

UNIVERSIDAD NACIONAL DE PIURA

Facultad de Ciencias

Escuela Profesional de Estadística



TESIS

“MODELO SARIMA Y RED NEURONAL RECURRENTE PARA EL PRONOSTICO DE LA PRODUCCIÓN DE MANGO EN EL VALLE DE SAN LORENZO, 2024-2026”

Presentado por:

Br. Jairon Kevin Ojeda Silupu

Br. Jahayra Sheryden Rodríguez Rodríguez

Asesor:

MSc. Lemin Abanto Cerna

PARA OPTAR EL TITULO PROFESIONAL DE LICENCIADO EN ESTADÍSTICA

Línea de investigación:

Matemática y estadística

Sub Línea de investigación:

Series de Tiempo

Piura - Perú

2024



UNIVERSIDAD NACIONAL DE PIURA
Facultad de Ciencias
Escuela Profesional de Estadística



TITULO

“MODELO SARIMA Y RED NEURONAL RECURRENTE PARA EL PRONOSTICO
DE LA PRODUCCIÓN DE MANGO EN EL VALLE DE SAN LORENZO, 2024-2026”

LÍNEA DE INVESTIGACIÓN:

Matemática y estadística

SUB LÍNEA DE INVESTIGACIÓN:

Series de Tiempo

Br. Jahayra Sheriden Rodríguez Rodríguez
(Tesis)

Br. Jairon Kevin Ojeda Silupu
(Tesis)

MSc. Lemin Abanto Cerna
(Asesor)

Piura-Perú
2024

DECLARACIÓN JURADA DE ORIGINALIDAD DE LA TESIS

Yo, **JAIRON KEVIN OJEDA SILUPU**, identificado con DNI N° 75411140, Bachiller de la Escuela Profesional de Estadística de la Facultad de Ciencias y domiciliado en el C.P. Pedregal, del Distrito de Tambo Grande, Provincia de Piura y Departamento de Piura, con celular N° 942507330 y E-mail jaironojeda99@gmail.com,

DECLARO BAJO JURAMENTO: Que la tesis que presento es original e inédita, no siendo copia parcial ni total de una tesis desarrollada y/o realizada en el Perú o en el Extranjero. En caso contrario, de resultar falsa la información que proporciono, me sujeto a los alcances de lo establecido en el Art. N° 411 del código penal, concordante con el Art. N° 32 de la Ley N° 27444 y Ley del Procedimiento Administrativo General y las Normas Legales de Protección a los Derechos de autor.

En fe de lo cual firmo la presente.

Piura, 1 de mayo de 2025

Jairon Kevin Ojeda Silupu
DNI N.º 75411140

Artículo N° 411.- El que, en un procedimiento administrativo, hace una falsa declaración en relación con hechos o circunstancias que le corresponde probar, violando la presunción de veracidad establecida por ley, será reprimido con pena privativa de libertad no menos de uno ni mayor de cuatro años.

Art. 4. Inciso 4.12 del Reglamento del Registro Nacional de Trabajos de investigación para optar grados académicos y títulos profesionales –RENATI Resolución de consejo directivo N° 033-2016-SUNEDU-CD.

DECLARACIÓN JURADA DE ORIGINALIDAD DE LA TESIS

Yo, **JAHAYRA SHERIDEN RODRIGUEZ RODRIGUEZ**, identificado con DNI N° 75148803, Bachiller de la Escuela Profesional de Estadística de la Facultad de Ciencias y domiciliado en AA.HH. Los Médanos, Mz H – Lt 18, del Distrito de Castilla, Provincia y Departamento de Piura, con celular N° 931172612 y E-mail jahayrasheriden@gmail.com,

DECLARO BAJO JURAMENTO: Que la tesis que presento es original e inédita, no siendo copia parcial ni total de una tesis desarrollada y/o realizada en el Perú o en el Extranjero. En caso contrario, de resultar falsa la información que proporciono, me sujeto a los alcances de lo establecido en el Art. N° 411 del código penal, concordante con el Art. N° 32 de la Ley N° 27444 y Ley del Procedimiento Administrativo General y las Normas Legales de Protección a los Derechos de autor.

En fe de lo cual firmo la presente.

Piura, 1 de mayo de 2025

Jahayra Sheriden Rodriguez Rodriguez
DNI N.º 75148803

Artículo N° 411.- El que, en un procedimiento administrativo, hace una falsa declaración en relación con hechos o circunstancias que le corresponde probar, violando la presunción de veracidad establecida por ley, será reprimido con pena privativa de libertad no menos de uno ni mayor de cuatro años.

Art. 4. Inciso 4.12 del Reglamento del Registro Nacional de Trabajos de investigación para optar grados académicos y títulos profesionales –RENATI Resolución de consejo directivo N° 033-2016-SUNEDU-CD.

DECLARACIÓN JURADA DE ORIGINALIDAD DE LA TESIS

TÍTULO DE LA TESIS

“MODELO SARIMA Y RED NEURONAL RECURRENTE PARA EL PRONÓSTICO DE LA PRODUCCIÓN DE MANGO EN EL VALLE DE SAN LORENZO, 2024–2026”

Yo, **MSC. LEMIN ABANTO CERNA**, asesor de tesis, identificado con Documento Nacional de Identidad DNI° 17930402, de la Facultad de Ciencias, Escuela Profesional de Estadística.

