Universidad Nacional de Ingeniería

Facultad de Ingeniería Económica,

Estadística y CC.SS.

"Enfoque híbrido de series temporales SARIMA-MLP con corrección PCA para la mejora de pronósticos aplicado a la producción de plata en el Perú. Periodo 2001-2023."

Maria del Cielo Lozada Pérez

Investigación Estadística

Docente: Demetrio Antonio Ruiz Olorte

Lima, 16 de Diciembre de 2023

Resumen

Este estudio se centra en la mejora de los pronósticos de la serie de producción de plata durante el periodo comprendido entre 2001 y 2023 mediante la implementación de un enfoque híbrido SARIMA-MLP con corrección PCA. Inicialmente, se aborda la corrección del quiebre estructural de la serie en el año 2020, considerando la sensibilidad del modelo SARIMA a datos atípicos y con la ayuda de los componentes principales se puede mitigar este efecto. Posteriormente, se ajusta un modelo SARIMA$(1,1,1)(2,0,0)$ a los datos de entrenamiento con el objetivo de capturar patrones lineales, y para garantizar una representación adecuada de la estructura de los datos, se aplica un modelo de aprendizaje profundo MLP a los residuos. Los resultados revelan que el modelo híbrido corregido exhibe un menor error de pronóstico, medido a través del MAPE, en comparación con el SARIMA corregido. **Palabras clave**: Enfoque híbrido, SARIMA, MLP, PCA.

Abstract

This study focuses on improving the forecasts of the silver production series over the period 2001 to 2023 by implementing a hybrid SARIMA-MLP approach with PCA correction. Initially, the correction of the structural break of the series in the year 2020 is addressed, considering the sensitivity of the SARIMA model to outlier data and with the help of principal components this effect can be mitigated. Subsequently, a SARIMA$(1,1,1)(2,0,0)$ model is fitted to the training data with the objective of capturing linear patterns, and to ensure an adequate representation of the data structure, an MLP deep learning model is applied to the residuals. The results reveal that the corrected hybrid model exhibits a lower forecast error, as measured through the MLP, compared to the corrected SARIMA. **Key words**: Hybrid approach, SARIMA, MLP, PCA.

Índice general

Capítulo 1

Marco referencial

1.1. Introducción

Los modelos SARIMA son una extensión de los modelos ARIMA que permiten modelar y predecir series de tiempo con comportamiento estacional, lo que es común en muchos campos, como la economía, la meteorología y la industria. Sin embargo, usar únicamente este modelo, en algunos casos, puede no tener capacidad efectiva de hacer pronósticos precisos y confiables. Una primera mejora a este problema sería la corrección de periodos anómalos dentro de la serie de tiempo, ya que, sabemos que una serie con datos atípicos puede afectar el desarrollo del modelo. Como segunda mejora podemos usar, adicionalmente, modelos de machine learning para mejorar la contribución del ruido blanco a la precisión de los pronósticos, como se hizo en un estudio Ruiz-Aguilar et al., 2014, donde se propone una metodología de hibridación basada en la integración de los datos obtenidos del modelo de promedios móviles integrados autorregresivos (SARIMA) en el modelo de red neuronal artificial (ANN) para predecir el número de inspecciones. Así como este trabajo hay otras investigaciones que optan por usar modelos híbridos, ya

que, al usar estos modelos, se busca que los residuos se comporten como ruido blanco, lo que sugiere que el modelo es una buena representación de los datos.

Para este trabajo se usarán estas dos mejoras aplicadas a la serie producción de plata en el Perú (Periodo 2001-2023), donde claramente el periodo anómalo es el año 2020 (época de pandemia), este periodo será corregido usando componentes principales. Y se usará adicionalmente el modelo Red Neuronal Multicapa (MLP), para capturar los patrones no lineales que el modelo SARIMA por si solo no lo puede hacer.

1.2. Descripción del problema

Existen muchas series de tiempo que presentan quiebres estructurales durante un periodo para luego retornar a su estado natural. Este quiebre estructural puede ocasionar predicciones erróneas si no se toma en cuenta. Es por ello que es necesario realizar una corrección para ese periodo y solucionar este problema. Al hacer esta corrección se puede mejorar la precisión en los pronósticos, sin embrago, esto no garantiza que, al usar el modelo, en este caso, SARIMA obtengamos errores aleatorios (Ruido blanco), es decir, el modelo puede que no esté capturando adecuadamente la estructura de los datos y por ende no dar pronósticos adecuados.

Como alternativa de solución a esta problemática se plantea un enfoque híbrido SARIMA-MLP, donde consideraremos que la parte que no esta capturando el modelo SARIMA puede ser capturada por el modelo MLP, es decir, SARIMA solo captura los patrones lineales, mientras que MLP estaría capturando los patrones no lineales.

1.3. Formulación del problema

Ya habiendo entendido por qué es importante tanto corregir quiebres estructurales como obtener errores aleatorios en el modelo podemos formular la siguiente interrogante: ¿Conseguiremos mejores pronósticos haciendo una corrección a la serie producción de plata en el Perú (2001-2023) por medio de PCA para luego ajustarla bajo un enfoque híbrido SARIMA-MLP? Esta interrogante será resuelta a lo largo del trabajo.

1.4. Objetivos de la investigación

1.4.1. Objetivo general

Mejorar los pronósticos de la producción de plata del Perú (2001-2023) haciendo uso del enfoque híbrido de series temporales SARIMA-MLP con corrección por PCA.

1.4.2. Objetivos específicos

- Corregir el periodo 2020 de la serie producción de plata en el Perú (2001-2023) usando PCA.
- Implementar el enfoque híbrido SARIMA-MLP en la serie producción de plata en el Perú (2001-2023).
- Comparar el enfoque propuesto con el modelo SARIMA.

