = \frac{j}{n}$, para $j = 1, \dots, k$. $W_m(f)$ permite promediar los valores de la densidad espectral muestral sobre pequeños intervalos de frecuencias centrados en f y reducir la varianza a cambio de aumentar ligeramente el sesgo, tal que, $W_m(k) = \frac{1}{2^{m+1}}$ para $-m \leq k \leq m$. Además, cumple con las siguientes propiedades: $W_m(k) \geq 0$, $W_m(k) = W_m(-k)$, y $\sum_{k=-m}^m W_m(k) = 1$.

De ahí que, la densidad espectral muestral suavizada $\bar{S}(f) = \sum_{k=-m}^m W_m(k) \hat{S}(f + \frac{k}{n})$ es un estimador consistente de $S(f)$. Por otro lado, la ventana espectral Daniell modificada presenta pesos reducidos a la mitad para los valores extremos, pero manteniendo la propiedad $\sum_{k=-m}^m W_m(k) = 1$, y permite suavizar los cambios severos en los extremos. Muy comúnmente, se realiza convoluciones de una misma ventana Daniell modificada para suavizar más de una vez al periodograma y obtener mejores estimaciones. Dado lo mencionado, se prefirió emplear la ventana Daniell modificada para estimar el periodograma suavizado. Adicionalmente, se utilizó el ancho de banda BW como medida del tamaño de la ventana Daniell modificada que mide el ancho de banda de frecuencias promediadas en la suavización del periodograma. Esta medida se tomó en consideración dado el riesgo de tomar una ventana relativamente ancha que suavice dos o más picos juntos, al encontrarse muy cercanos, y podrían formar parte de la densidad espectral teórica, por lo que estos picos no se visualizarían en un periodograma suavizado.

Si bien la definición de la densidad espectral muestral (periodograma) se desarrolla en frecuencias $-1/2 < f \leq 1/2$, se centra el estudio en las frecuencias $0 \leq f \leq 1/2$, ya que la

función $\hat{S}(f)$ es simétrica alrededor de cero y, por tanto, las frecuencias negativas no aportan con nueva información.

Por último, se empleó el método tapering, para solventar el fenómeno denominado leakage, que se refiere a la fuga de poder de frecuencias no Fourier hacia frecuencias Fourier cercanas, causando así que se presente picos mayores en frecuencias que no tienen tanto poder. Usualmente se aplica el taper h_t denominado como campana cosenoidal partida (Split cosine bell), tal que,

$$h_t = \begin{cases} \frac{1}{2} \left\{ 1 - \cos \left[\frac{2\pi(t - 0.5)}{n} \right] \right\}, & \text{para } 1 \leq t \leq m \\ 1, & \text{para } m + 1 \leq t \leq n - m \\ \frac{1}{2} \left\{ 1 - \cos \left[\frac{\pi(n - t + 0.5)}{m} \right] \right\}, & \text{para } n - m + 1 \leq t \leq n \end{cases}$$

Donde el taper cosenoidal es empleado únicamente en los extremos de la serie temporal. Así, la serie de tiempo aplicada el taper se representa como $\check{y}_t = h_t y_t$.

Las periodicidades de los componentes periódicos más importantes fueron determinados como: $T = \frac{1}{f}$, donde T es el período y f la frecuencia del componente oscilatorio.

CAPÍTULO 3

3. RESULTADOS Y ANÁLISIS

Para el presente capítulo se exponen los resultados del análisis exploratorio, imputación por bosques aleatorios, aplicación del filtro de Doodson, y análisis espectral singular efectuados para la serie de tiempo nivel del mar de la estación fija de La Libertad-Ecuador.

3.1 Análisis exploratorio

En esta sección se presentan los estadísticos descriptivos, diagrama de caja, histograma y distribución de los valores perdidos para la serie de tiempo nivel del mar horario de La Libertad. El nivel del mar se encuentra medido en *mm*, y el período de datos de estudio fue desde el 1949-09-01 a las 05:00:00 horas hasta el 2021-01-31 a las 23:00:00 horas, con un total de 626059 observaciones.

En la Figura 3.1.1 se visualiza a la serie de tiempo en análisis, la cual presentó 19047 valores perdidos (3.04%), y, por tanto, 607012 datos válidos (96.96%). Para dicha serie, se elaboró un diagrama de caja para evaluar potenciales valores atípicos, esto se presenta en la Figura 3.1.2, donde se observó que no hay presencia de posibles valores aberrantes; además, de manera gráfica, la media o nivel de la serie (círculo azul) fue aproximadamente igual a la mediana (línea que cruza por la caja), lo que dio un indicio de una distribución simétrica del nivel del mar horario. Dicha simetría en la distribución se confirmó en el histograma que se muestra en la Figura 3.1.3, donde también se evidenció una distribución bimodal.

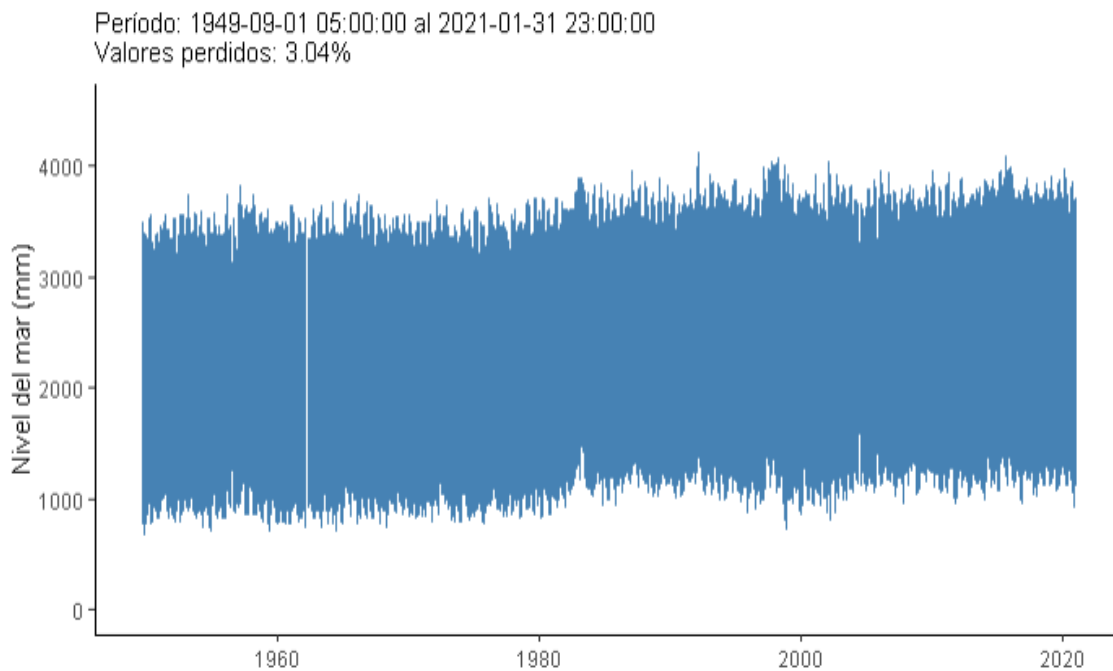


Figura 3.1.1 Serie de tiempo nivel del mar horario de La Libertad

Fuente: Elaboración propia

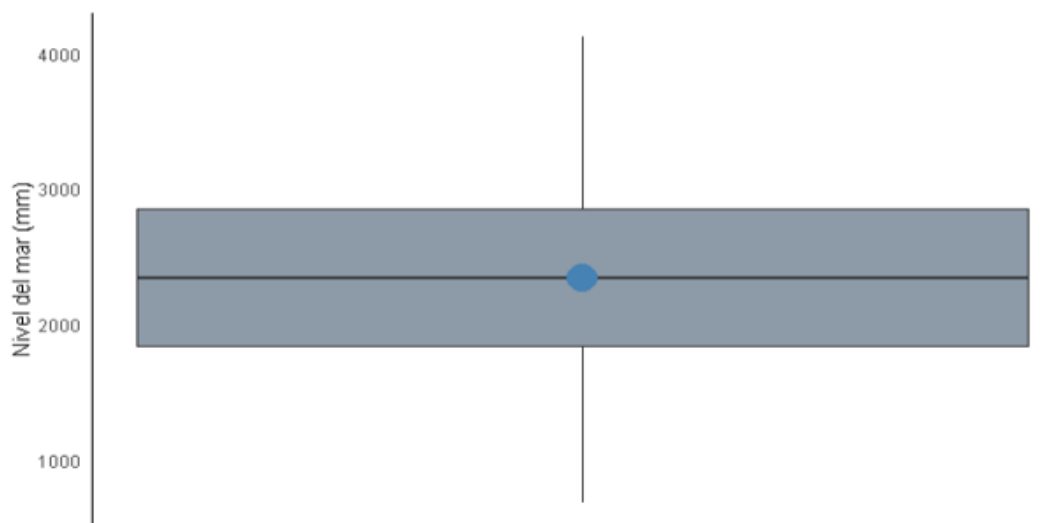


Figura 3.1.2 Diagrama de caja del nivel del mar horario

Fuente: Elaboración propia

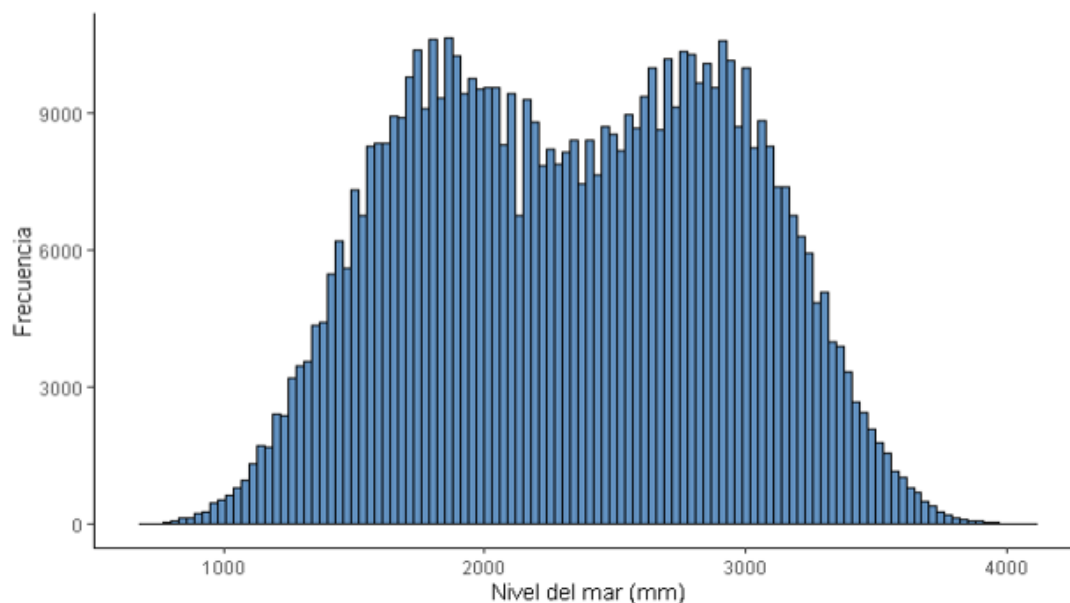


Figura 3.1.3 Histograma del nivel del mar horario

Fuente: Elaboración propia

Por otro lado, en la Tabla 3.1.1 se exponen los estadísticos descriptivos, y, se apreció, cuantitativamente, que la media y la mediana son aproximadamente idénticas, y debido a esto su distribución fue simétrica. Asimismo, contrastando junto con la Figura 3.1.2, el 50% de las observaciones se encontraron entre 1840 *mm* (1° cuartil) y 2847 *mm* (3° cuartil), y su desviación estándar indicó una variabilidad considerable en el nivel del mar horario respecto a su media.

Tabla 3.1.1 Estadísticos descriptivos del nivel del mar horario

Fuente: Elaboración propia

Mínimo	683 <i>mm</i>
Máximo	4120 <i>mm</i>
1° cuartil	1840 <i>mm</i>
2° cuartil (mediana)	2340 <i>mm</i>
3° cuartil	2847 <i>mm</i>
Media	2339 <i>mm</i>
Desviación estándar	612.1791 <i>mm</i>

Varianza	374 763.3 mm ²
----------	---------------------------

En lo que respecta a los valores perdidos, como se puede observar en la Figura 3.1.4, las décadas de los 50, 60, 70, y 2000 presentaron una mayor cantidad de valores perdidos. En cambio, los últimos meses de 1949, la década de los 80, el 2020, y el primer mes del 2021 no presentaron ausencia de datos.