DECLARO BAJO JURAMENTO: Que la tesis que se presenta es original e inédita, no siendo copia parcial ni total de una tesis desarrollada y/o realizada en el Perú o en el extranjero. En caso contrario, de resultar falsa la información que proporciono, me sujeto a los alcances de lo establecido en el Art. N° 411 del Código Penal, concordante con el Art. N° 32 de la Ley N° 27444 – Ley del Procedimiento Administrativo General, y las normas legales de protección a los derechos de autor.

En fe de lo cual firmo la presente.

Piura, 1 de mayo de 2025

**MSC. Lemin Abanto Cerna
DNI N.º 17930402**

Artículo N° 411.- El que, en un procedimiento administrativo, hace una falsa declaración en relación con hechos o circunstancias que le corresponde probar, violando la presunción de veracidad establecida por ley, será reprimido con pena privativa de libertad no menor de uno ni mayor de cuatro años.

Art. 4. Inciso 4.12 del Reglamento del Registro Nacional de Trabajos de Investigación para optar grados académicos y títulos profesionales – RENATI. Resolución de Consejo Directivo N° 033-2016-SUNEDU-CD.

UNIVERSIDAD NACIONAL DE PIURA
Facultad de Ciencias
Escuela Profesional de Estadística



TESIS

**“MODELO SARIMA Y RED NEURONAL RECURRENTE PARA EL
PRONOSTICO DE LA PRODUCCIÓN DE MANGO EN EL VALLE DE SAN
LORENZO, 2024-2026”**

APROBADA EN CONTENIDO Y ESTILO POR

Lic. Carlos Eduardo Cabrera Prieto, Dr.
Presidente

Ing. Ricardo Antonio Armas Juárez, M.Sc.
Secretario

Lic. Ronald Eduardo Minchola Alza, M.Sc.
Vocal



UNIVERSIDAD NACIONAL DE PIURA

FACULTAD DE CIENCIAS



“AÑO DE LA RECUPERACIÓN Y CONSOLIDACIÓN DE LA ECONOMÍA PERUANA”

ACTA DE SUSTENTACIÓN 057-2025-UI-FC-UNP

Los Miembros del Jurado Calificador que suscriben, reunidos para evaluar la Tesis denominada **“MODELO SARIMA Y RED NEURONAL RECURRENTE PARA EL PRONOSTICO DE LA PRODUCCIÓN DE MANGO EN EL VALLE DE SAN LORENZO, 2024-2026”**; presentada por el Señor Bachiller **JAIRON KEVIN OJEDA SILUPU**, con el asesoramiento del **Lic. Lemin Abanto Cerna M.Sc.**; oídas las observaciones y respuestas a las preguntas formuladas, y de conformidad al Reglamento de Tesis para obtener el Título Profesional en la Facultad de Ciencias, lo declaran:

APROBADO (X)

DESAPROBADO ()

Con la mención de:

Sobresaliente

(X) En consecuencia, queda en condición de ser ratificado por el Consejo de Facultad de Ciencias de la Universidad Nacional de Piura, y recibir el **TITULO PROFESIONAL DE LICENCIADO EN ESTADÍSTICA**.

() En consecuencia, queda en condición de ser ratificado por el Consejo Universitario de la Universidad Nacional de Piura, y recibir el **TITULO PROFESIONAL DE LICENCIADO EN ESTADÍSTICA**; después que el sustentante incorpore la sugerencia del Jurado Calificador.

Piura, 04 de junio del 2025.

Lic. CARLOS EDUARDO CABRERA PRIETO, Dr.
PRESIDENTE DE JURADO DE TESIS

Ing. RICARDO ANTONIO ARMAS JÚAREZ, M.Sc.
SECRETARIO DE JURADO DE TESIS



Lic. RONALD EDUARDO MINCHOLA ALZA, M.Sc.
VOCAL DE JURADO DE TESIS

Campus Universitario - Urb. Miraflores S/N. Castilla
PIURA – PERU



UNIVERSIDAD NACIONAL DE PIURA

FACULTAD DE CIENCIAS



"AÑO DE LA RECUPERACIÓN Y CONSOLIDACIÓN DE LA ECONOMÍA PERUANA"

ACTA DE SUSTENTACIÓN 058-2025-UI-FC-UNP

Los Miembros del Jurado Calificador que suscriben, reunidos para evaluar la Tesis denominada "**“MODELO SARIMA Y RED NEURONAL RECURRENTE PARA EL PRONOSTICO DE LA PRODUCCIÓN DE MANGO EN EL VALLE DE SAN LORENZO, 2024-2026”**"; presentada por la Señorita Bachiller **JAHAYRA SHERIDEN RODRÍGUEZ RODRÍGUEZ**, con el asesoramiento del **Lic. Lemin Abanto Cerna M.Sc.**; oídas las observaciones y respuestas a las preguntas formuladas, y de conformidad al Reglamento de Tesis para obtener el Título Profesional en la Facultad de Ciencias, la declaran:

APROBADA (X)

DESAPROBADA ()

Con la mención de:

SOBRESALIENTE

(X) En consecuencia, queda en condición de ser ratificado por el Consejo de Facultad de Ciencias de la Universidad Nacional de Piura, y recibir el **TITULO PROFESIONAL DE LICENCIADO EN ESTADÍSTICA**.