1.5. Justificación de la investigación

Los modelos SARIMA tienen limitaciones, una de ellas es que no siempre es adecuado cuando para capturar patrones no lineales de la serie y otra es que no puede ser eficaz

en la captura de cambios abruptos en la serie. Es por ello que en esta investigación tiene el propósito de proponer una alternativa tanto en la corrección de quiebres estructurales, así como una posible solución cuando se obtengan errores no aleatorios. Los quiebres estructurales pueden corregirse usando medias móviles comúnmente, sin embargo, no son la mejor herramienta para corregir estos quiebres. Usar PCA puede ser una solución. Por otro lado, usar un modelo hibrido ya con la serie corregida puede ayudar a que todos los patrones de los residuos sean capturados por el modelo. De acuerdo con la investigación de Xu et al., 2019 donde se propone un modelo SARIMA-MLP para pronosticar indicadores estadísticos en la industria de la aviación se concluye que uno de los modelos híbridos usados puede lograr una mayor precisión que otros métodos y demuestra que la incorporación de ruido blanco gaussiano puede aumentar la precisión de los pronósticos. Si corregimos ambos problemas en la serie podemos obtener tanto un modelo que haga adecuadas predicciones como capturar adecuadamente la estructura de los datos.

1.6. Alcances de la investigación

Esta investigación se enfocará en la implementación del modelo hibrido SARIMA-MLP con la finalidad de que el modelo sea considerado eficiente, además de usar componentes principales como método alternativo para la corrección de quiebres estructurales en una serie de tiempo. Finalmente se evaluará la eficacia de este modelo comparándolo con modelos SARIMA y suavización exponencial. La investigación usará datos proporcionados por el BCRP de la serie producción de plata en el Perú (Periodo 2001-2023).

1.7. Limitaciones de la investigación

- Esta investigación se limita a usar el modelo hibrido SARIMA-MLP, descartándose otros modelos que pueden ayudar capturar tanto los patrones lineales como no lineales.
- Este estudio no hace uso de variables externas, que, en ocasiones, puede influir en el estudio.
- Este estudio solo hace uso de un método para corregir el quiebre estructural en una serie en particular, limitando el uso de otros métodos que podrían resolver el mismo problema.

Capítulo 2

Antecedentes de la investigación

2.0.1. Antecedentes Nacionales

- Mercado, 2021

  Se han construido modelos híbridos ANN-ARIMA mediante remodelación, para realizar los pronósticos de los nuevos casos de contagios por Covid-19 en el Perú, para ello se extrajeron y utilizaron los casos confirmados de Covid-19 entre el periodo 03/06/20 hasta 28/02/21, desde la plataforma de datos abiertos del Ministerio de Salud. Los resultados encontrados indican que los 02 mejores modelos corresponden al modelo híbrido multiplicativo NNAR (27, 1, 6) * ARIMA (3, 0, 2) (1, 0, 1), y al modelo híbrido aditivo NNAR (27, 1, 6) + ARIMA (1, 0, 1), cuyos valores del error porcentual absoluto medio (MAPE) difieren sólo en un 0,575 %, proporcionando así casi las mismas previsiones. Considerando el promedio de los valores MAPE para los 03 mejores modelos de cada categoría de modelado, se ha determinado que los modelos híbridos NNAR-ARIMA son mejores que los modelos híbridos MLP-ARIMA, que los modelos híbridos aditivos NNAR + ARIMA tienen una superioridad de 1,20 % en los modelos híbridos multiplicativos NNAR * ARI-

MA; mientras que la superioridad del modelo híbrido aditivo MLP+ARIMA sobre el modelo híbrido multiplicativo MLP*ARIMA alcanza el 2,31 %.

- Amao Suxo, 2018

  En esta tesis se realiza una metodología de resumir-predecir-y-reconstruir, el método usado en esta tesis sugiere reducir la dimensionalidad de los datos usando un análisis de componentes principales para luego realizar pronósticos individuales sobre las componentes o auto vectores más significativos. Finalmente, un algoritmo recursivo, aplicado sobre la reconstrucción inversa espectral de los pronósticos individuales, brinda el pronóstico final de los mapas. Esta metodología es usada en esta investigación con el fin de corregir el quiebre estructural que presenta la serie de tiempo producción de plata en el Perú (Periodo 2001-2023).

2.0.2. Antecedentes Internacionales

- Ince y Trafalis, 2006

  En este estudio se propuso un modelo de pronóstico de dos etapas que incorpora técnicas paramétricas como la media móvil autorregresiva integrada (ARIMA), la vector autorregresiva (VAR) y técnicas de cointegración, y técnicas no paramétricas como la regresión de vectores de soporte (SVR) y las redes neuronales artificiales (ANN). La comparación de estos modelos mostró que la selección de insumos es muy importante. Además, nuestros hallazgos muestran que la técnica SVR supera a la ANN en dos métodos de selección de entrada.

- Xu et al., 2019

  En este estudio, se propone un novedoso modelo SARIMA-SVR para pronosticar indicadores estadísticos en la industria de la aviación que pueden usarse para fines

posteriores de gestión y planificación de la capacidad. Primero, SARIMA analiza la serie temporal. Luego, el ruido blanco gaussiano se calcula a la inversa. A continuación, se proponen y aplican cuatro modelos híbridos para pronosticar los indicadores estadísticos futuros en la industria de la aviación. Los resultados del estudio empírico sugieren que uno de los modelos propuestos, a saber, SARIMA-SVR3, puede lograr una mayor precisión que otros métodos y demuestra que la incorporación de ruido blanco gaussiano puede aumentar la precisión de los pronósticos.

- Sierra, Rodríguez et al., 2019

  El objetivo de esta investigación fue la construcción de modelos ARIMA-GARCH y ARIMAX-GARCH como herramienta para el pronóstico de la tasa de cambio en Colombia a partir de los retornos diarios de los precios de cierre USD/COP y su análisis de correlación dinámica con algunas variables de interés. Los resultados sugieren que la incorporación de variables exógenas significativas dentro de la modelación ARIMAX-GARCH con correlación persistente según el modelo DCC (por sus siglas en inglés Dinamic Conditional Correlation) al par USD/COP genera pronósticos fuera de muestra con mejor desempeño que los modelos univariados ARIMA-GARCH.

- Ruiz-Aguilar et al., 2014

  En este documento, el número de mercancías sujetas a inspección en el Puesto Europeo de Inspección Fronteriza se predice mediante un procedimiento híbrido de dos pasos. Se propone una metodología de hibridación basada en la integración de los datos obtenidos del modelo de promedios móviles integrados autorregresivos (SARIMA) en el modelo de red neuronal artificial (ANN) para predecir el número de inspecciones. Se comparan varios enfoques híbridos y los resultados indican

que los modelos híbridos superan a cualquiera de los modelos utilizados por separado. Esta metodología puede convertirse en una poderosa herramienta para la toma de decisiones en otras instalaciones de inspección de puertos o aeropuertos internacionales.