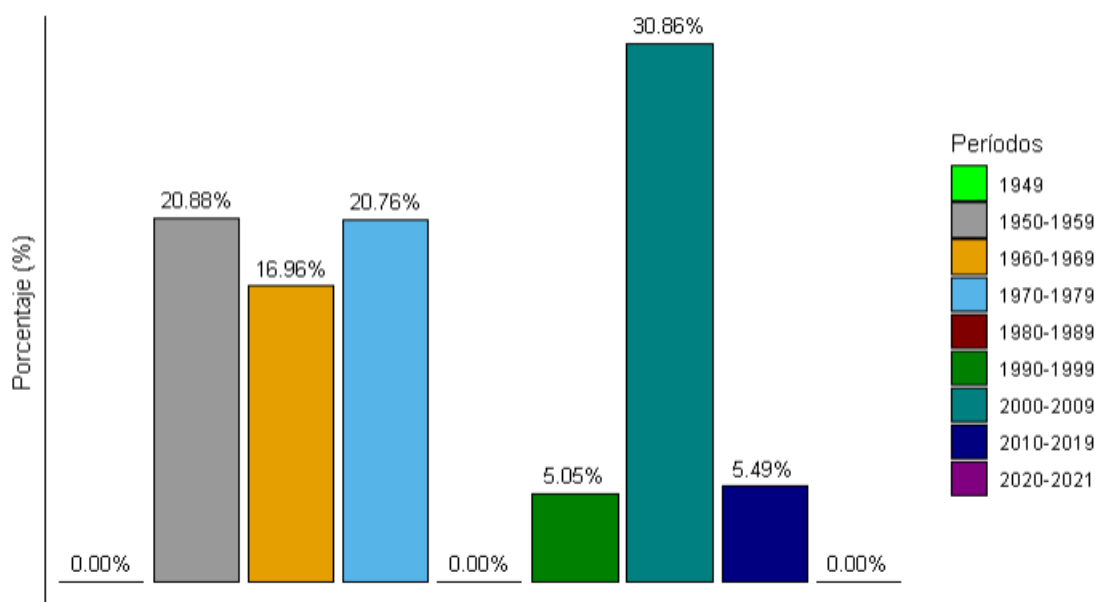


Figura 3.1.4 Valores perdidos por períodos

Fuente: Elaboración propia

Por otra parte, recordando lo expuesto en la sección 1.5.2.1.2, un gap se define como un intervalo de tiempo conformado por uno (valor perdido aislado) o más valores perdidos consecutivos. Estos se evidenciaron en la serie temporal con un total de 50 gaps, siendo el tamaño mínimo y máximo de los gaps, 25 y 2496 valores perdidos consecutivos, respectivamente. La Figura 3.1.5 muestra cómo se distribuyen estos gaps a lo largo de la serie. Así, se aprecia que las décadas de los 50, 70, 90, y 2000, fueron los períodos con mayor presencia de gaps. Más aún, al contrastar la Figura 3.1.4 con la Figura 3.1.5, se pudo determinar que la década de los 60 presentaba una cantidad reducida de gaps (2) y mayor

cantidad de valores perdidos, por lo que se notó la existencia de un gap con tamaño de 2496 valores perdidos (máximo tamaño posible de un gap) en el período, y que representó un poco más de 3 meses de ausencia de datos. Por el contrario, la década de los 90 contaba con una cantidad considerable de gaps (7) y menor cantidad de valores perdidos, lo que conllevó a la presencia de gaps con tamaños de 26 y 27 valores perdidos consecutivos, cercanos al tamaño mínimo, en el 42.86% de gaps encontrados en el período.

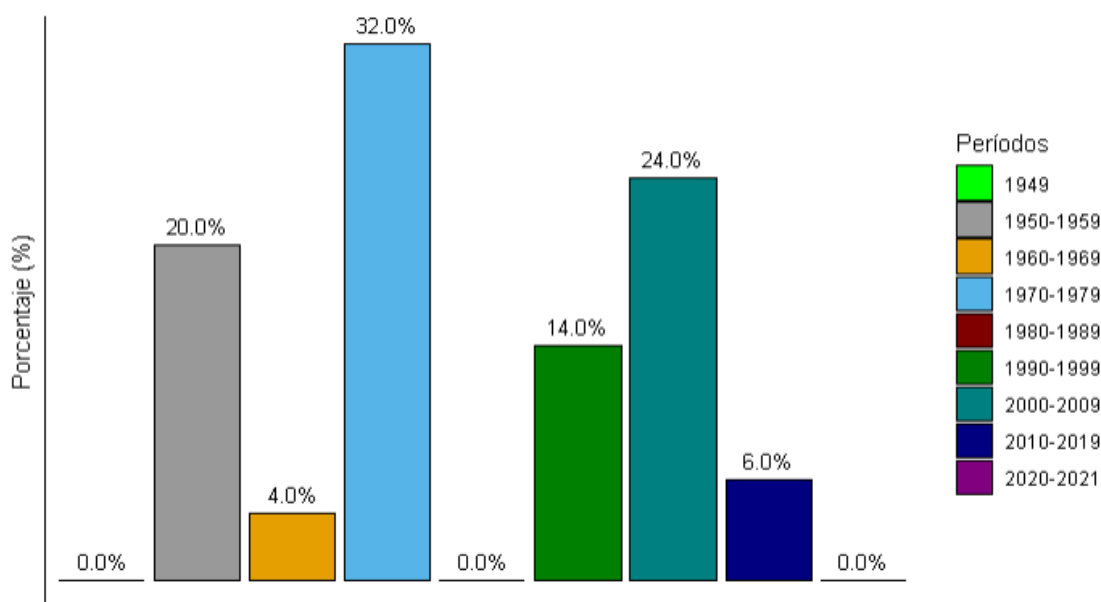


Figura 3.1.5 Presencia de gaps por períodos

Fuente: Elaboración propia

En la Figura 3.1.6 se aprecia que no hubo valores perdidos aislados, y, según lo estipulado en el apartado 1.5.2.1.2, los gaps en la serie de tiempo en análisis se categorizaron como gaps muy grandes. Adicionalmente, los tamaños de los gaps más frecuentes fueron: de 25 a 100, y de 101 a 500 valores perdidos consecutivos, representando así el 76% de los tamaños de los gaps; esta información fue posteriormente utilizada para la validación de la imputación para la serie temporal y se detalla en la siguiente sección.

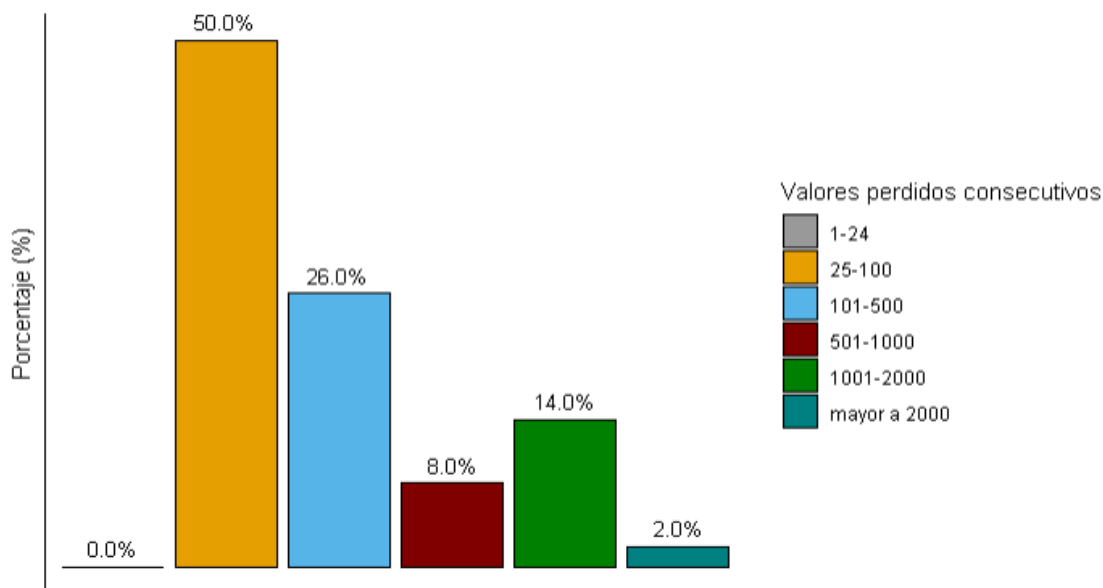


Figura 3.1.6 Tamaños de los gaps

Fuente: Elaboración propia

3.2 Imputación basada en bosques aleatorios

Para la presente sección se detallan los resultados de la imputación basada en bosques aleatorios para la serie de tiempo univariante nivel del mar horario de La Libertad y su respectiva validación a través de simulaciones de valores perdidos y métricas estadísticas.

Según lo descrito en el apartado 2.1 y 2.2, para realizar la imputación pertinente se empleó 500 árboles de decisión y muestras aleatorias de variables predictoras de tamaño $p/3$, donde p es el número de predictores, para efectuar las particiones en cada modelo estimado por bosques aleatorios. Además, se asumió que el mecanismo de pérdida de datos es del tipo MAR, puesto que los valores perdidos se presentaron de manera consecutiva y formando gaps de tamaños extensos. Por otra parte, se encontró casos particulares en los que las subseries antes del gap y después del gap, en el proceso de imputación, no contenían los datos suficientes como para entrenar los modelos según lo estipulado en la sección 2.2, por lo que se optó por extender las subseries haciendo uso de las imputaciones previas y así solventar dicho obstáculo.

Dada la gran cantidad de gaps que fueron imputados se presenta únicamente uno de ellos en la Figura 3.2.1, donde los valores imputados fueron graficados de color rojo. En la misma, se evidenció que los valores imputados tomaron la estructura subyacente de la serie de tiempo y, por tanto, visualmente, no se vio alterado el comportamiento intrínseco de la serie.

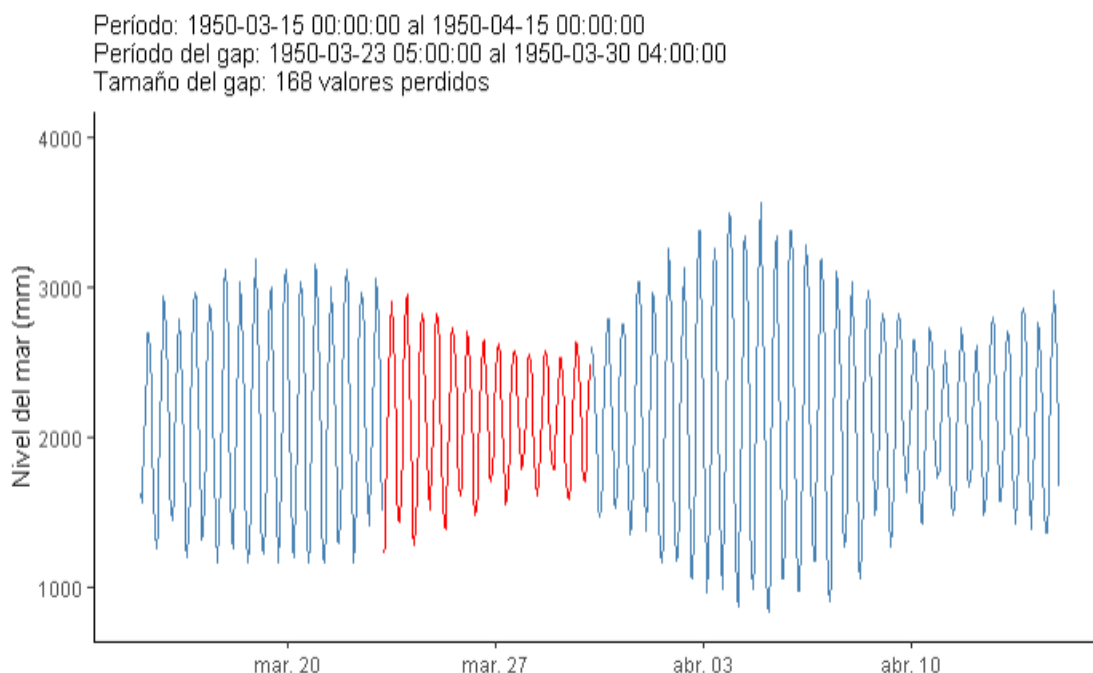


Figura 3.2.1 Serie de tiempo nivel del mar horario imputada

Fuente: Elaboración propia

Eventualmente, se validó las imputaciones realizadas según la información obtenida en la Figura 3.1.6, donde se apreció que los tamaños de los gaps con más frecuencia fueron: de 25 a 100, y de 101 a 500 valores perdidos consecutivos. Ya que el último valor perdido fue el 2016-05-01 a las 00:00:00 horas, entonces se empleó la serie temporal a partir del 2016-05-01 a las 01:00:00 horas hasta el 2021-01-31 a las 23:00:00 horas, con un total de 41686 datos para ejecutar la validación. Para dicha validación se simularon los gaps de tamaño 100 y 500 valores perdidos consecutivos de manera independiente, y consecuentemente se procedió con la imputación respectiva de cada gap. Las Figuras 3.2.2

y 3.2.3 muestran, de manera visual, cómo los valores imputados alcanzaron el comportamiento natural de la serie temporal, y, además, fueron muy cercanos a los valores reales en ambas simulaciones.