(X) En consecuencia, queda en condición de ser ratificado por el Consejo Universitario de la Universidad Nacional de Piura, y recibir el **TITULO PROFESIONAL DE LICENCIADO EN ESTADÍSTICA**; después que la sustentante incorpore la sugerencia del Jurado Calificador.

Piura, 04 de junio del 2025.

Lic. CARLOS EDUARDO CABRERA PRIETO, Dr.
PRESIDENTE DE JURADO DE TESIS

Ing. RICARDO ANTONIO ARMAS JÚAREZ, M.Sc.
SECRETARIO DE JURADO DE TESIS



Lic. RONALD EDUARDO MINCHOLA ALZA, M.Sc.
VOCAL DE JURADO DE TESIS

Campus Universitario - Urb. Miraflores S/N. Castilla
PIURA – PERU

DEDICATORIA

“A mis padres, Javier y Nelida, por ser mi apoyo constante en cada paso de este camino. Su amor, dedicación y ejemplo firme me acompañaron y fortalecieron incluso en los momentos más difíciles. A Dios, por regalarme vida, propósito y la fuerza necesaria para seguir adelante. Y a mis amigos sinceros, quienes con su presencia y palabras de aliento estuvieron a mi lado cuando más los necesité.”

Jairon Kevin Ojeda Silupu

“Con profundo orgullo y gratitud, dedico este trabajo a mis padres, Juan y Fanny, cuya entrega, apoyo incondicional y sacrificios han sido fundamentales para alcanzar este logro. Esta tesis es el reflejo de su amor, esfuerzo compartido y del ejemplo que han sido para mí a lo largo de la vida. A mis hermanos, Jean, Juan David, y a mi hermana Yumey, les dedico esta meta alcanzada con la esperanza de ser una fuente de inspiración y motivación para que persistan en la búsqueda de sus propios sueños. A Anderson, por estar presente en cada etapa de este camino, acompañándome con paciencia y cariño en los momentos buenos y en los desafíos. A mis abuelos, a mi tía Pascuala y mi primo Piero, por su apoyo.”

Jahayra Sheriden Rodríguez Rodríguez

AGRADECIMIENTO

Agradezco a Dios, por ser mi refugio en los momentos de duda y mi impulso en los de certeza; a mis padres, por su respaldo incondicional y por inculcarme los valores que han guiado mi formación académica y personal. A mi asesor de tesis MSc. Lemín Abanto Cerna, por su acompañamiento constante, su orientación clara y oportuna, y su alto compromiso académico, que fueron clave para el desarrollo riguroso y sólido de esta investigación.

Jairon Kevin Ojeda Silupu

Agradezco, en primer lugar, a Dios por brindarme salud, fortaleza y la oportunidad de culminar esta etapa tan importante. A mi familia, por ser el pilar fundamental en mi vida, por sus enseñanzas, valores y apoyo constante. A mi asesor, MSc. Lemín Abanto Cerna, por su guía y dedicación durante el desarrollo de esta investigación. Gracias a mi compañero de vida y a las personas cercanas que, con su cariño, comprensión y palabras de aliento, me han acompañado silenciosamente en los momentos de mayor esfuerzo, motivándome a seguir adelante tanto en lo personal como en lo académico.

Jahayra Sheriden Rodríguez Rodríguez

ÍNDICE

RESUMEN	xiii
ABSTRACT	xiv
INTRODUCCIÓN	1
CAPÍTULO I: ASPECTOS DE LA PROBLEMATICA	3
1.1 Descripción de la Realidad Problemática	3
1.2 Formulación del Problema de Investigación	5
1.2.1 Problema general	5
1.2.2 Problemas específicos	5
1.3 Justificación e Importancia de la Investigación	6
1.4 Objetivos de la Investigación	7
1.4.1 Objetivo General	7
1.4.2 Objetivos Específicos	7
1.5 Delimitación de la Investigación	8
1.5.1 Delimitación Espacial	8
1.5.2 Delimitación Temporal	8
1.5.3 Delimitación Económica	8
CAPÍTULO II: MARCO TEORICO	9
2.1 Antecedentes de la Investigación	9
2.1.1 Antecedentes Internacionales	9
2.1.2 Antecedentes Nacionales	10
2.1.3 Antecedentes Locales	12
2.2 Bases Teóricas	13
2.2.1 Series de tiempo	13
2.2.2 Procesos estocásticos estacionarios	16
2.2.3 Test estadísticos de estacionariedad.	17
2.2.4 Diferenciación	20
2.2.5 Notación de retroceso “Backshift notation”	22
2.2.6 Familia ARIMA y sus Extensiones	23