- Berberich, 2020

  Esta tesis muestra que la combinación de un método de Holt-Winters y una red de memoria a largo plazo es prometedora cuando la periodicidad de una serie de tiempo se puede especificar con precisión. La especificación precisa permite que el método Holt-Winter simplifique la tarea de pronóstico para la red de memoria a corto plazo y, en consecuencia, facilita que el método híbrido obtenga pronósticos precisos. La pregunta de investigación que debe responderse es qué características de una serie temporal determinan la superioridad de los enfoques estadísticos, de aprendizaje automático o híbridos. El resultado del experimento realizado muestra que esta pregunta de investigación no puede responderse de manera general. Sin embargo, los resultados proponen hallazgos para métodos de pronóstico específicos. El método de Holt-Winters proporciona pronósticos confiables cuando la periodicidad se puede determinar con precisión. ARIMA, sin embargo, maneja mejor las sonalidades estacionales superpuestas que el método Holt-Winters debido a su enfoque autorregresivo. Además, los resultados sugieren la hipótesis de que los métodos de aprendizaje automático tienen dificultades para extrapolar series de tiempo con tendencia. Finalmente, el perceptrón multicapa puede realizar pronósticos precisos para varias series temporales a pesar de su simplicidad, y la red de memoria a corto plazo demuestra que necesita conjuntos de datos relevantes de longitud adecuada para realizar pronósticos precisos.

Capítulo 3

Base teórica

3.1. Serie de tiempo

Es una sucesión de n realizaciones ordenadas y equidistantes cronológicamente sobre una características de interés o sobre varias características.

3.1.1. Tipos de variaciones

Presenta diversos tipos de variaciones, para este análisis solo hablaremos de 3:

1. Tendencia T_t: Esto puede definirse vagamente como "cambio a largo plazo en el nivel medio".citar Es decir, la tendencia existe cuando, a largo plazo, el nivel medio de la serie varía.

2. Estación E_t: Variaciones periódicas a corto plazo con periodos regulares y alrededor de la media.

3. Irregular a_t: Variaciones aleatorias, no tienen un patrón de variación.

3.1.2. Procesos autorregresivos

Según Peña Sánchez de Rivera, 2005 diremos que una serie temporal z_t sigue un proceso autorregresivo de orden p si:

$$\tilde{z}_t = \phi_1 \tilde{z_{t-1}} + \cdots + \phi_p \tilde{z_{t-p}} + a_t$$

Donde $\tilde{z_{t-1}} = z_t - \mu$, siendo μ la media del proceso z_t y a_t un proceso de ruido blanco. Utilizando la notación de operadores, la ecuación de un AR(p) es:

$$(1 - \phi B - \cdots - \phi_p B^p)\tilde{z}_t = a_t$$

Y llamado $\phi(B) = 1 - \phi B - \cdots - \phi_p B^p$ al polinomio de grado p en el operador de retardo, cuyo primer término es la unidad.Sean $G_1^{-1}, \ldots, G_p^{-1}$ las raíces de la función $\phi(B) = 0$, este proceso será estacionario si $|G_i| < 1$, para todo i.

3.1.3. Medias moviles

Segun Peña Sánchez de Rivera, 2005, diremos que una serie temporal sigue un proceso de medias moviles si MA(q) si:

$$\tilde{z}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$$

Introduciendo la notación de operadores:

$$\tilde{z}_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)a_t$$

Se puede escribirse de forma más compacta como: $\tilde{z}_t = \theta_q(B)a_t$. Un proceso MA($q$) es

siempre estacionario, por ser la suma de procesos estacionarios. Diremos que el proceso es invertible si las raices del operador $\theta_q(B) = 0$ son, en módulo, mayores que la unidad.

3.1.4. Procesos ARMA

Según Peña Sánchez de Rivera, 2005 el proceso ARMA(p, q) se define como:

$$(1 - \phi B - \cdots - \phi_p B^p)\tilde{z}_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)a_t$$

o en notación compacta:

$$\phi_q(B)\tilde{z}_t = \theta_q(B)a_t$$

En proceso será estacionario si las raíces de $\phi_q(B) = 0$ estan fuera del círculo de la unidad, e invertible si lo están de las $\theta_q(B)$

3.1.5. Procesos integrados ARIMA

Es frecuente que los procesos en las series no tengan un compotamiento homogéneo a lo largo del tiempo, además lo procesos pueden no mostrar estacionariedad en la pendiente. Para correguir ello se aplican diferencias hasta que la serie sea estacionaria. Entonces llamaremos a z, un proceso de media móvil integrada autorregresiva(ARIMA) de órdenes p, d y q, es decir ARIMA(p, d, q) donde d es la diferencia máxima que produce la que serie sea estacionaria. Boshnakov, 2016

. Un ARIMA(p, d, q) puede escribirse de la siguiente manera

$$\Phi(B)(1 - B)^d z_t = \Theta(B)a_t$$

3.1.6. Modelo ARIMA estacional

Según Peña Sánchez de Rivera, 2005 cuando una serie presenta un componente periódico igual o inferior a un año es necesario añadir el componente estacional al modelo con el fin de captar su verdadero movimiento.

Box et al., 2015 proponen el modelo estacional multiplicativos denotado como SARIMA$(p,d,q)(P,D,Q)_s$ y es representado por:

$$\phi_p(B)\Phi_P(B^s)\Delta^d\Delta_s^D z_t = \Theta_Q(B^s)\theta_q(B)a_t$$

Donde:

- $\phi_p(B)$,es el polinomio autorregresivo de orden p.
- $\Phi_P(B^s)$, es el polinomio autorregresivo estacional de orden P.
- $\Delta^d = (1-B)^d$, es el operador diferencia y d es el número de diferencias necesarias para eliminar la tendencia de la serie.
- $\Delta^D = (1-B)^D$, es el operador diferencia generalizado, cuando dos observaciones están distantes entre sí de s intervalos de tiempo que presentan alguna semejanza y D es el número de diferencias de rezagos s necesarias para eliminar la estacionalidad de la serie.
- $\theta_q(B)$, es el polinomio de medias móviles de orden q.
- $\Theta_Q(B^s)$, es el polinomio de medias móviles estacional de orden Q.