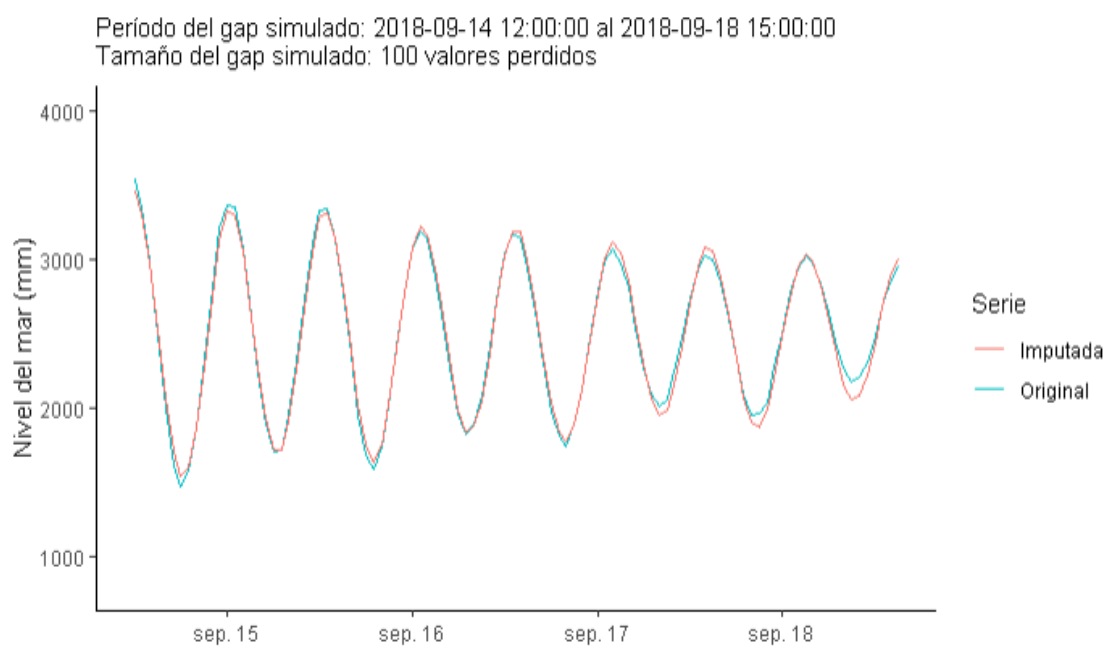


Figura 3.2.2 Serie de tiempo nivel del mar horario-Validación con 100 valores perdidos consecutivos simulados

Fuente: Elaboración propia

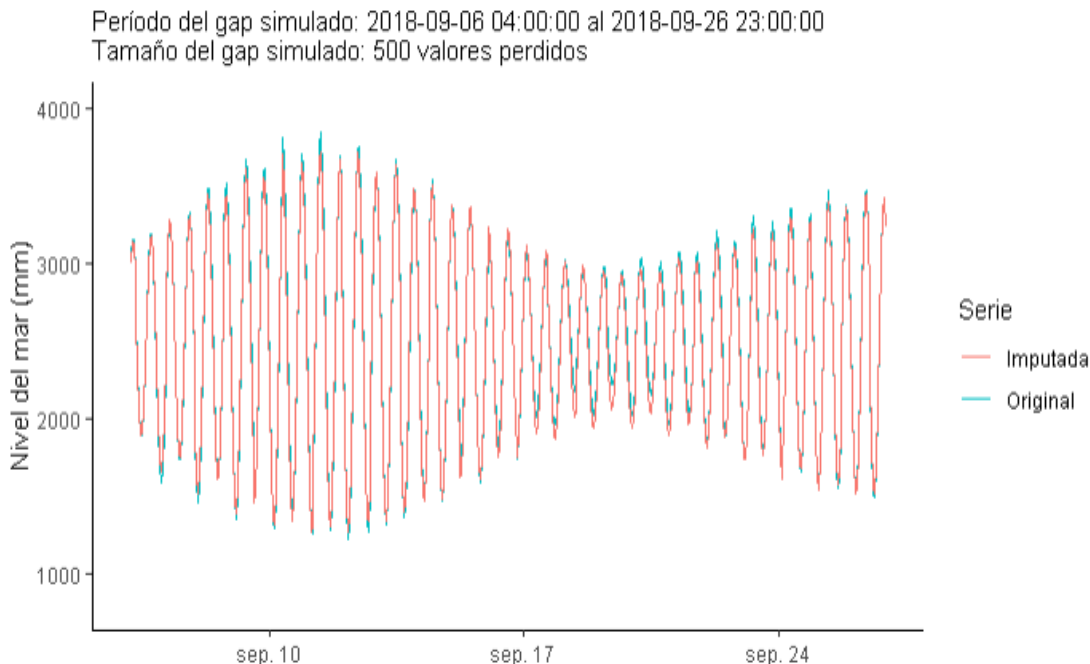


Figura 3.2.3 Serie de tiempo nivel del mar horario-Validación con 500 valores perdidos consecutivos simulados

Fuente: Elaboración propia

Para cuantificar el desempeño del método de imputación se emplearon métricas estadísticas que se detallan en la Tabla 3.2.1. Por un lado, se observó que la similaridad fue bastante alta e idéntica a pesar de que el tamaño del gap simulado aumentó. En cambio, el MAE y el RMSE aumentaron ligeramente al incrementar el tamaño del gap, estos comportamientos en estas métricas en particular son totalmente naturales de acuerdo a sus definiciones, pues hubo una mayor cantidad de valores perdidos imputados, aun así, estas medidas fueron bastantes buenas, pues demuestran la precisión del método de imputación incluso para tamaños grandes de gaps. Por otro lado, el FB y el FSD también presentaron un leve aumento al incrementar el tamaño del gap, no obstante, recordando lo enunciado en el apartado 2.4, los valores cercanos a cero indicaron que los modelos de imputación mostraron un buen desempeño y los valores imputados capturaron la estructura subyacente de la serie original.

Tabla 3.2.1 Métricas para la validación de la imputación*Fuente: Elaboración propia*

Métricas de validación	Tamaño del gap simulado	
	100	500
Similaridad	0.9810124	0.9830628
MAE	40.39693	45.49603
RMSE	48.78186	58.32874
FB	0.001793809	0.009770325
FSD	0.005102561	0.0165937

Por último, se evaluaron los estadísticos descriptivos de la serie temporal nivel del mar horario y la serie resultante de la imputación efectuada; esto se evidencia en la Tabla 3.2.2, donde se identificó que los valores de los estadísticos prácticamente no sufrieron cambios severos, por el contrario, son valores muy similares.

Tabla 3.2.2 Estadísticos descriptivos de la serie de tiempo nivel del mar horario y su imputación*Fuente: Elaboración propia*

Estadísticos descriptivos	Serie original	Serie imputada
Mínimo	683 <i>mm</i>	683 <i>mm</i>
Máximo	4120 <i>mm</i>	4120 <i>mm</i>
1° cuartil	1840 <i>mm</i>	1833 <i>mm</i>
2° cuartil (mediana)	2340 <i>mm</i>	2340 <i>mm</i>
3° cuartil	2847 <i>mm</i>	2847 <i>mm</i>
Media	2339 <i>mm</i>	2338 <i>mm</i>
Desviación estándar	612.1791 <i>mm</i>	611.911 <i>mm</i>
Varianza	374 763.3 <i>mm</i> ²	374 435 <i>mm</i> ²

3.3 Filtro de Doodson

En esta sección se presenta el resultado de aplicar el filtro de Doodson en la determinación de la serie de tiempo nivel medio del mar diario de La Libertad, que posteriormente dio paso para la obtención de la serie nivel medio del mar mensual, y, finalmente, la serie anomalías del nivel medio del mar mensual.

Basado en lo expuesto en la sección 2.3, se determinó la serie temporal nivel medio del mar diario a través del filtro de Doodson, el resultado se presenta en la Figura 3.3.1. Al aplicar el filtro el período de datos se vio ligeramente modificado, es decir, pasó de un período horario del 1949-09-01 a las 05:00:00 horas al 2021-01-31 a las 23:00:00 horas, al período diario resultante del 1949-09-02 al 2021-01-30, por lo que se perdió dos días en el proceso de filtrado. Así, la serie de tiempo nivel medio del mar diario contó con 26084 observaciones.

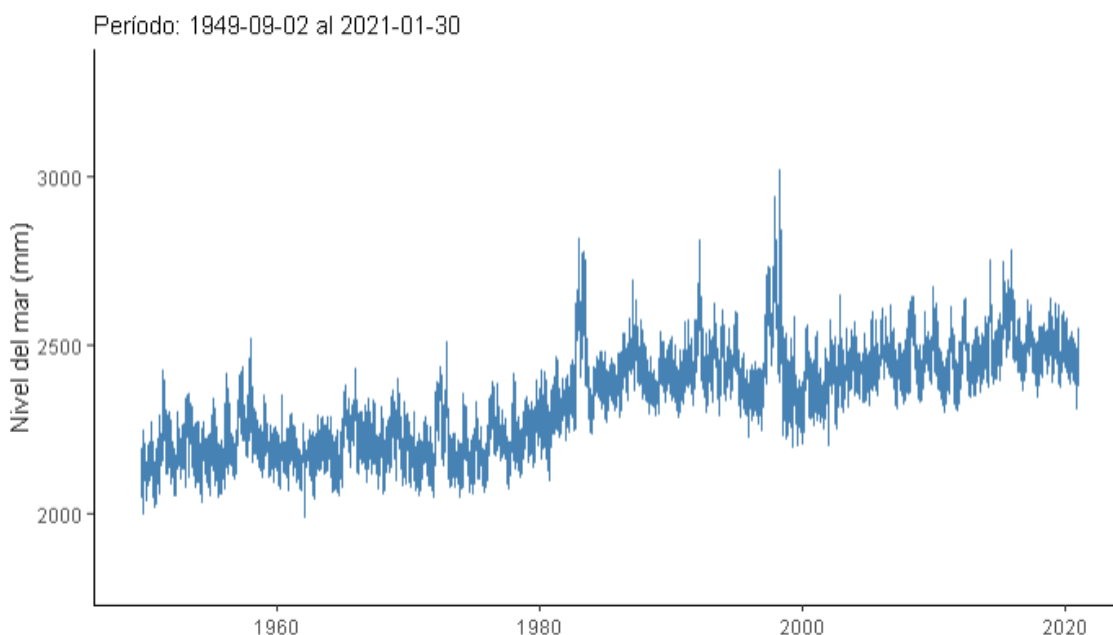


Figura 3.3.1 Serie de tiempo nivel medio del mar diario mediante el filtro de Doodson

Fuente: Elaboración propia

Recordando lo enunciado en el apartado 1.5.4, el filtro de Doodson fue utilizado debido a lo recomendado por UNESCO/IOC, de tal manera que, se comparó la serie resultante tras

aplicar el filtro de Doodson con la serie de registros diarios obtenidos de la UHSLC. El período de datos diarios de la UHSLC comprende del 1949-09-02 al 2021-01-31, con un total de 26085 observaciones, de las cuales 25249 fueron datos válidos (96.80%) y 836 fueron valores perdidos (3.2%). Para realizar la comparación se aisló la observación del día 2021-01-31 en la serie determinada por la UHSLC para que tanto la serie por el filtro de la UHSLC como por el filtro de Doodson dispongan de la misma cantidad de observaciones. En la Figura 3.3.2 se aprecia que la estructura y los valores de la serie por el filtro de Doodson fueron prácticamente idénticos, visualmente, que los proporcionados por la UHSLC.



Figura 3.3.2 Serie de tiempo nivel medio del mar diario-Filtro de Doodson vs UHSLC

Fuente: Elaboración propia

Los 25248 datos válidos de la UHSLC fueron comparados con los datos resultantes por el filtro de Doodson según la indexación en el tiempo y se cuantificó dicha comparación entre los filtros empleando las mismas métricas estadísticas que en la sección 3.2, y se expone en la Tabla 3.3.1. En la misma, se evidenció muy alta similitud, y medidas considerablemente cercanas a cero por parte del FB y FSD, indicando que las estructuras de las series

determinadas por los dos filtros fueron aproximadamente idénticas; en cambio, el MAE y el RMSE resultaron bastante bajos e indicaron que los valores de ambas series fueron muy parecidos. Más aún, el valor del RMSE obtenido, corroboró lo señalado en el apartado 1.5.4, donde UNESCO/IOC (2002) afirmaban que la desviación estándar de la diferencia entre los dos filtros, para los promedios diarios, fue de 4.7 *mm* en sus comparaciones; para este caso en particular, fue de aproximadamente 4.14 *mm*.

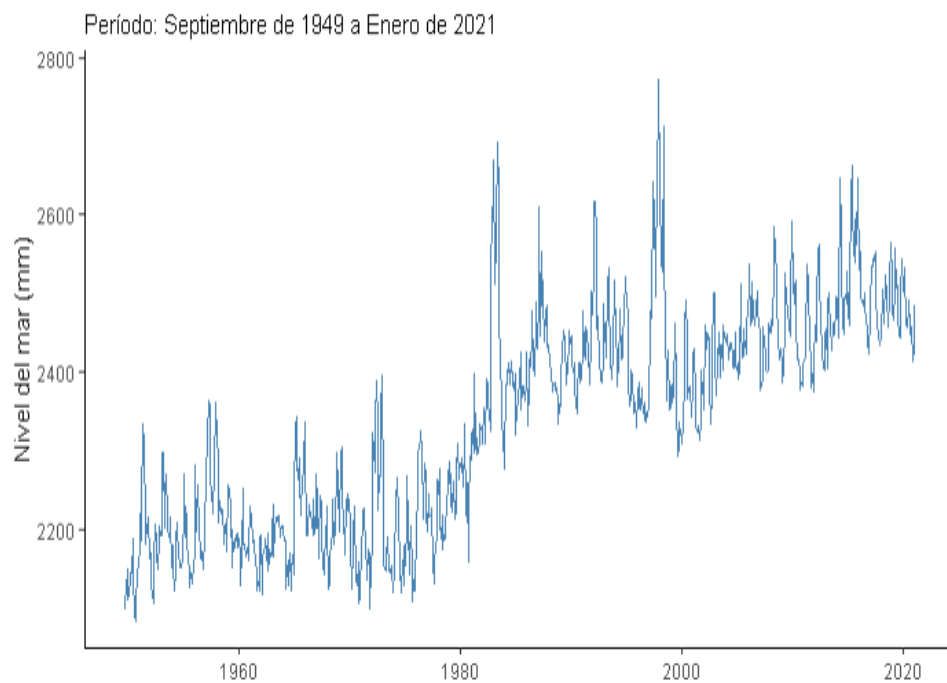
Tabla 3.3.1 Métricas para la comparación entre el filtro de Doodson y la UHSLC

Fuente: Elaboración propia

Similaridad	0.9971867
MAE	2.883303
RMSE	4.136394
FB	0.00004082749
FSD	0.000671696

Por último, se determinaron las series de tiempo nivel medio del mar mensual y anomalías del nivel medio del mar mensual según lo estipulado en la sección 2.3, y se visualizan en la Figura 3.3.3. Ambas series temporales presentaron 857 observaciones.

a)



b)

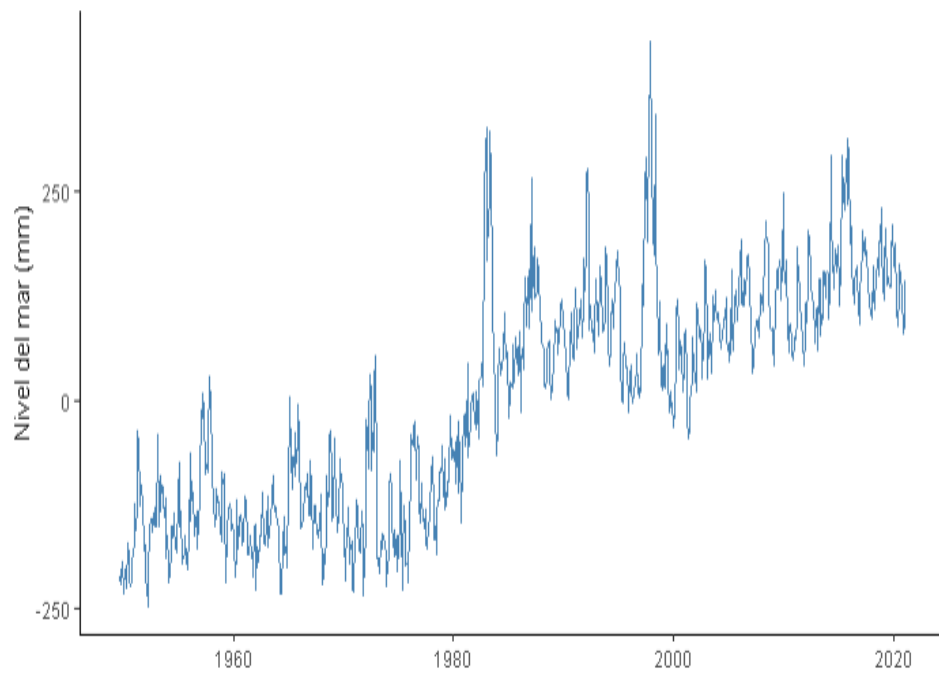


Figura 3.3.3 (a) Serie de tiempo nivel medio del mar mensual, (b) Serie de tiempo anomalías del nivel medio del mar mensual

Fuente: Elaboración propia

3.4 Análisis Espectral Singular

En esta sección se exponen los resultados del análisis espectral singular (SSA) efectuado para la serie de tiempo anomalías del nivel medio del mar mensual de La Libertad.