2.2.7 Metodología de Box y Jenkins	26
2.2.8 Redes Neuronales Artificiales	33
2.2.9 Algunas RNA'S Especializadas en Series de Tiempo	36
2.2.10 Long-Short Term Memory (LSTM)	38
2.2.11 Recorrido general que realiza una LSTM	41
2.2.12 Alternativas de pronósticos usando Redes LSTM	44
2.2.13 Producción de mango	46
2.2.14 Fases fenológicas	47
2.2.15 Los factores participes en la producción de mango	49
2.2.16 Zonas clave de producción en el Perú	52
2.2.17 Producción de mango en el Valle de San Lorenzo	53
2.3 Glosario de Términos Básicos	54
2.4 Hipótesis	55
2.4.1 Hipótesis General	55
2.5 Operacionalización de Variables	55
CAPÍTULO III: MARCO METODOLÓGICO	56
3.1 Enfoque	56
3.2 Diseño	56
3.3 Nivel	56
3.4 Tipo	56
3.5 Sujetos de Investigación	56
3.5.1 Población	56
3.5.2 Muestra	57
3.6 Métodos y Procedimientos	57
3.7 Técnicas e Instrumentos	58
3.8 Aspectos Éticos	58
CAPÍTULO IV: RESULTADOS Y DISCUSIÓN	59
4.1 Resultados	59
4.2 Discusión	91
CONCLUSIONES	95
RECOMENDACIONES	96

REFERENCIAS BIBLIOGRÁFICAS 97

ANEXOS	102
ANEXO 01: Matriz de Consistencia	102
ANEXO 02: Carta de solicitud de datos	103
ANEXO 03: Datos de la serie de tiempo empleada	105
ANEXO 05: Código R utilizado	106
ANEXO 06: Informe de Turnitin	118
ANEXO 06: Seguimiento del modelo SARIMA	119

ÍNDICE DE TABLAS

Tabla 1	Operacionalización de la variable	55
Tabla 2	Medidas descriptivas de la variable Producción de Mango	59
Tabla 3	Test de Dickey-Fuller Aumentada (ADF)	64
Tabla 4	Test de Kwiatkowski Phillips Schmidt Shin (KPSS)	65
Tabla 5	Modelos Estimados para la Serie de Producción de Mango con Diferencia Estacional	68
Tabla 6	Test's de Diagnóstico del modelo SARIMA (1,0,0)(2,1,0)[12]	70
Tabla 7	Test's de Diagnóstico del modelo SARIMA (0,0,1)(2,1,0)[12]	71
Tabla 8	Test's de Diagnóstico del modelo SARIMA (1,0,0)(0,1,1)[12]	72
Tabla 9	Test's de Diagnóstico del modelo SARIMA(0,0,1)(0,1,1)[12]	74
Tabla 10	Comparación entre modelos SARIMA	75
Tabla 11	Configuración de los modelos candidatos LSTM.	78
Tabla 12	Resumen de la Arquitectura del Modelo LSTM (A) creado en keras3 . . .	79
Tabla 13	Resumen de la Arquitectura del Modelo LSTM (B) creado en Keras3. . . .	81
Tabla 14	Resumen de la Arquitectura del Modelo LSTM (C) creado en Keras3. . . .	83
Tabla 15	Resumen de la Arquitectura del Modelo LSTM (D) creado en Keras3. . . .	85
Tabla 16	Monitoreo del Desempeño de los Modelos en Función de los Parámetros e Hiperparámetros.	87
Tabla 17	Evaluación de los modelos LSTM sobre conjunto de prueba (Test vs Predicciones).	87
Tabla 18	Contrastación de los modelos mediante métricas basadas en los errores del modelo.	88
Tabla 19	Pronósticos mensuales del modelo SARIMA(1,0,0)(2,1,0)[12] con intervalos de confianza.	90

ÍNDICE DE FIGURAS

Figura 1	Realizaciones de un Proceso Estocástico y serie de tiempo	13
Figura 2	Componentes de una serie de tiempo	16
Figura 3	Serie estacionaria vs no estacionaria.	17
Figura 4	Diagrama de flujo de la metodología Box-Jenkins	27
Figura 5	Ejemplo de correlograma ACF y PACF	30
Figura 6	Neurona Artificial genérica y su correspondencia biológica.	33
Figura 7	Principales funciones de activación para redes neuronales artificiales. . .	35
Figura 8	Arquitectura de una NNAR básica.	36
Figura 9	Estructura de los 3 tipos de RNN's	38
Figura 10	Módulo repetido en un LSTM con cuatro capas en interacción	39
Figura 11	Cinta transportadora	40
Figura 12	Compuertas en la LSTM	40
Figura 13	Primer recorrido de la LSTM estándar	41
Figura 14	Segundo recorrido de la LSTM estándar	41
Figura 15	Tercer recorrido de la LSTM estándar	42
Figura 16	Cuarto recorrido de la LSTM estándar	42
Figura 17	Primera variante de la LSTM	43
Figura 18	Segunda variante de la LSTM	43
Figura 19	Tercera variante de la LSTM	44
Figura 20	Fenología general del mango peruano según los meses del año	49
Figura 21	Rangos generales de temperaturas para la producción de mango	51
Figura 22	Provincias y distritos de Piura que producen mango según VBP.	52
Figura 23	Producción de mango en el Valle de San Lorenzo, y Provincias de Piura	57
Figura 24	Serie Temporal de la producción de mango (t)	60
Figura 25	Estacionalidad mensual de la producción de mango (t)	61
Figura 26	Distribución mensual de la producción de mango (t)	62
Figura 27	Componentes de la serie de producción de mango (t)	63
Figura 28	Diferencia estacional ($D=1$) de la producción mensual de mango	66
Figura 29	Correlogramas de la serie con diferencia estacional	67
Figura 30	Diagnóstico visual del modelo SARIMA $(1,0,0)(2,1,0)[12]$	69
Figura 31	Diagnóstico visual del modelo SARIMA $(0,0,1)(2,1,0)[12]$	71