3.2. Perceptrón multicapa (MLP)

Es un algoritmo de aprendizaje supervisado que aprende una función $f(.) : R^m \rightarrow R^o$ mediante entrenamiento en un conjunto de datos, donde m es el número de dimensiones para la entrada y o es el número de dimensiones para la salida. Dado un conjunto de características $X = x_1, x_2, x_3, \ldots, x_m$ y el target y puede aprender un aproximador de funciones no lineales para clasificación o regresión. Se diferencia de la regresión logística en que entre la capa de entrada y la de salida puede haber una o más capas no lineales, llamadas capas ocultas. La Figura muestra un MLP de una capa oculta con salida escalar.

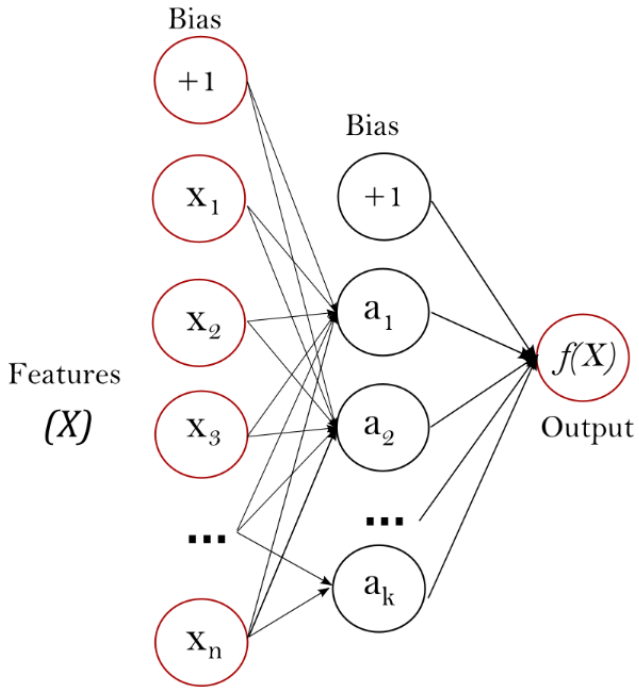

Figura 3.1: MLP de una capa oculta.

La capa más a la izquierda, conocida como capa de entrada, consta de un conjunto de neuronas $x_i|x_1, x_2, \ldots, x_m$ que representan las características de entrada. Cada neurona de la capa oculta transforma los valores de la capa anterior con una suma lineal ponderada $w_1x_1 + w_2x_2 + \cdots + w_mx_m$ seguido de una función de activación no lineal $g(.) : R \rightarrow R$ como la función tan hiperbólica. La capa de salida recibe los valores de la última capa

oculta y los transforma en valores de salida. Pedregosa et al., 2011

3.3. Análisis de componentes principales

Este método encuentra la manera de reducir la dimensionalidad de la data y que esta contenga la mayor variación total. La idea es que cada una de las n observaciones que viven en un espacio p-dimensional sea reducido. ACP busca el menor número de dimensiones donde el concepto de interés varíe de acuerdo al grado de variabilidad que cada dimensión me aporte.

La primera componente del conjunto de datos $X_1, X_2, \ldots, X_p$ es la siguiente combinación lineal normalizada:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

Y así sucesivamente hasta obtener la misma cantidad de componentes que de variables. Donde las primeras componentes son las que presentan mayor varianza, estas son representadas por los autovalores de la matriz X ordenadas de forma descendente.

3.4. Estructura híbrida en paralelo

Para esta estructura inicialmente se debe definir el número de modelos k que se desea ajustar en serie. En particular, cuando $k = 2$ se trata de capturar en cada uno de los modelos el comportamiento lineal L_t y no lineal N_t de una serie temporal respectivamente. En la literatura no está establecido el orden de la secuencia para modelar y pronosticar dichos patrones, sin embargo Ballesteros López et al., 2023 menciona que Zhang (2003) y Zeng et al. (2008) establecen que bajo esta configuración el orden que se debe tener

en cuenta para aprovechar las capacidades predictivas de los modelos de interés es: (1) lineal- no lineal y (2) no lineal- lineal. El primero consiste en capturar las características lineales que presentan los datos de estudio, a través de un modelo de pronóstico f_1. De esta manera, los residuos resultantes $e_{t+h,1} = f(e_{t-1}, \ldots, e_{t-n})$ contiene las relaciones no lineales presentes en los datos y estos se utilizan como entrada para un segundo modelo de pronóstico f_2. Así, pronóstico final está dado por la suma de los pronósticos individuales como se sigue:

$$\hat{f}_{\text{combinado}} = \hat{f}_{t+h,1} + \hat{f}_{t+h,2}$$

Donde $\hat{f}_{t+h,1}$ es el pronóstico de la parte final y $\hat{f}_{t+h,2}$ el pronóstico no lineal.

3.5. Marco conceptual

3.5.1. Producción de plata en el Perú

Perú, conocido como el tercer productor mundial de plata, ha experimentado una ligera disminución en su producción de este preciado metal en mayo de 2023. Según datos proporcionados por el Ministerio de Energía y Minas de Perú (MINEM), la producción de plata fue de 242,527 kg (7,8 moz), lo que significa un descenso del $4,7\,\%$ en comparación con los 254,493 kg (8,2 moz) extraídos en el mismo mes del año anterior.

El MINEM atribuye este desempeño negativo principalmente a la disminución de la producción reportada por Compañía Minera Antamina S.A. ($-29{,}4\,\%$) y Compañía Minera Ares ($-24{,}1\,\%$).

La disminución en la producción de plata de Perú destaca la vulnerabilidad del sector minero a fluctuaciones y cambios en el rendimiento de las empresas productoras. A pesar de esta disminución, Perú sigue siendo un jugador clave en la producción mundial de

plata.

Es esencial seguir de cerca el desempeño de las compañías mineras y las regiones productoras para entender cómo pueden influir en las cifras nacionales y mundiales. Además, se deben explorar más estrategias para mitigar los posibles impactos de las disminuciones en la producción en el futuro. MineriaenLinea, Julio-2023

Capítulo 4

Metodología

4.1. Hipótesis

4.1.1. Hipótesis General

Se obtendrá mejores pronósticos empleando el enfoque híbrido de series temporales SARIMA-MLP con corrección por PCA.