Como se observa en la Figura 3.3.3 (b), la serie muestra una potencial tendencia compleja, por lo que se siguió lo estipulado por Golyandina et al. (2018), donde mencionaron que para este tipo de series se recomienda, en primera instancia, extraer la tendencia mediante el SSA con un parámetro L mínimo, para, en una segunda instancia, aplicar el SSA con un parámetro L grande en la serie residual resultante de la extracción de la tendencia a la serie original para estimar los componentes periódicos subyacentes en la misma.

Por lo mencionado, se aplicó el SSA para la serie temporal anomalías del nivel medio del mar mensual según lo enunciado en el apartado 2.5 con un tamaño de ventana $L = 48$. En la figura 3.4.1 se presentan los valores propios asociados a los eigentriples de los componentes reconstruidos de la serie a partir del SSA aplicado. En la misma, se evidenció que el primer valor propio se encontró bastante alejado de los demás, por lo que el componente reconstruido asociado a dicho valor propio correspondía al componente de tendencia de la serie; esto se corroboró, más adelante, al evaluar su vector propio asociado.

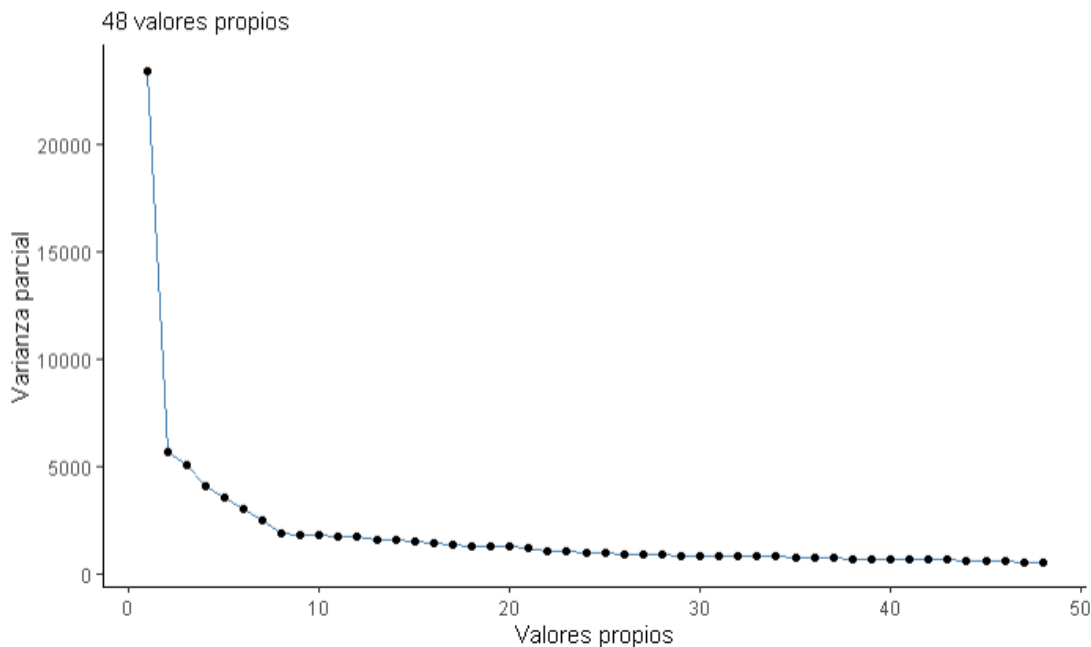


Figura 3.4.1 Valores propios para el SSA con $L=48$

Fuente: Elaboración propia

El vector propio asociado al primer valor propio se visualiza en la Figura 3.4.2, y se evidenció un comportamiento donde varía de manera paulatina (slowly-varying); dicho comportamiento fue descrito por Golyandina et al. (2018) como una característica del componente de tendencia. Por ello, a partir de la matriz correspondiente al eigentriple asociado al primer valor y vector propio se estimó el componente de tendencia de la serie de tiempo anomalías del nivel medio del mar mensual; esto se muestra en la Figura 3.4.3, donde se observó que la tendencia estimada capturaba adecuadamente el comportamiento de la serie y correspondía a una tendencia no lineal.

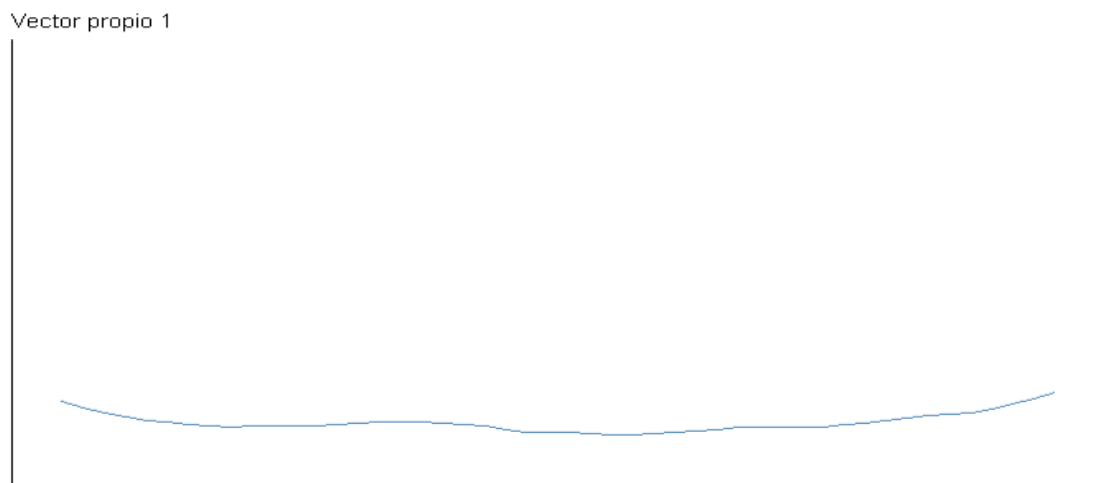


Figura 3.4.2 Primer vector propio para el SSA con $L=48$

Fuente: Elaboración propia

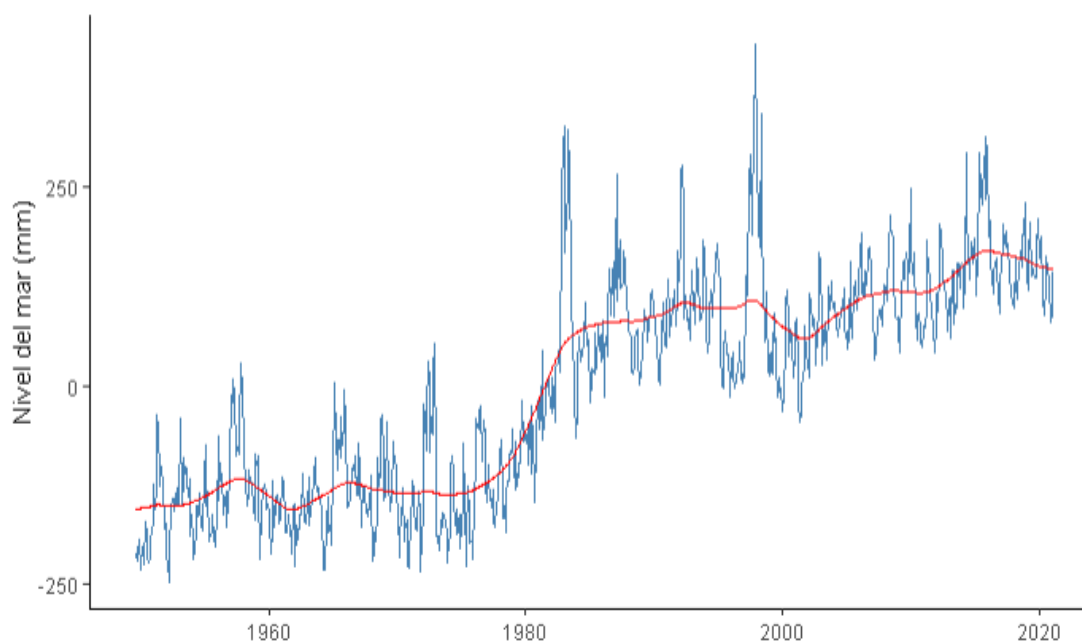


Figura 3.4.3 Tendencia no lineal estimada para la serie anomalías del nivel medio del mar mensual

Fuente: Elaboración propia

Sin embargo, para evaluar las tasas de cambio en el nivel del mar se determinó sus tendencias mediante regresión lineal a partir de la tendencia no lineal estimada, según lo descrito en la sección 2.6, correspondientes a los períodos: a) enero de 1993 a diciembre de

2020 (azul), y b) enero de 2010 a diciembre de 2020 (anaranjado). Estos períodos fueron analizados puesto que, desde la llegada de los datos de altimetría por satélite en 1993 se han realizado estimaciones más precisas y confiables sobre las tasas de cambio en el nivel medio global del mar, entre ellas la estimación realizada por Church & White (2011) con la cual se pudo evaluar la condición actual del nivel del mar en la estación fija de La Libertad. Lo mencionado se aprecia en la Figura 3.4.4, donde, a pesar de que no hubo un ajuste lineal óptimo para la serie en los períodos evaluados, la tendencia lineal estimada para el período (a) permitió identificar un incremento aproximado de $3.2 \pm 0.12 \text{ mm/año}$ en el nivel del mar esperado. Esta estimación fue similar a la tasa de aumento estimada de $3.2 \pm 0.4 \text{ mm/año}$ en el nivel medio global del mar para el período de 1993 a 2009 con datos de altimetría expuesto por Church & White (2011). En cambio, para el período (b) se estimó un incremento aproximado de $4.05 \pm 0.35 \text{ mm/año}$ en el nivel del mar esperado. En consecuencia, se espera que el nivel del mar haya aumentado a una tasa más rápida en la última década (2010-2020) en comparación con los últimos 27 años.

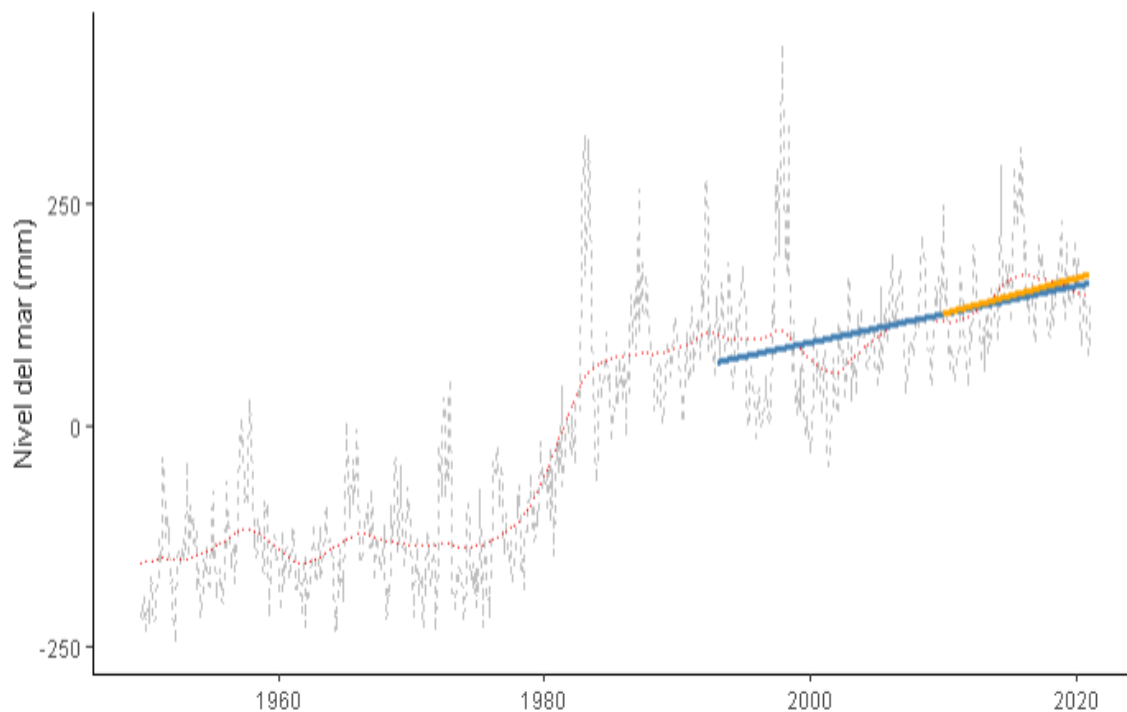


Figura 3.4.4 Tendencias lineales estimadas para la serie anomalías del nivel medio del mar mensual

Fuente: Elaboración propia

Posteriormente, en la Figura 3.4.5 se expone la serie residual que resultó de la extracción de la tendencia no lineal estimada de la serie de tiempo anomalías del nivel medio del mar mensual. A dicha serie residual se le aplicó el SSA con $L = 432$ según lo establecido en las secciones 2.5 y 1.5.5.1. Se limitó la cantidad de valores y vectores propios a calcular a 50, esto debido a la carga computacional a la que el algoritmo está inmerso al determinar cantidades grandes de valores y vectores propios.