Figura 32	Diagnóstico visual del modelo SARIMA (1,0,0)(0,1,1)[12]	72
Figura 33	Diagnóstico visual del modelo SARIMA(0,0,1)(0,1,1)[12]	73
Figura 34	Ajuste de modelos SARIMA preseleccionados sobre producción de mango (t)	75
Figura 35	Partición del conjunto de datos para encontrar el mejor LSTM.	77
Figura 36	Entrenamiento y validación de modelo LSTM (A).	79
Figura 37	Prueba del modelo LSTM A con 24 inputs.	80
Figura 38	Entrenamiento y validación de modelo LSTM (B).	81
Figura 39	Prueba del modelo LSTM B con 12 inputs.	82
Figura 40	Entrenamiento y validación de modelo LSTM (C)	83
Figura 41	Prueba del modelo LSTM C con 12 inputs.	84
Figura 42	Entrenamiento y validación de modelo LSTM (D)	85
Figura 43	Prueba del modelo LSTM D con 12 inputs.	86
Figura 44	Comparación de los mejores modelos LSTM (A) y SARIMA frente al test	88
Figura 45	Pronóstico de la producción (t) de mango usando SARIMA (1,0,0) (2,1,0) [12], desde agosto de 2024 a enero de 2026.	90

RESUMEN

El presente estudio, titulado “Modelo SARIMA y red neuronal recurrente para el pronóstico de la producción de mango en el valle de San Lorenzo, 2024-2026”, tuvo como objetivo general determinar la eficiencia de los modelos SARIMA y redes neuronales LSTM en el pronóstico mensual de la producción de mango. La investigación adoptó un enfoque cuantitativo, con diseño no experimental, de tipo aplicada y nivel predictivo, por otro lado la población comprendió todos los meses con registros de producción desde el inicio de la actividad hasta la actualidad, extrayéndose como muestra el periodo comprendido entre enero de 2000 y agosto de 2024. La información fue obtenida mediante búsqueda electrónica, solicitada a MIDAGRI y descargada en hojas de cálculo de Excel. En el análisis descriptivo, la serie de producción mensual de mango mostró alta variabilidad, con presencia de valores extremos que elevaron la media (15,148 t) frente a una mediana considerablemente menor (235 t), lo que evidencia una distribución sesgada. Se identificó una clara estacionalidad, con picos en diciembre y enero, y baja producción entre abril y julio debido a las temporadas. Siguiendo la metodología Box-Jenkins, se evaluaron siete modelos SARIMA, destacando el modelo SARIMA(1,0,0)(2,1,0)[12] como el más eficiente (MAE = 8,217; MASE = 0.833). Paralelamente, se entrenaron 4 modelos LSTM utilizando Keras3, destacando el modelo A con una arquitectura de dos capas ocultas (256 y 128 unidades), 24 entradas y 17 salidas. Posteriormente al comparar ambos modelos sobre un conjunto de prueba de 21 meses, el modelo SARIMA presentó un mejor desempeño en métricas clave (MAE y RMSE), especialmente en la predicción de valores extremos. Se concluye que el modelo SARIMA(1,0,0)(2,1,0)[12] es el más eficiente para ajustar y predecir la serie de tiempo de producción de mango en el valle de San Lorenzo. La proyección muestra un crecimiento progresivo desde agosto (1,472.69 t) hasta noviembre (10,520.33 t), seguido de una ligera caída en diciembre (8,214.21 t) y un repunte estimado en enero de 2026 (9,547.92 t).

Palabras clave: Estacionalidad, estacionariedad, SARIMA, Red neuronal recurrente.

ABSTRACT

The present study, entitled “SARIMA Model and Recurrent Neural Network for Mango Production Forecasting in the San Lorenzo Valley, 2024-2026,” had the general objective of determining the efficiency of SARIMA models and LSTM neural networks in monthly mango production forecasting. The research adopted a quantitative approach, with a non-experimental, applied, and predictive design. The population included all months with production records from the beginning of the activity to the present, extracting the period from January 2000 to August 2024 as a sample. The information was obtained through an electronic search, requested from MIDAGRI and downloaded into Excel spreadsheets. In the descriptive analysis, the monthly mango production series showed high variability, with the presence of extreme values that raised the mean (15,148 t) compared to a considerably lower median (235 t), evidencing a skewed distribution. A clear seasonality was identified, with peaks in December and January, and low production between April and July due to the seasons. Following the Box-Jenkins methodology, seven SARIMA models were evaluated, highlighting the SARIMA(1,0,0)(2,1,0)[12] model as the most efficient (MAE = 8.217; MASE = 0.833). In parallel, four LSTM models were trained using Keras3, highlighting model A with a two-hidden-layer architecture (256 and 128 units), 24 inputs, and 17 outputs. Subsequently, when comparing both models on a 21-month test set, the SARIMA model showed better performance in key metrics (MAE and RMSE), especially in predicting extreme values. It is concluded that the SARIMA(1,0,0)(2,1,0)[12] model is the most efficient to fit and predict the mango production time series in the San Lorenzo Valley. The projection shows a progressive growth from August (1,472.69 t) to November (10,520.33 t), followed by a slight drop in December (8,214.21 t) and an estimated rebound in January 2026 (9,547.92 t).