4.1.2. Hipótesis Específicas

- El análisis de componentes principales corrige y mejora la predicción en la serie de tiempo.
- Los residuos, aplicando el modelo híbrido SARIMA-MLP, se comportan como ruido blanco.

4.2. Tipo, Nivel y Diseño de la investigación

4.2.1. Tipo de investigación

Esta investigación se ubica en el contexto del enfoque aplicado. Como indica Lozada, 2014 La investigación aplicada busca la generación de conocimiento con aplicación directa a los problemas de la sociedad o el sector productivo. Bajo este concepto, esta investigación busca mejorar el pronóstico de la serie producción de plata del Perú (2001-2023) haciendo uso del enfoque híbrido de series temporales SARIMA-MLP con corrección por PCA.

4.2.2. Nivel de investigación

El nivel de investigación es de tipo explicativa, ya que se aplican técnicas cuantitativas basadas en el análisis de datos históricos para predecir el comportamiento futuro de una variable.

En esta investigación usaremos métodos estadísticos y de machine learning para predecir el comportamiento futuro de la serie de tiempo producción de plata del Perú (2001-2023).

4.2.3. Diseño de investigación

El diseño de la presente investigación es cuantitativa, ya que el pronóstico de series de tiempo se basan en modelos estadísticos y matemáticos, además permiten realizar pronósticos basados en patrones estadísticos identificados en los datos históricos.

Es no experimental ya que según Bautista, 2019 , este tipo de diseño no se manipulan deliberadamente variables independientes, sino el investigador observa y analiza los

fenómenos tal como ocurren naturalmente, sin intervenir o modificar el entorno.

Además es longitudinal, ya que los datos fueron recolectados a través del tiempo

4.3. Población y muestra

Se trabajará con la serie de tiempo mensual "Producción de plata del Perú" durante los periodos Enero-2001 y Junio-2023. Los datos de esta serie fueron obtenidos de la base de datos del BCRP.

4.4. Técnicas de análisis e instrumentos

Ya que los datos están limpios, primero se realizará un análisis exploratorio de la serie, después se corregirá el quiebre que presenta en un periodo con la ayuda del análisis de componentes principales. Luego se hará particionamiento de la serie en train-test y posteriormente su predicción usando el modelo híbrido, haciendo uso

4.5. Cuadro de operacionalización de variables

Variable	Tipo	Escala de medición	Descripción
Z	Cuantitativo discreta	Razón	Producción de plata total (kg. f) del Perú durante los periodos Enero2001 y Junio2023

Figura 4.1: Descripción de la variable usada en la investigación

Capítulo 5

Resultados

El propósito de esta investigación es proponer una alternativa en la corrección de quiebres estructurales, así como una posible solución cuando se obtengan errores no aleatorios de tal manera que se pueda obtener adecuadas predicciones como también se pueda capturar adecuadamente la estructura de los datos. A continuación en la figura 5.3 se muestra el comportamiento de la serie a lo largo del periodo 2001 al 2023.

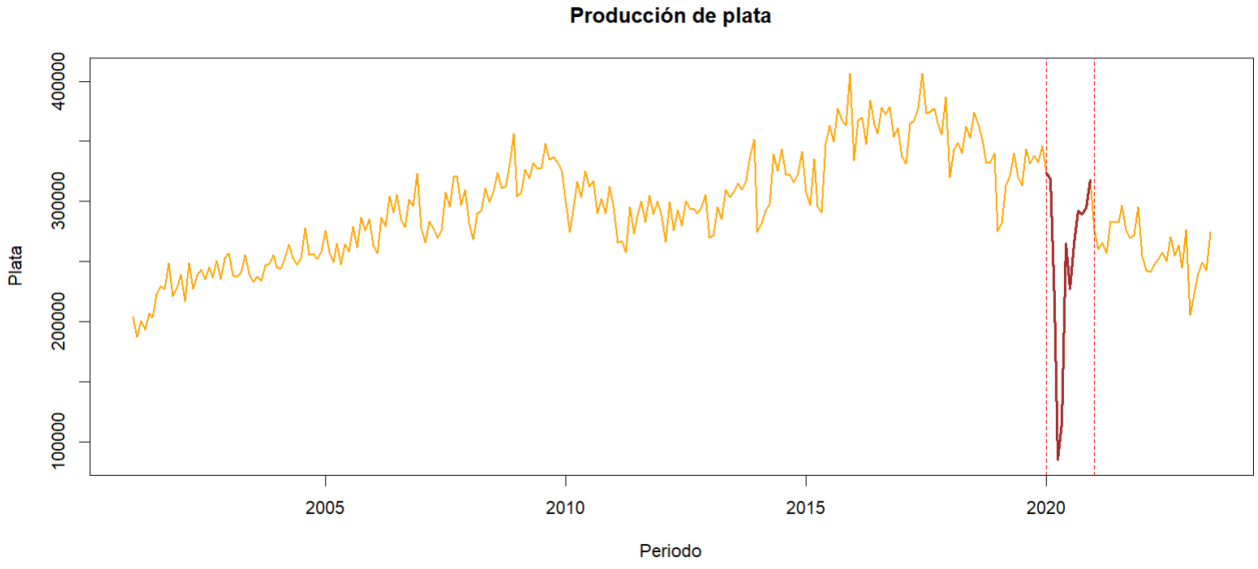

Figura 5.1: Serie producción de plata entre los años 2001 a 2023.

Se hace notar que en el periodo 2020 la serie presenta un quiebre estructural debido al contexto (covid-19) que se estuvo viviendo en este entonces. Ya que este es un periodo anómalo se tiene que corregir.