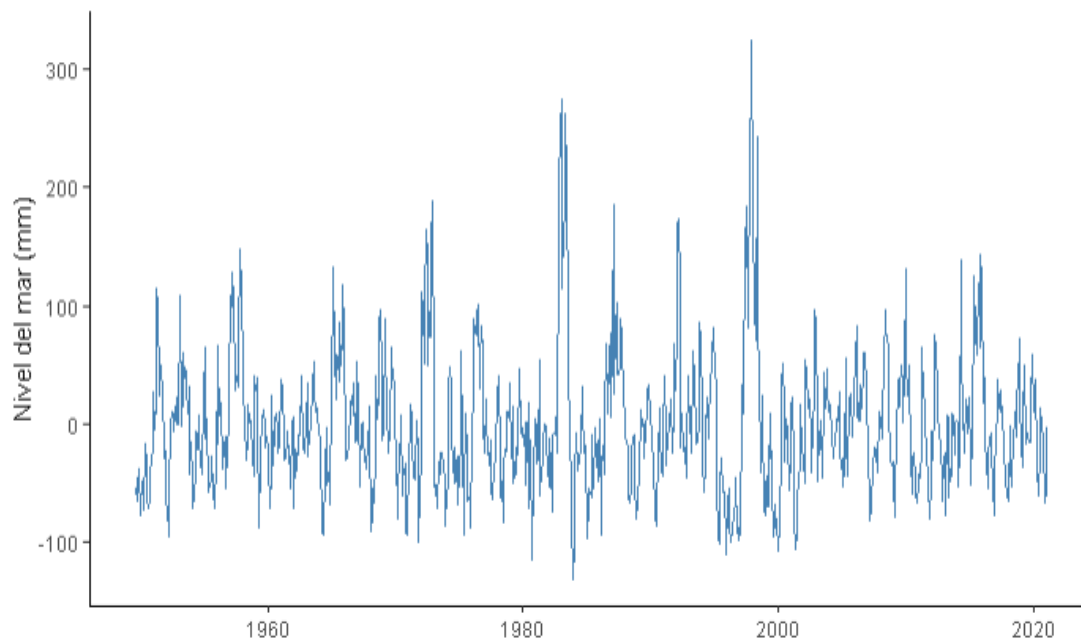


Figura 3.4.5 Serie de tiempo residual

Fuente: Elaboración propia

Así, en la Figura 3.4.6 se muestran los valores propios asociados a los eigentriples de los componentes reconstruidos de la serie residual a partir del SSA aplicado, y se visualizó que los componentes reconstruidos correspondientes a los valores propios 1 al 18 explicaban mayor varianza parcial de la serie residual en comparación a los demás. Por consiguiente, se realizó un prefiltrado para la serie residual a través del SSA mencionado, obteniendo así, una serie resultante de la adición de los componentes reconstruidos asociados a los 18 valores propios más representativos según su varianza parcial explicada.

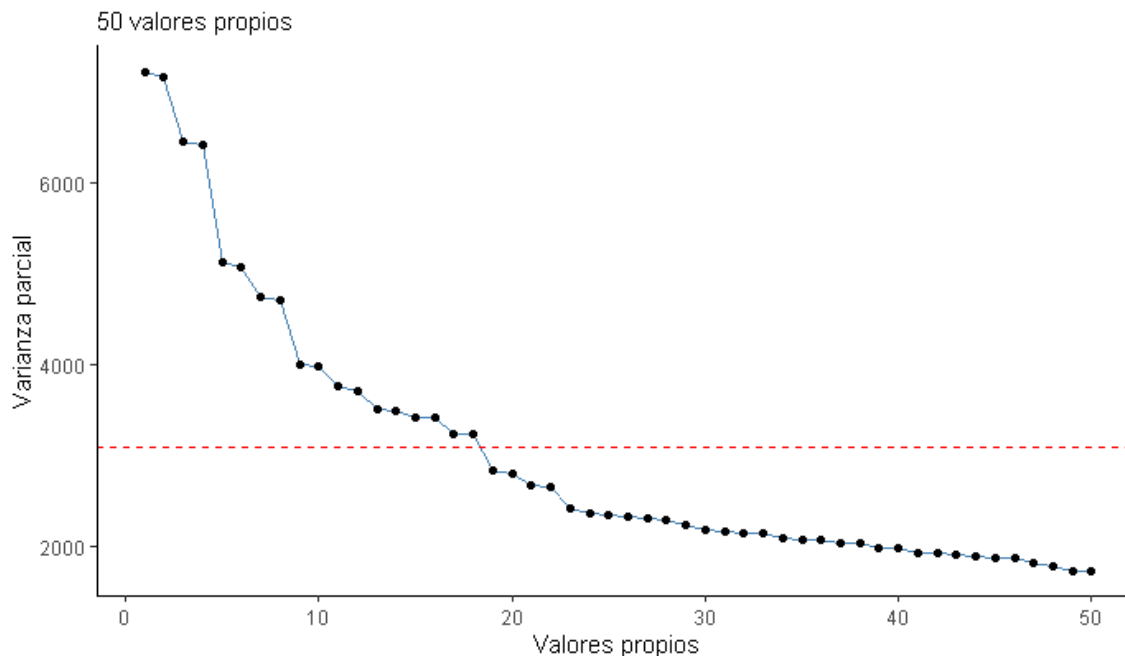


Figura 3.4.6 Valores propios para el SSA con $L=432$

Fuente: Elaboración propia

Se estimó la densidad espectral mediante el periodograma suavizado para dicha serie prefiltrada según lo estipulado en el apartado 2.7, y se expone en la Figura 3.4.7. Para determinar el periodograma suavizado se empleó un taper correspondiente al 10% para cada extremo de la serie, y una convolución de dos ventanas espectrales Daniell modificadas con $m = 5$ y ancho de banda $BW = 0.00203$. El periodograma mostró 7 componentes periódicos, de los cuales solo 5 de ellos obtuvieron espectros de potencia (energía) considerablemente altos, y, por ende, fueron los más representativos.

El objetivo de este prefiltrado para la serie residual fue de aislar el ruido subyacente en la serie para mejorar la identificación de los componentes periódicos en el periodograma suavizado.

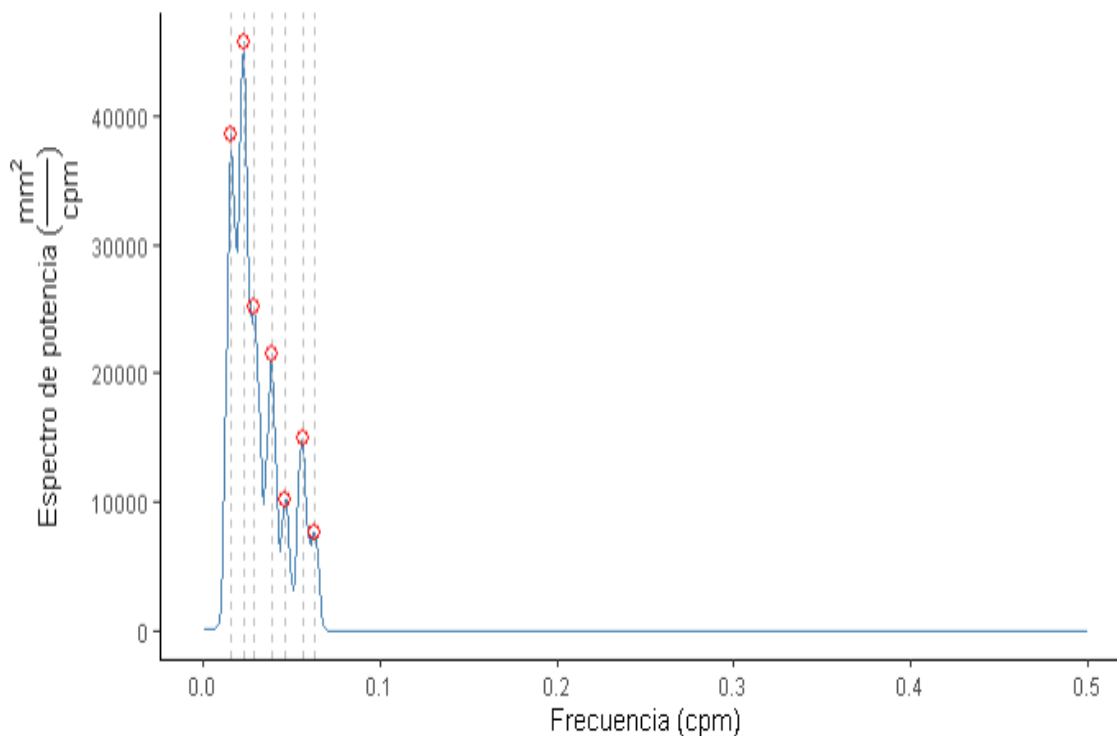


Figura 3.4.7 Periodograma suavizado

Fuente: Elaboración propia

En la Tabla 3.4.1 se aprecian las frecuencias, periodicidades y espectros de potencia asociados a los componentes periódicos encontrados en el análisis. Por lo tanto, según lo detallado en la Tabla 3.4.1 y Figura 3.4.7, se identificó a la frecuencia dominante como la frecuencia correspondiente al componente periódico 1 con periodicidad 3.6 años. Además, se notó que la frecuencia 0.0462963 *cpm*, correspondiente al componente periódico 6, es un armónico de la frecuencia dominante.

La NCEI (s.f.) ha estipulado que El Niño-Oscilación del Sur (ENOS) es un fenómeno que se presenta en el Pacífico tropical con una periodicidad irregular aproximada de entre 2 y 7 años, causando fluctuaciones climatológicas alrededor del mundo. Contrastando lo mencionado, se evidenció que los 5 componentes periódicos más representativos corresponden potencialmente al fenómeno ENOS, según su periodicidad. En particular, el

componente periódico 1 correspondiente a la frecuencia dominante se encontró asociado al fenómeno ENOS.

Tabla 3.4.1 Frecuencias, periodicidades y espectros de potencia de los componentes periódicos ordenados según su importancia

Fuente: Elaboración propia

Componente periódico	Espectro de potencia (mm^2/cpm)	Frecuencia (cpm)	Periodicidad
1	45675.19	(*) 0.02314815	3.6 años
2	38506.29	0.0162037	5.14 años
3	25152.32	0.03935185	2.12 años
4	21419.18	0.02893519	2.88 años
5	14988.95	0.05671296	17.63 meses
6	10229.16	0.0462963	21.6 meses
7	7642.376	0.06365741	15.7 meses

(*) *Frecuencia dominante*

Adicionalmente, en la Figura 3.4.6 se observó que los valores propios se agrupaban, de cierta forma, en pares, lo que dio un indicio a su potencial agrupamiento por eigentriples en la estimación particular de los componentes periódicos, que se corroboró al evaluar sus vectores propios asociados. Esto se aprecia en la Figura 3.4.8, donde cada par de vectores propios mostraron un comportamiento muy parecido, con la misma frecuencia, y su diferenciación pasaba por cierto desfase, lo que permitió agrupar las matrices correspondientes a los eigentriples asociados a estos pares de vectores para reconstruir los

componentes periódicos. En la Figura 3.4.9 y Figura 3.4.10 se exponen los 5 componentes periódicos más importantes reconstruidos.

Más aún, la Figura 3.4.6 contrastó la importancia de cada componente periódico según su varianza parcial explicada por cada par de valores propios, un claro ejemplo es el componente periódico 1 que fue reconstruido al agrupar las matrices asociadas al primer par de valores propios, donde este par de valores propios representaban la mayor cantidad de varianza parcial capturada, por ende, el componente periódico 1 con periodicidad 3.6 años, potencialmente generado por el fenómeno ENOS, es el más relevante en la variabilidad de la serie de tiempo anomalías del nivel medio del mar mensual.