Keywords: Seasonality, stationarity, SARIMA, Recurrent neural network.

INTRODUCCIÓN

Uno de los pilares de la economía y de la seguridad alimentaria en el mundo es la producción agrícola, porque contribuye de manera significativa a las economías en muchas partes del mundo. En este contexto, la producción del mango en el Perú, así como en Piura, emerge como uno de los cultivos frutales de mayor importancia, tanto por su valor nutricional, como por la versatilidad culinaria, lo que lo convierte en un producto con bastante demanda en el mercado nacional e internacional. No obstante, su producción está sujeta a una serie de factores que pueden condicionar su rendimiento, incluyendo las condiciones climáticas, enfermedades, prácticas de cultivo, entre otros. La complejidad de estos factores y una gestión deficiente de los recursos agrícolas ha generado un bajo rendimiento por hectárea en la zona de Piura y en particular en el Valle San Lorenzo.

Es destacable que, en el entorno de la producción agrícola, del Valle de San Lorenzo, los factores, la crisis política actual y las decisiones a veces desacertadas de las autoridades competentes tienen un impacto directo y negativo en los agricultores y la economía regional y al ser una importante zona productora y exportadora de mango en el país, dicho impacto también se manifiesta en la economía nacional.

Pese a la gran demanda de este producto y de los beneficios económicos que genera a la región y al erario nacional, no se encuentra entre los cultivos con prioridad por el PNC (Plan Nacional de Cultivos) 2023-2024. Tampoco hay mucho apoyo de las instituciones que conforman el Sistema Nacional de Innovación Agraria (SNIA), y demás actores involucrados, que no hacen uso de la tecnología para mejorar su productividad y abordar las disparidades entre la oferta y la demanda de este producto, tanto a mediano como en el largo plazo. Estos aspectos y la falta de planificación de la producción no permiten anticipar la futura demanda del producto y tampoco establecer metas a futuro. Es aquí donde los pronósticos juegan un papel fundamental, para tomar conocimiento de las necesidades de producción en los próximos años y de esta manera satisfacer la demanda local, nacional e internacional. El uso de modelos predictivos se ha convertido en una herramienta fundamental para que los responsables de tomar decisiones agrícolas puedan anticipar la evolución de la producción de mango, identificar posibles riesgos y diseñar estrategias de mitigación.

El presente estudio se enmarca en este contexto, y está orientado a evaluar dos metodologías basadas en el Modelo SARIMA y en la Red Neuronal (LSTM) que han

demonstrado bastante eficiencia en la elaboración de pronósticos según la literatura científica, a fin de pronosticar la producción de mango en el Valle de San Lorenzo.

El estudio consta de cuatro (04) capítulos, en el primero se fundamenta los aspectos de la problemática, donde se menciona la realidad problemática, así como formulación del problema que se pretende estudiar, así mismo, se menciona la justificación e importancia del estudio como la producción de mango, los objetivos que se buscan medir y la delimitación de la investigación.

El segundo capítulo, lo conforma el marco teórico, donde se mencionan los estudios anteriores a nivel internacional, nacional y local; además, las bases teóricas la conforman las definiciones y lo que implica las variables a medir; así como el glosario de términos; la hipótesis y definición y operacionalización de variables también forman parte de este segundo capítulo.

El tercer capítulo consiste en el marco metodológico, destacando el enfoque, así como el diseño, nivel, tipo y sujetos de la investigación, además se detalla los métodos y procedimientos a seguir según las técnicas e instrumentos de la investigación, destacando los aspectos éticos del estudio.

Por último, el cuarto capítulo consistió en los resultados y discusión, seguido de las conclusiones y recomendaciones del estudio.

CAPÍTULO I: ASPECTOS DE LA PROBLEMATICA

1.1 Descripción de la Realidad Problemática

En el área agrícola, según el Ministerio de Desarrollo Agrario y Riego (MIDAGRI, 2023), el Perú presenta una superficie agrícola de 11.6 millones de has a nivel nacional; mientras que la región Piura tiene un área potencial agrícola de 277.372 has, de las cuales 157.798 has son irrigadas y se ubican en los valles de Alto, Medio y Bajo Piura; Chira y San Lorenzo.