5.1. Corrección del quiebre estructural

A continuación, se presenta el análisis realizado en la corrección del quiebre estructural:

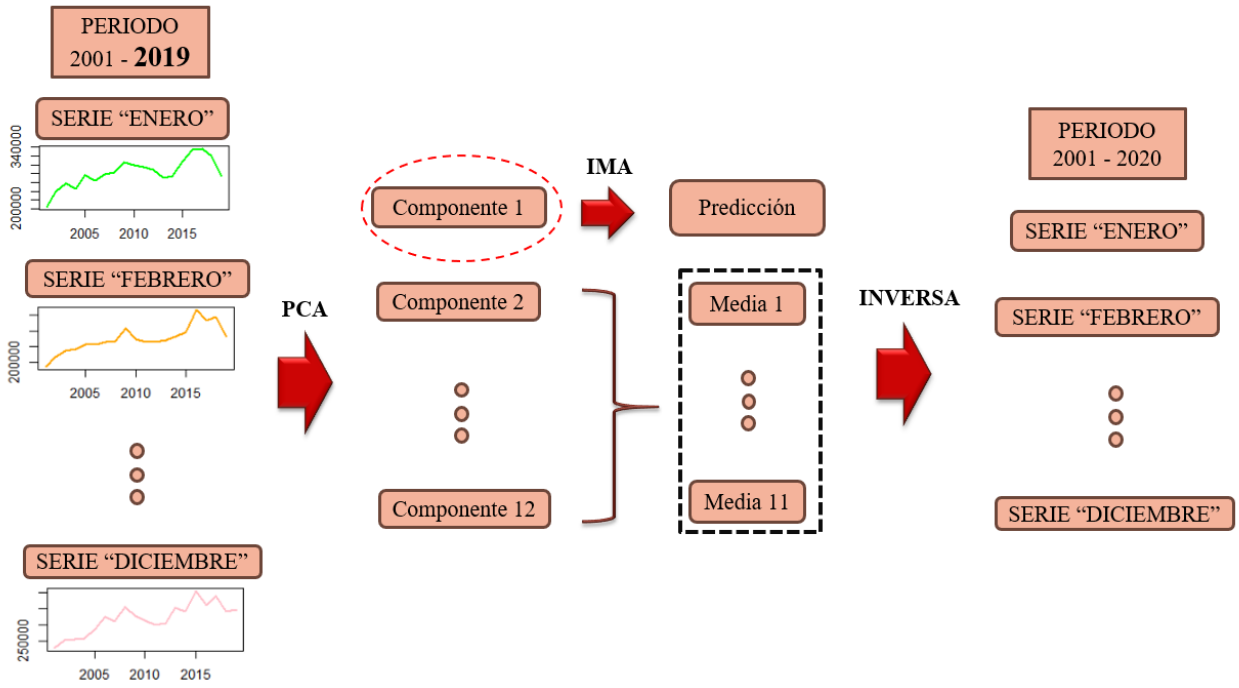

Figura 5.2: Esquema de investigación para la corrección de la serie de tiempo producción de plata entre los años 2001 a 2019.

- La producción de plata, para cada mes, a lo largo de los años $2001 - 2019$ presenta cierto patrón. Entonces se puede corregir el periodo 2020 usando series individuales por mes en el periodo $2001 - 2019$ haciendo pronósticos para cada una de ellas. Sin embargo, se evidencia que estas series no son independientes.

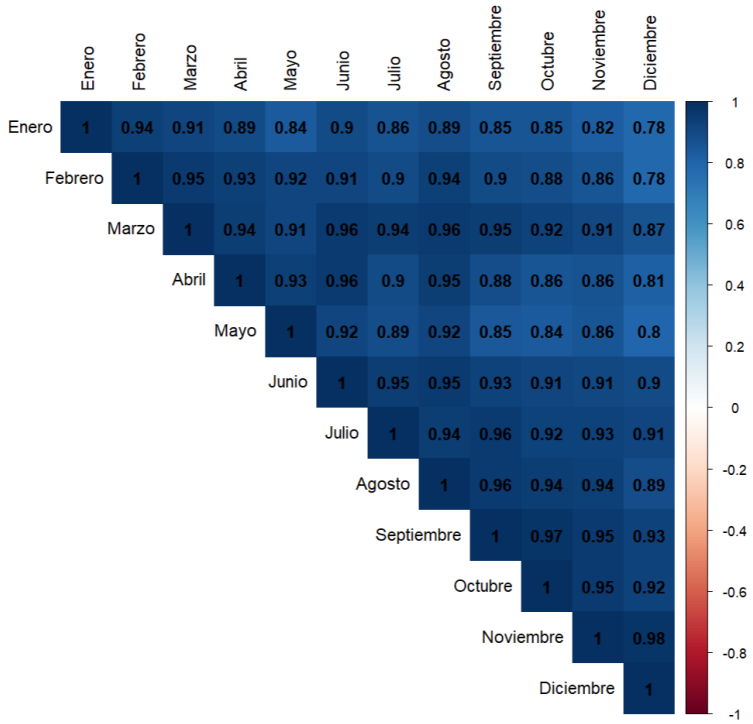

Figura 5.3: Matriz de correlación para cada una de las series individuales por mes en el periodo $2001 - 2019$

- Se observa en la figura 5.3 que las correlaciones no bajan de 0,78, esto indica que no se puede hacer pronósticos individuales. Por esta razón se hace el uso de los componentes principales.

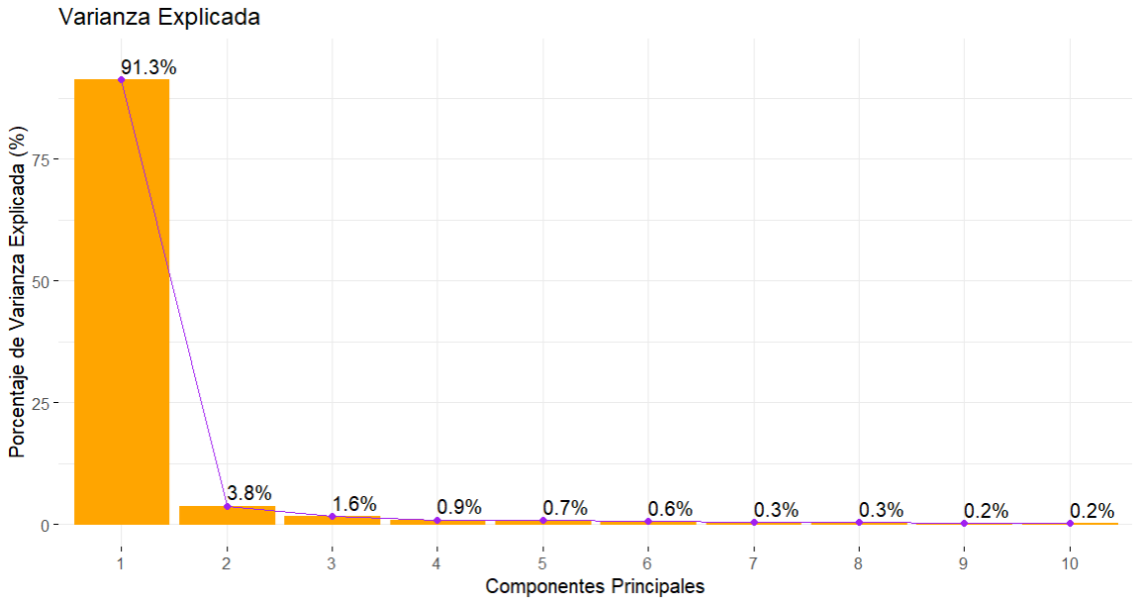

Figura 5.4: Varianza explicada por cada una de las 12 componentes principales.