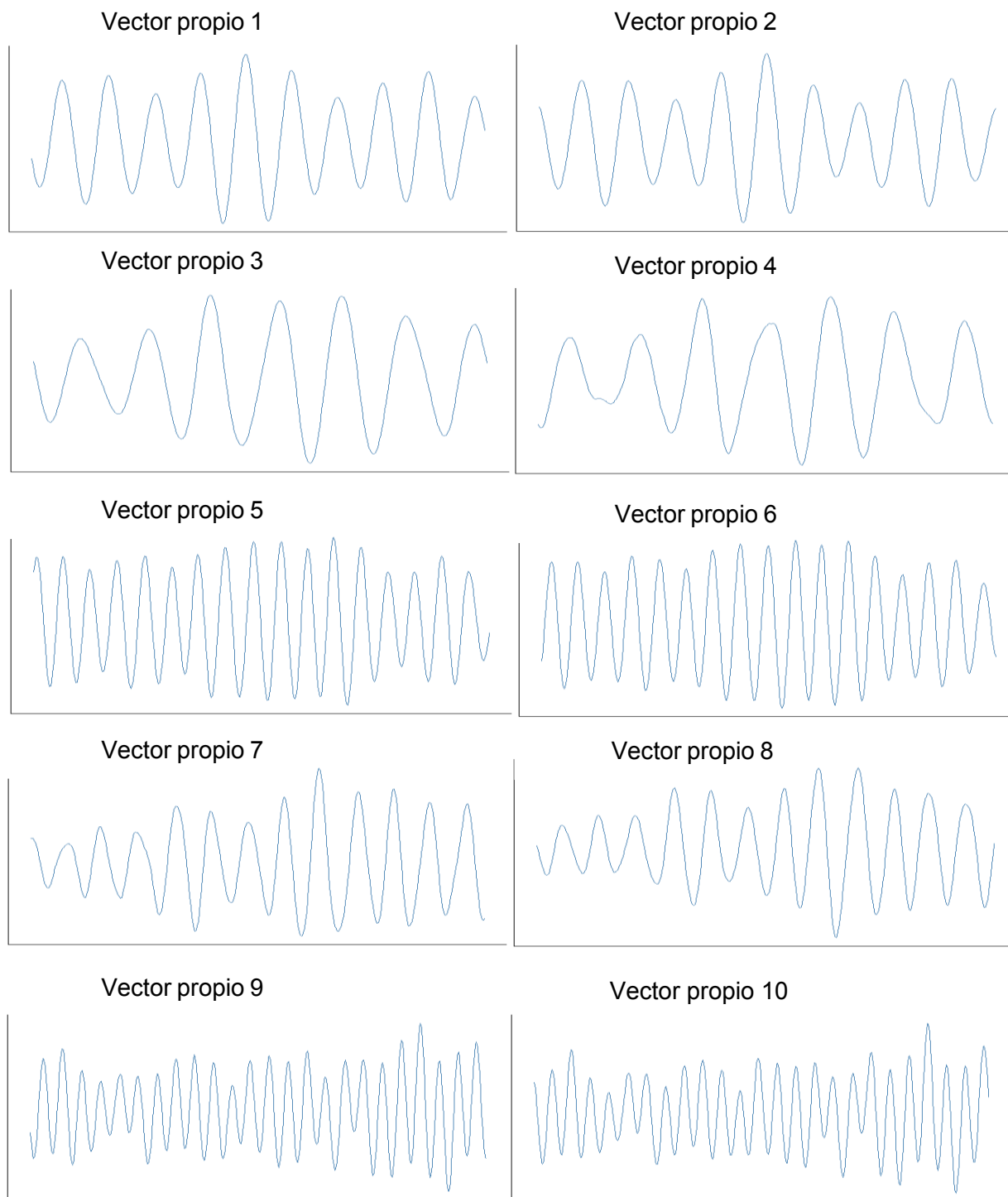


Figura 3.4.8 Vectores propios para el SSA con $L=432$

Fuente: Elaboración propia

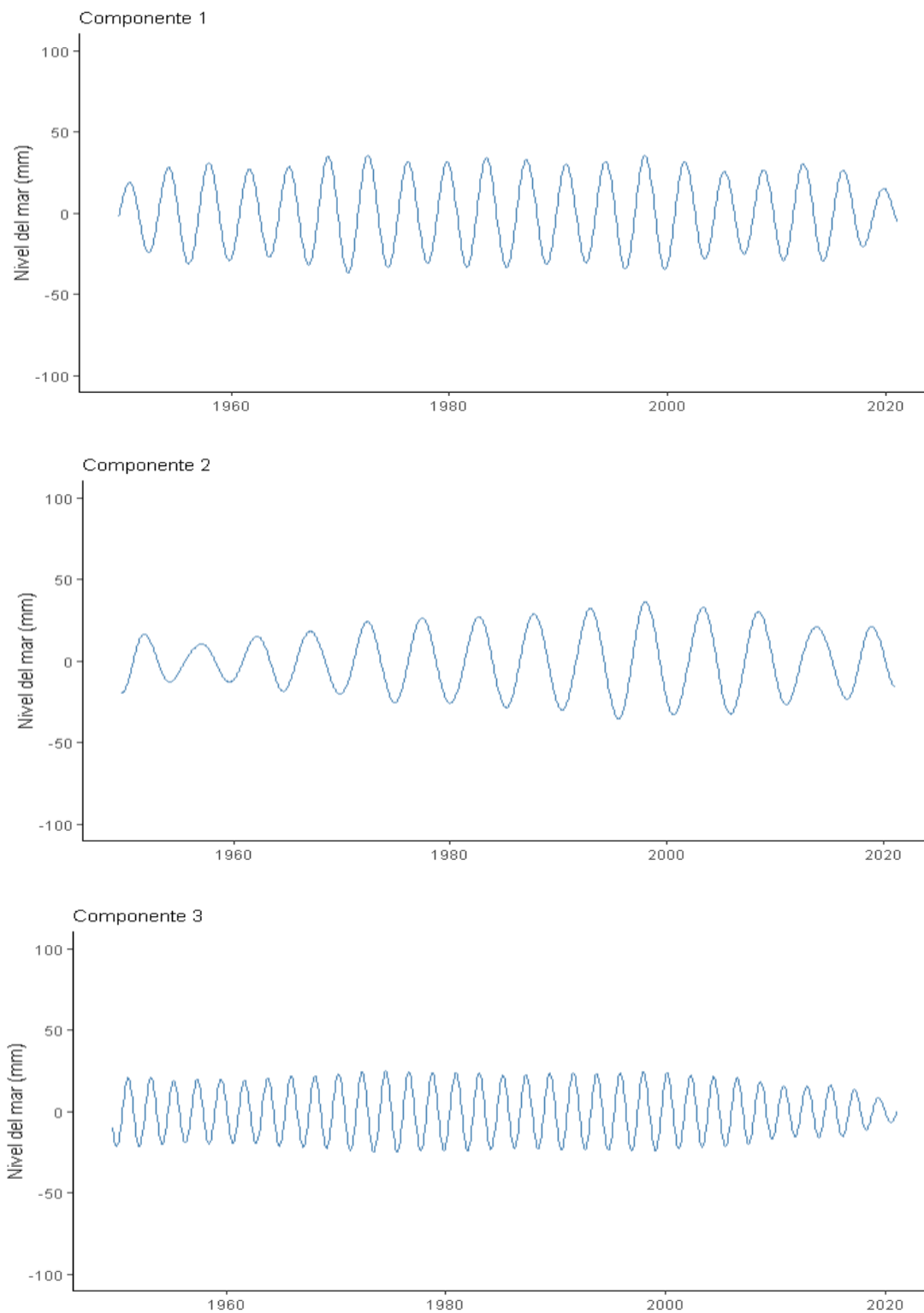


Figura 3.4.9 Componentes periódicos 1,2 y 3 estimados para la serie anomalías del nivel medio del mar mensual

Fuente: Elaboración propia

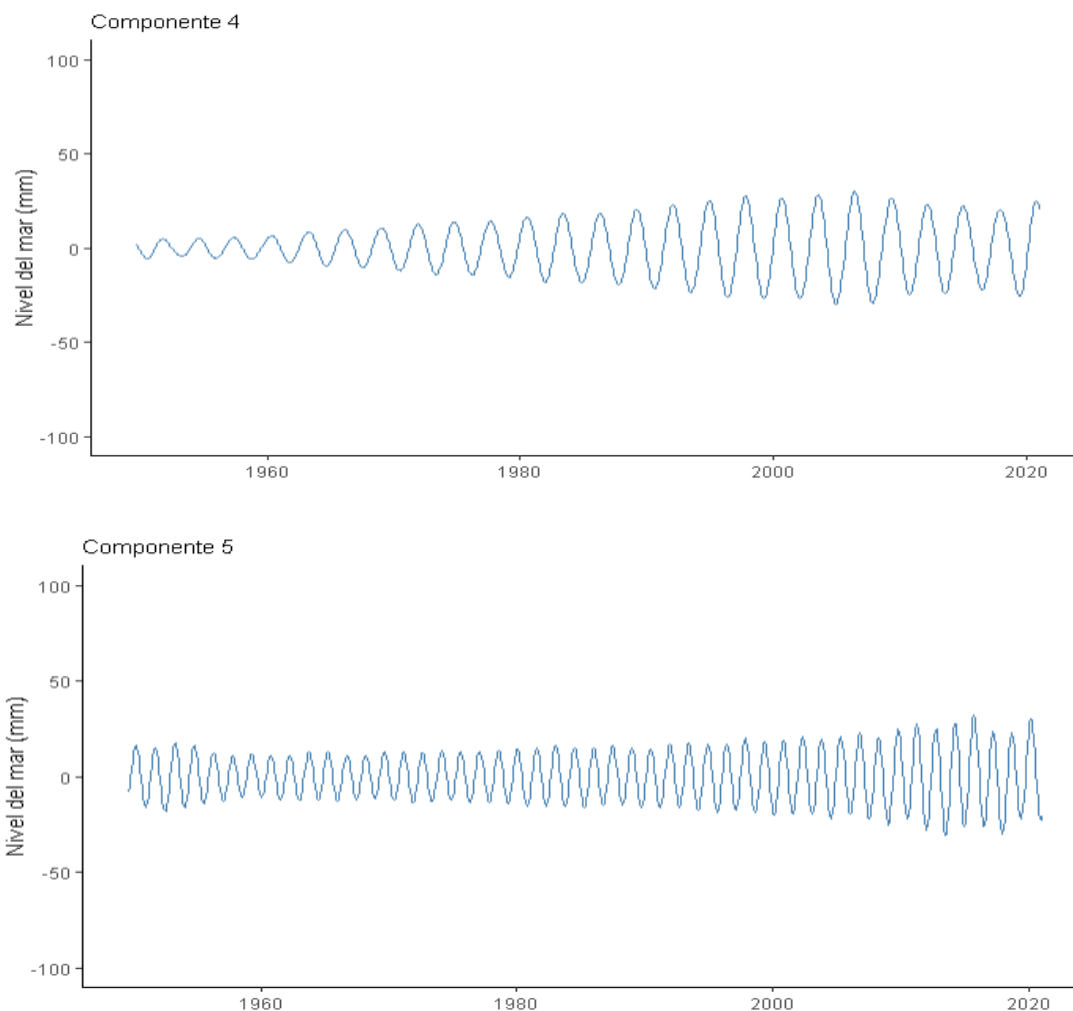


Figura 3.4.10 Componentes periódicos 4 y 5 estimados para la serie anomalías del nivel medio del mar mensual

Fuente: Elaboración propia

Finalmente, los 5 componentes periódicos estimados más representativos se adicionaron con la tendencia no lineal estimada para obtener la serie reconstruida final y fue comparada con la serie de tiempo anomalías del nivel medio del mar mensual; esto se visualiza en la Figura 3.4.11. En la misma, se observó que la serie reconstruida mediante los componentes estimados por SSA representó y capturó apropiadamente la estructura inherente de la serie temporal anomalías del nivel medio del mar mensual. Más aún, la serie reconstruida capturó adecuadamente los ENOS más destacados históricamente de los años

1957-1958, 1972-1973, 1982-1983, y 1997-1998 en el Ecuador (Organización Panamericana de la Salud, 2000).

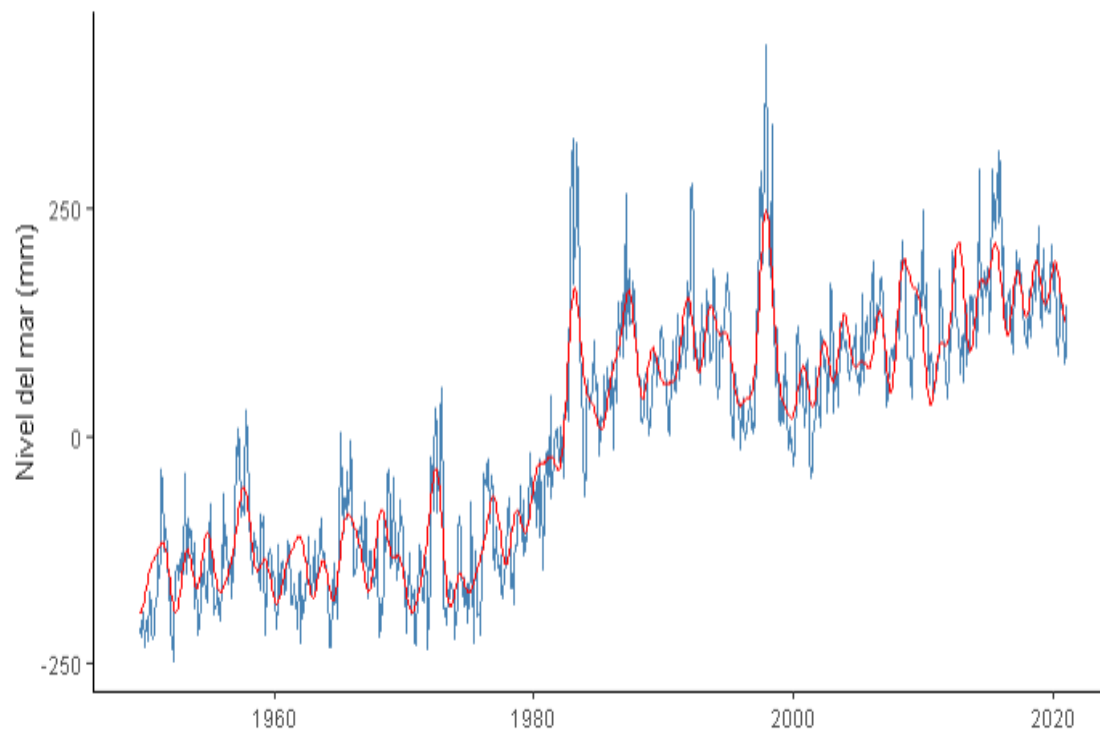


Figura 3.4.11 Serie reconstruida a partir de los componentes estimados por SSA

Fuente: Elaboración propia

CAPÍTULO 4

4. CONCLUSIONES Y RECOMENDACIONES

4.1 Conclusiones

- El método de imputación propuesto basado en bosques aleatorios capturó la estructura subyacente de la serie de tiempo nivel del mar horario de La Libertad y, por tanto, no alteró el comportamiento natural de la serie y permitió realizar el análisis espectral singular pertinente.
- El filtro de Doodson fue eficaz en la determinación de la serie de tiempo nivel medio del mar diario, pues su desviación estándar con respecto al filtro de la UHSLC fue de alrededor de 4.14 mm , y, tanto las estructuras, como los valores de las series determinadas por ambos filtros, fueron aproximadamente idénticas.
- La tendencia estimada para la serie de tiempo anomalías del nivel medio del mar mensual de La Libertad correspondía a una tendencia no lineal compleja.
- La tendencia lineal estimada para la serie de tiempo anomalías del nivel medio del mar mensual correspondiente al período de 1993-2020, permitió identificar un incremento aproximado de $3.2 \pm 0.12 \text{ mm/año}$ en el nivel del mar esperado, el cual fue similar a la tasa de aumento en el nivel medio global del mar estimado por datos de altimetría para el período de 1993-2009 por Church & White (2011).
- Se espera que el nivel del mar haya aumentado a una tasa más rápida en la última década en comparación a los últimos 27 años, pues se estimó un incremento aproximado de $4.05 \pm 0.35 \text{ mm/año}$ para el período de 2010-2020 en el nivel del mar esperado de La Libertad.