En el contexto de la producción agrícola, como es el caso del cultivo de mango, se presenta la imperiosa necesidad de prever con precisión los cambios en la producción. Esta tarea guarda paralelismos con el desafío de anticipar y planificar recursos en diversos sectores, volviéndose esencial para afrontar los retos de la variabilidad en este cultivo y maximizar su potencial. La progresión de la producción, el área cultivada y el rendimiento del mango en el territorio peruano ha experimentado una mejora constante, en consonancia con las tendencias a nivel global. La creciente importancia de una alimentación de mayor calidad a nivel mundial constituye el principal incentivo para aumentar de forma sostenida la oferta de mangos. Según MIDAGRI (2023), debido a su excelente desempeño y alta calidad, el mango peruano logra alcanzar rendimientos de hasta 17 toneladas por hectárea, superando significativamente el promedio mundial de 7.2 toneladas por hectárea, por consiguiente, en el 2021, la producción media de mango en Perú alcanzó las 14.8 toneladas por hectárea, situándolo en la quinta posición a nivel mundial en términos de rendimiento. En ese sentido, aún subsisten diversas naciones con el potencial de duplicar e incluso triplicar sus rendimientos nacionales, lo que subraya la necesidad de introducir innovaciones y aplicar tecnologías que permitan mantenerse al nivel de los importantes productores de mango en el ámbito global, sin embargo, se considera que uno de los mayores desafíos mundiales del siglo para la agricultura es el cambio climático.

Perú, es uno de los países que más exporta este producto, teniendo como principal mercado Estados Unidos y Netherlands; no obstante, las estadísticas generan preocupación en el sector agrícola, debido a las fluctuaciones de producción constantes que se dan en cada año; según el sistema integrado de estadísticas agrarias, plataforma implementada por el MIDAGRI (2023), a nivel nacional la producción de mango ha experimentado fluctuaciones significativas desde siempre, por ejemplo, en los últimos años, específicamente en diciembre

de 2019, la producción alcanzó las 162,719.01 toneladas, al año siguiente, en diciembre de 2020, la producción aumentó a 176,791.53 toneladas. Sin embargo, en 2021, la producción disminuyó a 169,091.97 toneladas en el mismo mes, para diciembre de 2022, la producción continuó cayendo a 158,745.39 toneladas, con un rendimiento de 16.47 toneladas por hectárea. Luego, en septiembre de 2023, el mango ocupó el sexto lugar entre las principales partidas de frutos de exportación, no obstante, en noviembre del mismo año, la producción se redujo a 11,658.87 toneladas decayendo históricamente.

Aragón (2022) en un estudio de tendencias de mercado – Mango, elaborado para el instituto nacional de innovación agraria, menciona que la producción de mangos en el territorio peruano abarca prácticamente la totalidad de la nación, con sus niveles más significativos de producción registrados en la región costera del norte, particularmente en las zonas de Piura y Lambayeque, así mismo, la extensión de tierra agrícola cultivada en nuestro país alcanza las 4,155,678 hectáreas, donde el 0.94% de ellas, está dedicado al cultivo del mango. Hoy en día se estima que hay alrededor de 29 a 30 mil hectáreas de terreno destinado al cultivo de mangos, destacándose la variedad Kent como la más predominante en esta área.

Por consiguiente, en la región Piura, el MIDAGRI (2023) señala en su dashboard interactivo que se llevó a cabo una producción de mango de 3,821.00(t) para el mes de noviembre del año 2023, mientras que para el mismo mes del año anterior fue de 24,237.00(t), lo cual demuestra una seria caída de la producción para ese mes en un (45.3%) respecto al 2022. Sin embargo, al igual lo que sucedió a nivel nacional, la región Piura presentó una elevada producción en el 2019 de 175.729.00, la cual cayó a 122.960.00(t) en el mismo mes del 2020, evidenciando una drástica reducción de la producción en un -30.0%.

La región Piura presenta el 80.7% de valor bruto de producción de mango, donde en provincias destaca Piura (93.6%), Morropón (5.7) y Sullana (0.64%). Así mismo, a nivel distrital de la provincia de Piura, destaca el distrito de Tambogrande (80.9%) y Las Lomas (16.9%) de VBP de mango. Dicho ello, los 2 últimos mencionados conforman la parte media del Valle de San Lorenzo. Oscar Cornejo, presidente de la Mesa Técnica Regional de Cultivo de Mango 2023, mencionó que en el Valle de San Lorenzo hay 28 mil hectáreas de producción de mango sin producción, viéndose afectados cerca de 12 mil agricultores. No obstante, la crisis climática ocasionó muchas perdidas que afecta la economía en el Valle de San Lorenzo, donde la producción en su momento fue del 5% en el producto del mango de exportación.

Ahora bien, la producción de mango en la mayoría de los casos se da por pequeños agricultores que se encargan de la gestión de riego, que en algunas ocasiones tiende a

escasearse, la fertilización, labores culturales y manejo fitosanitario, implicando una serie de gastos que en algunas veces suelen ser en vano, ya que, las instituciones competentes muchas veces no pronostican de manera óptima la producción incurriendo a la mala toma de decisiones y evitando una capacitación adecuada para ellos, es por ello la vital importancia del crecimiento científico en el sector para mitigar una serie de problemas y favorecer directamente a los entes involucrados.