- De la figura 5.4 se observa que la primera componente capta el 91,3 % de varianza explicada, mientras que las demás aportan no más de 4 %.

- Luego de analizar las componentes, ajustamos la primera componente a un modelo IMA$(1, 1)$ y, de las demás componentes obtenemos su media para no perder información. Este es un método que se aplicó por Amao Suxo, 2018 pag.74 por la misma razón, no perder información de las demás componentes.

- Finalmente sacamos la inversa para obtener nuevamente nuestra matriz de series por mes. Cabe resaltar que esta matriz tiene una fila de más, donde, aquella fila representa las nuevas observaciones para el periodo 2020 de la serie producción de plata. La serie corregida se muestra en la figura 5.6

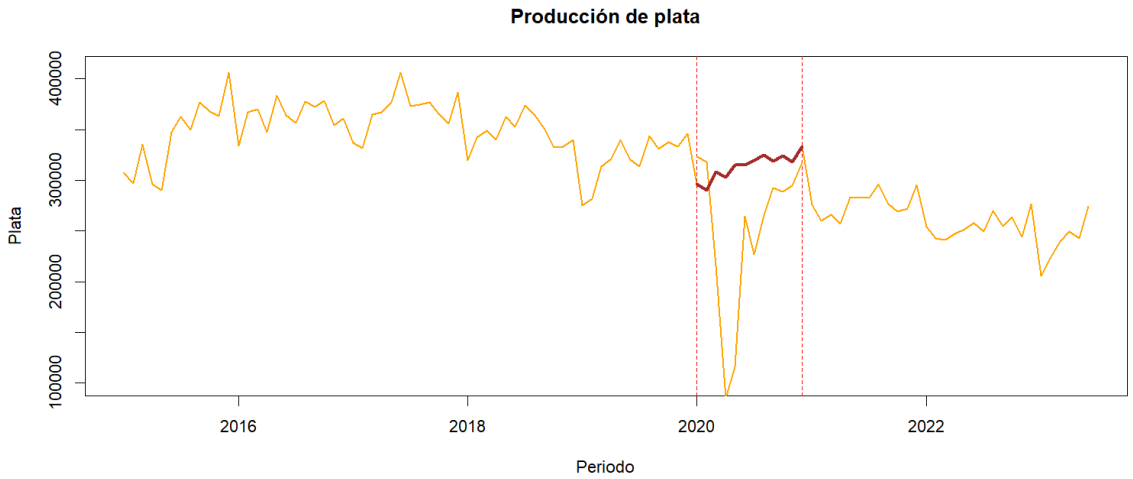

Figura 5.5: Serie de tiempo producción de plata, donde la línea azul representa la corrección del periodo 2020.

5.2. Análisis exploratorio de la serie corregida

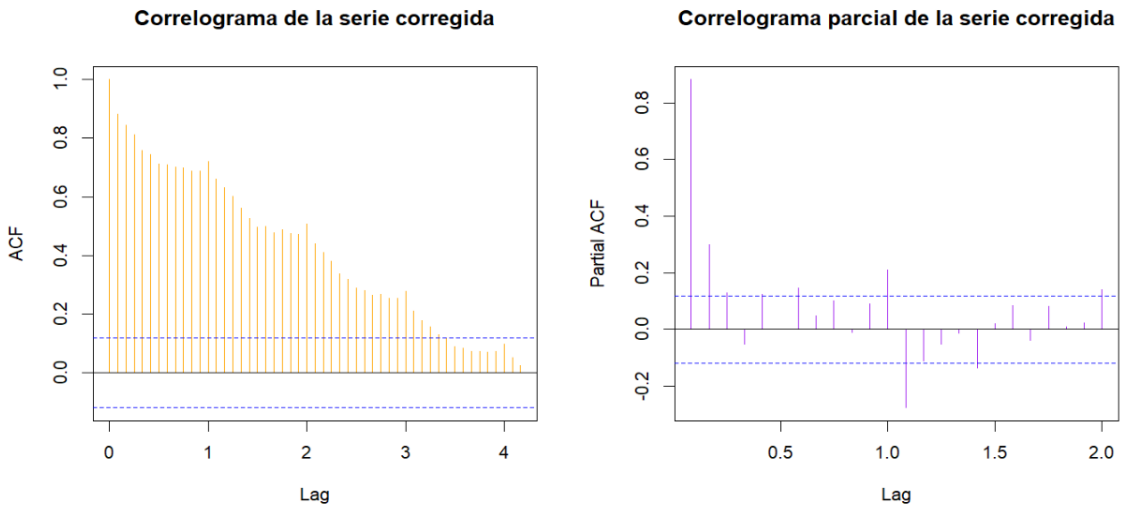

Figura 5.6: Correlograma y correlograma parcial de la serie corregida producción de plata en el Perú (2001-2023).

- **Tendencia:** Se puede observar en el correlograma que existe tendencia intensa en la parte de medias móviles.
- **Estación:** También se observa estación, aunque parece que está dominado por la tendencia.
- **Estacionariedad:** La serie no es estacionaria.