- Los 5 componentes periódicos estimados más representativos para la serie de tiempo anomalías del nivel medio del mar mensual de La Libertad fueron asociados al fenómeno ENOS.
- El nivel del mar de La Libertad se encontró predominado por el componente periódico con periodicidad de 3.6 años asociado al fenómeno interanual ENOS.
- La serie reconstruida mediante los componentes estimados por SSA capturó adecuadamente la estructura intrínseca de la serie de tiempo anomalías del nivel medio del mar mensual de La Libertad, e identificó los ENOS más destacados históricamente en el Ecuador.

4.2 Recomendaciones

- La carga computacional para realizar las imputaciones según el método propuesto es bastante alta, sobre todo cuando se trata con series de tiempo extensas, como con la que se trabajó para este proyecto, que presentan una gran cantidad de gaps con tamaños muy grandes. Por lo cual, se recomienda emplear equipos computacionales con un buen procesador y un mínimo de memoria RAM de 8 GB.
- Para casos particulares en los que las subseries antes del gap y después del gap, en el proceso de imputación, no contengan los datos suficientes como para entrenar los modelos de machine learning (bosques aleatorios), se recomienda extender las subseries empleando imputaciones de gaps previas, ya sea extender la subserie antes del gap, después del gap, o ambas, según lo amerite.
- Dada la compleja tendencia no lineal que presentó la serie de tiempo anomalías del nivel medio del mar mensual, se recomienda ajustar tendencias cuadráticas que permitan identificar potenciales aceleraciones en el aumento del nivel del mar y profundizar en el estudio de sus tendencias.

- El tamaño de ventana L para la estimación de la tendencia a través del SSA en series de tiempo con tendencias complejas como con la que se ha trabajado en el presente proyecto, influye en la adecuada descomposición de dicho componente, pues al emplear un tamaño de L muy pequeño puede causar que el componente de tendencia se vea mezclado con ciertos componentes periódicos, y estos no podrían ser identificados apropiadamente al analizar la serie residual. Por lo mencionado, se recomienda, para la estimación de la tendencia, emplear un L mínimo de partida para paulatinamente incrementarlo hasta capturar la estructura natural de la serie en análisis sin “absorber” sus componentes periódicos.
- Se recomienda realizar análisis espectral para las series de tiempo nivel medio del mar diario y nivel medio del mar anual de la estación fija de La Libertad, para identificar frecuencias dominantes y sus potenciales fenómenos generadores de fluctuaciones a nivel intra-anual e inter-decadal, respectivamente.

BIBLIOGRAFÍA

- Banco de Desarrollo de América Latina. (2017). *Diagnóstico y proyección de vulnerabilidades frente a la variabilidad y cambio climático en la ciudad de Guayaquil*.
- Bayot, B., & Cornejo Rodríguez, M. (1996). Evidencia de ondas ecuatoriales en Salinas y Galápagos. *Acta Oceanográfica Del Pacífico*, 8(1).
- Beşel, C., & Tanır Kayıkçı, E. (2020). Investigation of Black Sea Mean Sea Level Variability By Singular Spectrum Analysis. *International Journal of Engineering and Geosciences (IJEG)*, V(1), 33–41. <https://doi.org/10.26833/ijeg.580510>
- Bokde, N., Beck, M. W., Martínez Álvarez, F., & Kulat, K. (2018). A novel imputation methodology for time series based on pattern sequence forecasting. *Pattern Recognition Letters*, 116, 88–96. <https://doi.org/10.1016/j.patrec.2018.09.020>
- Braker, J. . (1994). Singular Spectrum Analysis of North Sea mean sea level time series. In *Rijkswaterstaat Report Database*. https://puc.overheid.nl/doc/PUC_131942_31/1
- Caldwell, P. C., Merrifield, M. A., & Thompson, P. R. (2015). *Sea level measured by tide gauges from global oceans--the Joint Archive for Sea Level holdings (NCEI Accession 0019568), Version 5.5, NOAA National Centers for Environmental Information, Dataset. Fast Delivery data obtained on March 3, 2021*. <https://doi.org/10.7289/V5V40S7W>
- Cazenave, A., Dieng, H. B., Meyssignac, B., Von Schuckmann, K., Decharme, B., & Berthier, E. (2014). The rate of sea-level rise. *Nature Climate Change*, 4(5), 358–361. <https://doi.org/10.1038/nclimate2159>
- Church, J. A., & White, N. J. (2011). Sea-Level Rise from the Late 19th to the Early 21st Century. *Surveys in Geophysics*, 32(4–5), 585–602. <https://doi.org/10.1007/s10712-011-9119-1>
- Cryer, J. D., & Chan, K.-S. (2008). *Time Series Analysis With Applications in R* (2nd ed.). Springer.
- Flores, A., Tito, H., & Centty, D. (2019). Model for time series imputation based on average of historical vectors, fitting and smoothing. *International Journal of Advanced Computer Science and Applications*, 10(10), 346–352. <https://doi.org/10.14569/ijacsa.2019.0101049>
- Flores, A., Tito, H., & Silva, C. (2019). Local average of nearest neighbors: Univariate time series imputation. *International Journal of Advanced Computer Science and Applications*, 10(8), 45–50. <https://doi.org/10.14569/ijacsa.2019.0100807>

- Golyandina, N., Korobeynikov, A., & Zhigljavsky, A. (2018). Singular Spectrum Analysis with R. In *Springer* (1st ed.). <https://doi.org/10.1007/978-3-662-57380-8>
- Hallegatte, S., Green, C., Nicholls, R. J., & Corfee-Morlot, J. (2013). Future flood losses in major coastal cities. *Nature Climate Change*, 3(9), 802–806. <https://doi.org/10.1038/nclimate1979>
- Hawthorne, G., & Elliott, P. (2005). Imputing cross-sectional missing data: Comparison of common techniques. *Australian and New Zealand Journal of Psychiatry*, 39(7), 583–590. <https://doi.org/10.1111/j.1440-1614.2005.01630.x>
- Ivanov, A., Georgiev, I., & Dimitrov, N. (2019). Single spectrum analysis for monthly sea level data from Varna tide gauge station. *10th Congress of Balkan Geophysical Society, BGS 2019, September*. <https://doi.org/10.3997/2214-4609.201902664>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer. <https://doi.org/10.1007/978-1-4614-7138-7>
- Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, 38(18), 2895–2907. <https://doi.org/10.1016/j.atmosenv.2004.02.026>
- Khelifa, S., Gourine, B., Rami, A., & Taibi, H. (2016). Assessment of nonlinear trends and seasonal variations in global sea level using singular spectrum analysis and wavelet multiresolution analysis. *Arabian Journal of Geosciences*, 9(10), 1–8. <https://doi.org/10.1007/s12517-016-2584-6>
- Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2004). *Applied Linear Statistical Models*. McGraw-Hill/Irwin.
- Ministerio del Ambiente del Ecuador (MAE). (2019). Primera Contribución Determinada a nivel nacional para el Acuerdo de París bajo la Convención Marco de Naciones Unidas sobre Cambio Climático. *Gobierno de Ecuador*, 1–44.
- Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., & Stork, J. (2015). *Comparison of different Methods for Univariate Time Series Imputation in R*. <http://arxiv.org/abs/1510.03924>
- NCEI. (n.d.). *El Niño/Southern Oscillation (ENSO)*. Retrieved July 26, 2021, from <https://www.ncdc.noaa.gov/teleconnections/enso/>
- Noor, N. M., Al Bakri Abdullah, M. M., Yahaya, A. S., & Ramli, N. A. (2015). Comparison of linear interpolation method and mean method to replace the missing values in environmental data set. *Materials Science Forum*, 803, 278–281. <https://doi.org/10.4028/www.scientific.net/MSF.803.278>

- Oppenheimer, M., Glavovic, B. C., Hinkel, J., van de Wal, R., Magnan, A. K., Abd-Elgawad, A., Cai, R., Cifuentes-Jara, M., DeConto, R. M., Ghosh, T., Hay, J., Isla, F., Marzeion, B., Meyssignac, B., & Sebesvari, Z. (2019). Sea Level Rise and Implications for Low-Lying Islands, Coasts and Communities. *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*, 321–445. <https://www.ipcc.ch/srocc/download/>
- Organización Panamericana de la Salud. (2000). *Fenómeno El Niño, 1997-1998* (Issue 8).
- Ozsoy, O., Haigh, I. D., Wadey, M. P., Nicholls, R. J., & Wells, N. C. (2016). High-frequency sea level variations and implications for coastal flooding: A case study of the Solent, UK. *Continental Shelf Research*, 122, 1–13. <https://doi.org/10.1016/j.csr.2016.03.021>
- Phan, T. T. H. (2020). Machine Learning for Univariate Time Series Imputation. *2020 International Conference on Multimedia Analysis and Pattern Recognition, MAPR 2020*, 1–6. <https://doi.org/10.1109/MAPR49794.2020.9237768>
- Phan, T. T. H., Poisson Caillault, É., & Bigand, A. (2020). eDTWBI: Effective Imputation Method for Univariate Time Series. *Advances in Intelligent Systems and Computing, 1121 AISC*, 121–132. https://doi.org/10.1007/978-3-030-38364-0_11
- Phan, T. T. H., Poisson Caillault, É., Lefebvre, A., & Bigand, A. (2017). Dynamic time warping-based imputation for univariate time series data. *Pattern Recognition Letters*, 139, 139–147. <https://doi.org/10.1016/j.patrec.2017.08.019>
- Telgársky, R. (2013). Dominant Frequency Extraction. *Eprint ArXiv:1306.0103*, 1–12. <http://arxiv.org/abs/1306.0103>
- UNESCO/IOC. (1985). Manual on Sea Level Measurement and Interpretation. *Intergovernmental Oceanographic Commission Manual and Guides No. 14, I*, 83. https://www.psmsl.org/train_and_info/training/manuals/ioc_14i.pdf
- UNESCO/IOC. (2002). Manual on Sea Level Measurement and Interpretation. *Intergovernmental Oceanographic Commission Manual and Guides No. 14, III*, 55. <https://unesdoc.unesco.org/ark:/48223/pf0000125129>
- UNESCO/IOC. (2016). Manual on Sea Level Measurement and Interpretation: Radar Gauges. *Intergovernmental Oceanographic Commission Manual and Guides No. 14, V*, 104. <https://unesdoc.unesco.org/ark:/48223/pf0000246981>
- University of Hawaii Sea Level Center. (n.d.-a). *The science of sea level change*. Retrieved June 7, 2021, from <https://uhslc.soest.hawaii.edu/research/>
- University of Hawaii Sea Level Center. (n.d.-b). *What datasets are available?*. Retrieved

June 7, 2021, from <https://uhslc.soest.hawaii.edu/datainfo/>

Wong, Z. (2011). *Relación entre las oscilaciones del nivel del mar del Océano Pacífico y las variaciones del nivel del mar en la costa del Ecuador*. Escuela Superior Politécnica del Litoral.