5.3. Aplicación del modelo híbrido SARIMA-MLP

A continuación, se presenta el análisis realizado en la aplicación del modelo híbrido SARIMA-MLP

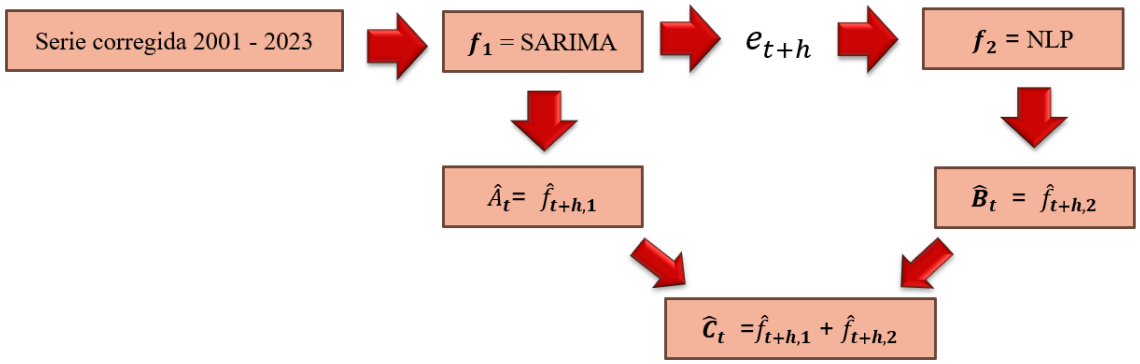

Figura 5.7: Esquema de investigación del modelo híbrido SARIMA-MLP para la serie corregida producción de plata entre los años 2001 a 2023.

- Primero separamos los datos en entrenamiento (periodo 2001-2021) y prueba (periodo 2022-2023) para luego ajustar los datos de entrenamiento a un modelo SARIMA$(1,1,1)(2,0,0)$.

	N° de observaciones	Periodo de inicio	Periodo de final
TRAIN	252	2001-1	2021-12
TEST	18	2022-1	2023-6

Figura 5.8: Separación de los datos en muestra de entrenamiento y testeo.

- Con los residuos obtenidos ajustamos un modelo de MLP con el objetivo de capturar los patrones no lineales de la serie.

- Finalmente, se suman los pronósticos de ambos ajustes en el test obteniéndose una nueva predicción, la cual se compara con los valores originales de los datos de prueba y se evalúa su rendimiento.

Por último, se presenta, en la figura 5.9, las métricas obtenidas para los modelos SARIMA sin corrección, SARIMA corregido y el modelo híbrido SARIMA-MLP corregido.

	MAE	MAPE	RMSE
SARIMA SIN CORRECCIÓN	23843.05	10.03	29868.36
SARIMA CORREGIDO	10383.87	4.31	14036.24
SARIMA-MLP CORREGIDO	9877.024	4.10	13163.49

Figura 5.9: Métricas para cada uno de los modelos.

Capítulo 6

Conclusiones y recomendaciones

6.1. Conclusiones

- Los pronósticos en la serie producción de plata en el Perú (2001-2023) mejoraron al usar el enfoque híbrido SARIMA-MLP con corrección PCA.
- El uso de PCA ayudó a corregir el quiebre estructural de la serie de tiempo producción de plata del Perú en el periodo 2020.
- Se implementó el modelo híbrido SARIMA-MLP con la serie corregida obteniendo un MAPE igual a 4,1.
- Como se observó en la figura 5.9 haciendo uso del enfoque híbrido SARIMA-MLP con corrección PCA se obtiene un MAPE igual a 4,1, siendo este menor en comparación al SARIMA.

6.2. Recomendaciones

- Se puede mejorar la implementación del modelo híbrido SARIMA-MLP haciendo uso de más capas ocultas en el proceso de la red neuronal.

- Se recomienda comparar la serie corregida con PCA con otras formas de corrección de una serie. Ya sea con medias móviles o añadiendo una variable indicadora al modelo. Y compararlas para evaluar su rendimiento.

- Cada serie de tiempo tiene sus peculiaridades, en ese sentido, el enfoque propuesto aquí puede como no favorecer a las predicciones. Se recomienda hacer uso de otro enfoque híbrido tal sea el caso.

Bibliografía

Amao Suxo, C. (2018). Forescat modeling of spatio-temporal raster data using principal component analysis and a neural networks-wavelet decomposition model.

Ballesteros López, F., et al. (2023). Estructuras híbridas para el modelado y pronóstico de series temporales: metodologías y aplicaciones.

Bautista, E. F. (2019). PERCEPCIÓN DEL ÉXITO EMPRESARIAL, DESDE LA PERSPECTIVA DE PLANEACIÓN EN EL MICROEMPRESARIO EN PUEBLA. *Hitos de Ciencias Económico Administrativas*, *25*(72), 271-285.

Berberich, D. (2020). *ggSAC NOT* [Tesis doctoral, Karlsruhe Institute of Technology].

Boshnakov, G. N. (2016). Introduction to Time Series Analysis and Forecasting, Wiley Series in Probability and Statistics, by Douglas C. Montgomery, Cheryl L. Jennings and Murat Kulahci (eds). Published by John Wiley and Sons, Hoboken, NJ, USA, 2015. Total number of pages: 672 Hardcover: ISBN: 978-1-118-74511-3, ebook: ISBN: 978-1-118-74515-1, etext: ISBN: 978-1-118-74495-6.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Ince, H., & Trafalis, T. B. (2006). A hybrid model for exchange rate prediction. *Decision Support Systems*, *42*(2), 1054-1062.

Lozada, J. (2014). Investigación aplicada: Definición, propiedad intelectual e industria. *CienciAmérica: Revista de divulgación científica de la Universidad Tecnológica Indoamérica*, *3*(1), 47-50.

Mercado, A. F. O. (2021). Modelos híbridos SARIMA-ANN para pronósticos de la COVID-19 en el Perú. *Revista IECOS*, *22*(1), 7-22.

MineriaenLinea. (Julio-2023). Producción de Plata en Perú 2023: Análisis de la Caída en la Minería de Plata [Fecha de acceso: DD de Mes de Año].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830. https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Peña Sánchez de Rivera, D. (2005). *Análisis de series temporales*. Editorial Alianza.

Ruiz-Aguilar, J., Turias, I., & Jiménez-Come, M. (2014). Hybrid approaches based on SARIMA and artificial neural networks for inspection time series forecasting. *Transportation Research Part E: Logistics and Transportation Review*, *67*, 1-13.

Sierra, G. M., Rodríguez, N. J. M., et al. (2019). Modelación y comovimientos de la tasa de cambio colombiana, 2011-2017. *Revista de Metodos Cuantitativos para la Economia y la Empresa*, *28*, 301-341.

Xu, S., Chan, H. K., & Zhang, T. (2019). Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach. *Transportation Research Part E: Logistics and Transportation Review*, *122*, 169-180.