APÉNDICES

APÉNDICE A

Gaps en la serie de tiempo nivel del mar horario de La Libertad

Fuente: Elaboración propia

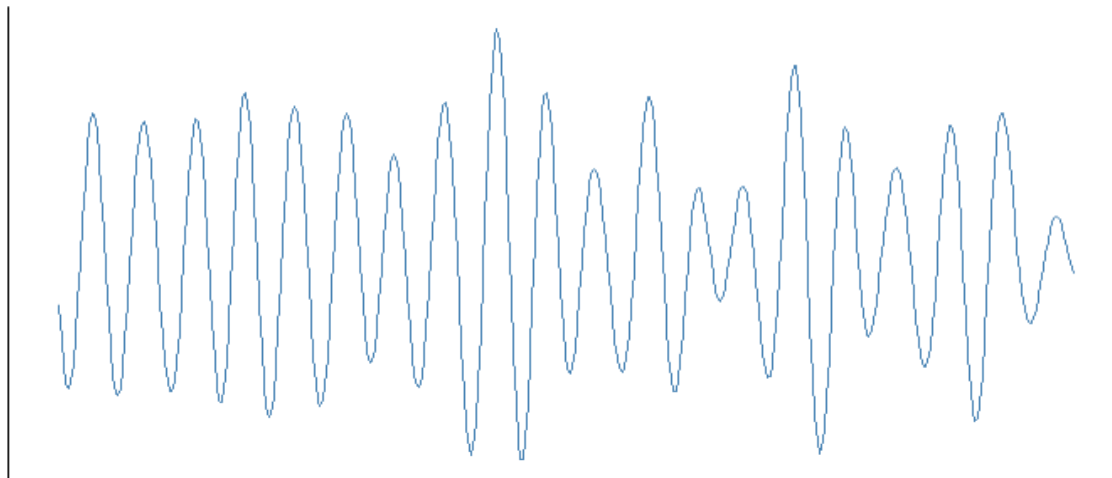
Gap	Tamaño	Inicio	Final
1	168	1950-03-23 05:00:00	1950-03-30 04:00:00
2	41	1953-07-31 05:00:00	1953-08-01 21:00:00
3	165	1953-08-07 18:00:00	1953-08-14 14:00:00
4	152	1956-06-01 05:00:00	1956-06-07 12:00:00
5	37	1956-06-16 09:00:00	1956-06-17 21:00:00
6	34	1956-06-24 13:00:00	1956-06-25 22:00:00
7	1068	1956-06-29 12:00:00	1956-08-12 23:00:00
8	1224	1956-11-11 05:00:00	1957-01-01 04:00:00
9	25	1957-03-24 16:00:00	1957-03-25 16:00:00
10	1063	1957-06-08 17:00:00	1957-07-22 23:00:00
11	2496	1962-02-19 00:00:00	1962-06-02 23:00:00
12	734	1963-12-01 15:00:00	1964-01-01 04:00:00
13	373	1970-05-13 11:00:00	1970-05-28 23:00:00

14	75	1970-08-18 02:00:00	1970-08-21 04:00:00
15	37	1970-12-25 04:00:00	1970-12-26 16:00:00
16	263	1971-02-18 17:00:00	1971-03-01 15:00:00
17	48	1971-03-27 05:00:00	1971-03-29 04:00:00
18	744	1971-10-01 05:00:00	1971-11-01 04:00:00
19	42	1971-11-18 01:00:00	1971-11-19 18:00:00
20	38	1971-12-01 05:00:00	1971-12-02 18:00:00
21	32	1972-02-23 21:00:00	1972-02-25 04:00:00
22	27	1972-03-19 21:00:00	1972-03-20 23:00:00
23	485	1972-06-27 01:00:00	1972-07-17 05:00:00
24	32	1972-08-29 21:00:00	1972-08-31 04:00:00
25	64	1974-08-20 21:00:00	1974-08-23 12:00:00
26	1508	1977-10-01 05:00:00	1977-12-03 00:00:00
27	31	1978-01-10 16:00:00	1978-01-11 22:00:00
28	156	1978-10-31 07:00:00	1978-11-06 18:00:00
29	78	1995-01-26 11:00:00	1995-01-29 16:00:00
30	27	1995-03-14 17:00:00	1995-03-15 19:00:00
31	72	1998-10-15 05:00:00	1998-10-18 04:00:00
32	26	1998-10-23 04:00:00	1998-10-24 05:00:00

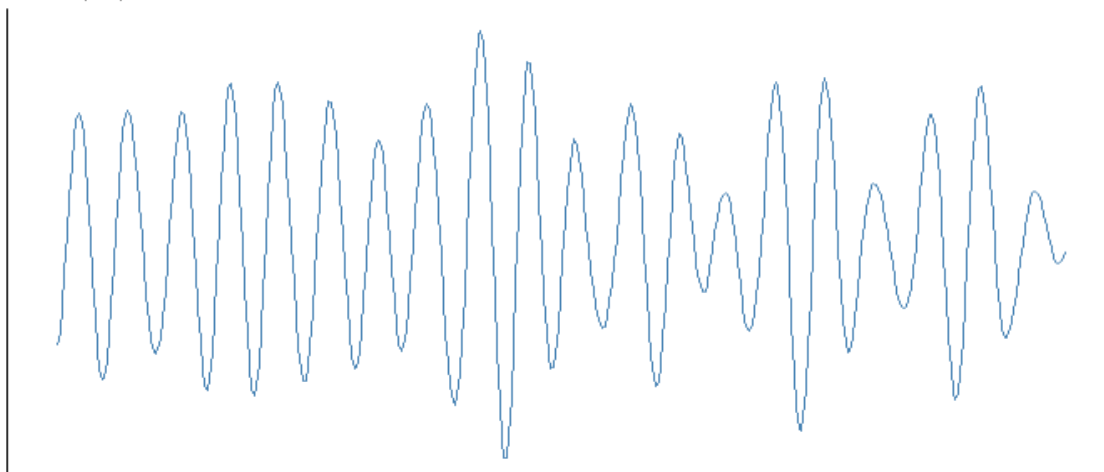
33	26	1998-10-31 03:00:00	1998-11-01 04:00:00
34	216	1999-10-01 05:00:00	1999-10-10 04:00:00
35	517	1999-10-11 05:00:00	1999-11-01 17:00:00
36	25	2000-02-16 04:00:00	2000-02-17 04:00:00
37	480	2000-03-03 05:00:00	2000-03-23 04:00:00
38	25	2000-04-30 05:00:00	2000-05-01 05:00:00
39	120	2000-06-07 05:00:00	2000-06-12 04:00:00
40	384	2000-06-14 05:00:00	2000-06-30 04:00:00
41	29	2000-08-16 00:00:00	2000-08-17 04:00:00
42	144	2001-08-17 05:00:00	2001-08-23 04:00:00
43	1464	2003-11-01 05:00:00	2004-01-01 04:00:00
44	1682	2004-05-14 03:00:00	2004-07-23 04:00:00
45	46	2004-09-29 07:00:00	2004-10-01 04:00:00
46	1438	2005-11-02 07:00:00	2006-01-01 04:00:00
47	40	2008-12-27 13:00:00	2008-12-29 04:00:00
48	234	2011-10-01 22:00:00	2011-10-11 15:00:00
49	91	2011-10-14 22:00:00	2011-10-18 16:00:00
50	721	2016-04-01 00:00:00	2016-05-01 00:00:00

APÉNDICE B

Vector propio 13



Vector propio 14



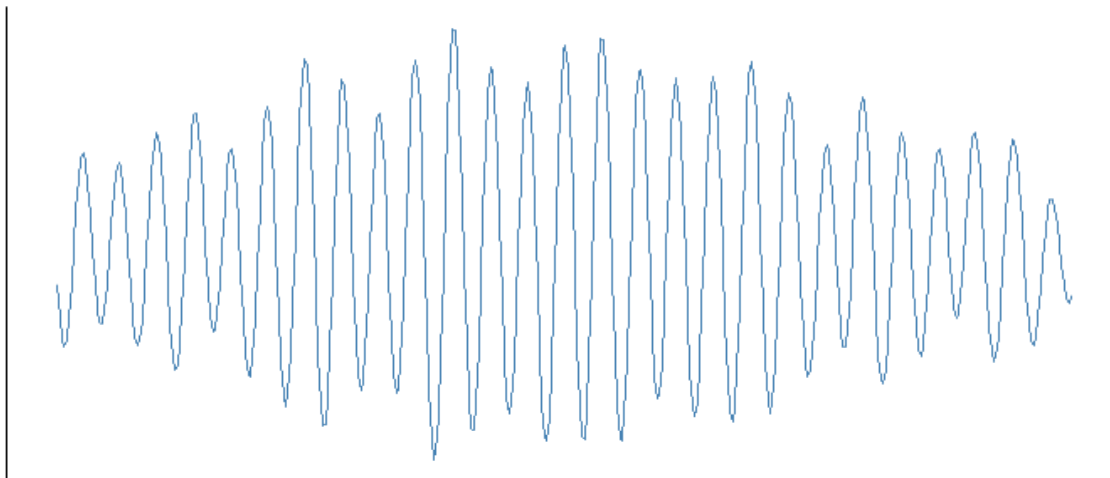
Vectores propios correspondientes al componente periódico 6 para el SSA con

L=432

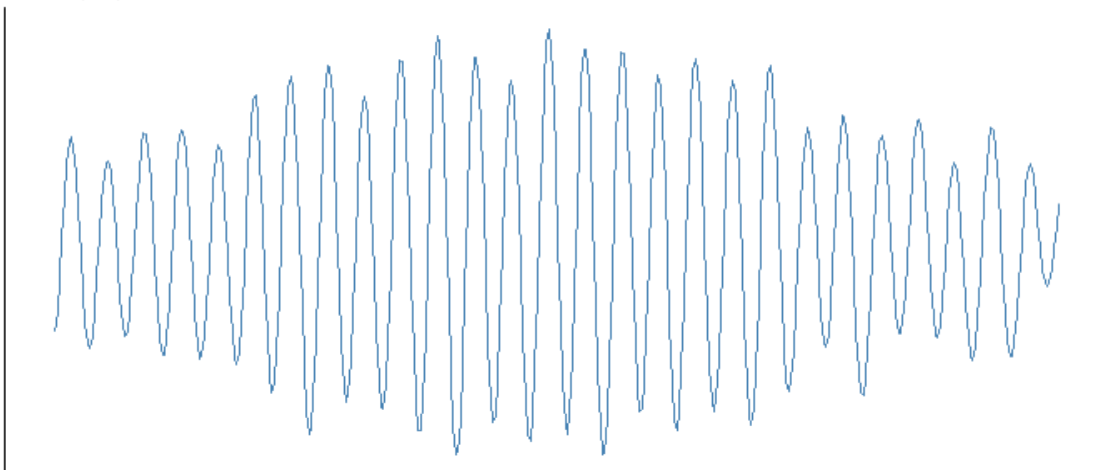
Fuente: Elaboración propia

APÉNDICE C

Vector propio 17



Vector propio 18

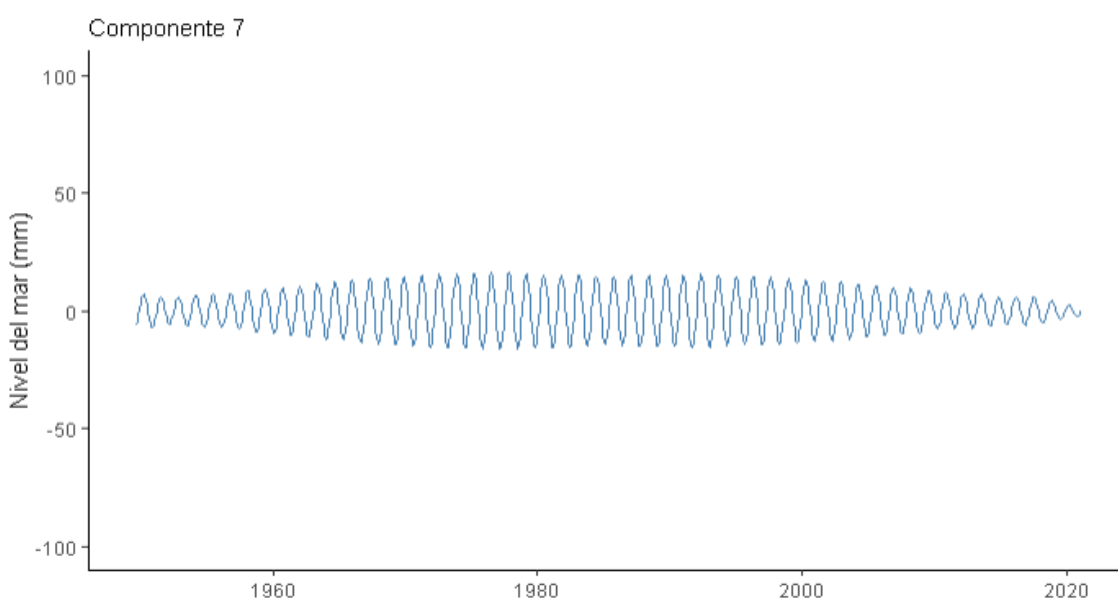
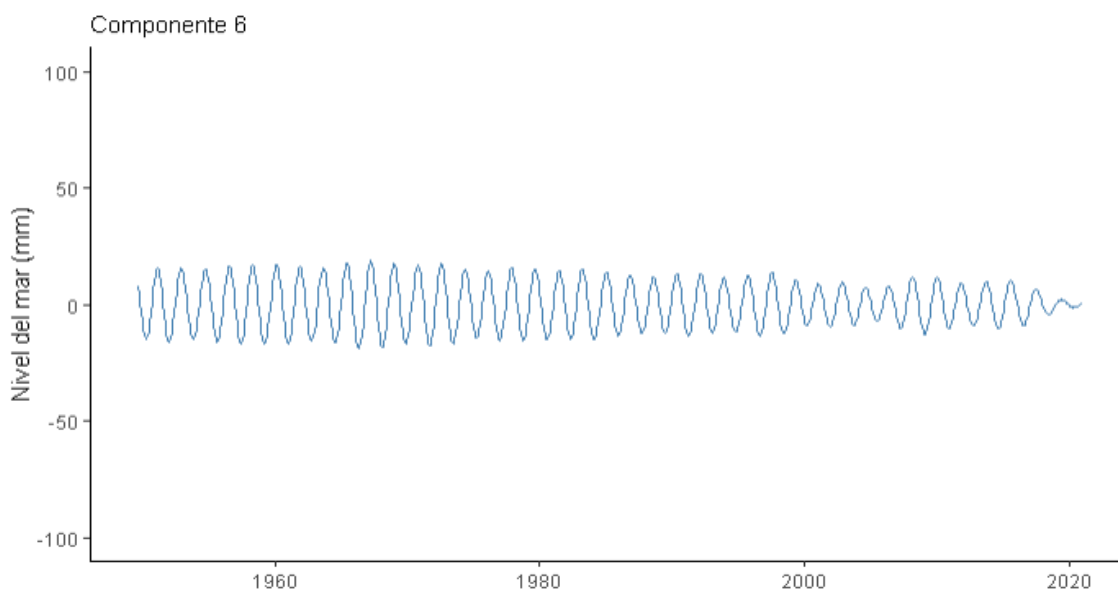


Vectores propios correspondientes al componente periódico 7 para el SSA con

$L=432$

Fuente: Elaboración propia

APÉNDICE D



Componentes periódicos 6 y 7 estimados para la serie anomalías del nivel medio del mar mensual

Fuente: Elaboración propia