Como menciona Aragón (2022) en el estudio de tendencias de mercado del mango dado el significativo impacto económico y social que tiene el cultivo a nivel nacional y departamental (Piura), así como su posición clave dentro de las exportaciones agrícolas de Perú, resulta preocupante que este producto no sea incluido en el Marco Orientador de Cultivos 2022-2023 del MIDAGRI. Sorprendentemente, el mango no está entre los cultivos priorizados por el PNC (Plan Nacional de Cultivos) 2023-2024. Esto es notable considerando que el mango es un cultivo constante, lo que significa que su ciclo de crecimiento excede los doce meses y que su etapa de cosecha se prolonga a lo largo de varios años. Además, su instalación requiere una inversión considerable debido a sus altos costos iniciales.

Ante ello, las instituciones que forman parte del Sistema Nacional de Innovación Agraria (SNIA), y demás actores involucrados no hacen uso correcto de métodos científicos para realizar un pronóstico adecuado en la producción, lo cual puede desencadenar efectos significativos en la asignación de recursos financieros, tecnología y en el manejo de información, destinándose a esfuerzos innecesarios. Esto adquiere un contexto relevante en relación con la presente investigación que se centra en evaluar la eficacia y la precisión del Modelo SARIMA y las Redes Neuronales (LSTM) en la realización de pronósticos a periodos prolongados de la producción de mango que permitan anticipar la producción futura de mango en esa área geográfica teniendo en cuenta los desafíos del cambio climático.

1.2 Formulación del Problema de Investigación

1.2.1 Problema general

¿Qué modelo, entre SARIMA y LSTM, es más eficiente para el pronóstico mensual de la producción de mango en el Valle de San Lorenzo, Piura, 2024 – 2026?

1.2.2 Problemas específicos

- ¿Cómo se comporta la serie temporal de la producción mensual de mango en la parte media del Valle de San Lorenzo desde enero del 2000 hasta agosto de 2024?

- ¿Cuáles son los estimadores del modelo SARIMA aplicado a la serie temporal de la producción mensual de mango en la parte media del Valle de San Lorenzo, utilizando la metodología de Box Jenkins desde enero del 2000 hasta agosto de 2024?
- ¿Cuáles son los estimadores óptimos de los parámetros de las redes neuronales LSTM que permiten pronóstica la producción mensual de mango en la parte media del Valle de San Lorenzo, utilizando los datos históricos desde enero del 2000 hasta agosto de 2024?
- ¿Cuál de los dos modelos presenta mejor precisión en términos de error del modelo de la producción mensual de mango setiembre de 2024 hasta enero de 2026?

1.3 Justificación e Importancia de la Investigación

En el contexto de la producción agrícola, específicamente en el cultivo de mango en el Valle de San Lorenzo, los factores climáticos constantes, la crisis política actual y las decisiones a veces desacertadas de las autoridades competentes tienen un impacto directo y negativo en la economía de los agricultores. En este escenario, la obtención de pronósticos precisos y confiables se vuelve de vital importancia para una adecuada planificación agrícola, asignación de recursos y desarrollo de políticas efectivas, a fin de adoptar estrategias apropiadas y mitigar los impactos adversos.

El presente estudio se justifica de manera práctica, ya que sus resultados pueden beneficiar a diversos actores del sistema agroproductivo del Valle de San Lorenzo, como productores, cooperativas, empresas exportadoras y entidades gubernamentales. Los pronósticos generados optimizan la toma de decisiones estratégicas, la planificación productiva y la gestión de recursos, reduciendo riesgos económicos. Como complemento, se desarrolló una aplicación interactiva en Shiny que permite visualizar y dar seguimiento a los pronósticos de manera accesible para usuarios no especializados, reforzando la utilidad práctica de la investigación.

Así mismo, se presenta una justificación social, ya que los resultados impactarán positivamente en la sociedad en general. Los pronósticos de producción permitirán un suministro constante de alimentos frescos y nutritivos para la población, contribuyendo a una dieta saludable, así como en el bienestar de la calidad de vida, además, al asegurar la disponibilidad de mango a lo largo del tiempo, se fortalecerá la seguridad alimentaria a los consumidores en general.

Adicionalmente, la investigación responde a una justificación económica clara y

concreta. La precisión en los pronósticos de la producción de mango tendrá un impacto directo en la planificación financiera de los agricultores. Al prever con mayor exactitud la cantidad de mango que se espera cosechar, los agricultores podrán tomar decisiones financieras más informadas, ajustando sus estrategias de inversión y presupuesto. Además, la estabilidad en la oferta de mango resultante de los pronósticos también puede influir en el equilibrio de los costos en el mercado, evitando fluctuaciones abruptas que puedan afectar los ingresos de los productores y la economía local en general.

Presenta una justificación metodológica la cual cobra relevancia al proponer una comparación rigurosa entre dos enfoques de pronóstico: los modelos de Box Jenkins en específico el SARIMA y las redes neuronales recurrentes LSTM (Long short-term memory), para la predicción de la producción de mango en el valle de San Lorenzo. La utilización del modelo SARIMA implica una estructura de análisis de series temporales que es ampliamente usada cuando la serie considera componentes de estacionalidad en los datos en este caso recalmando que la producción se da por temporadas. Por otro lado, las redes neuronales (LSTM) son herramientas de aprendizaje automático que demuestran su utilidad en varios estudios previos en el sector agrícola con en el pronóstico de secuencias temporales complejas debido a su memoria a largo plazo, adaptabilidad y captura de dependencias temporales.

Es importante mencionar que el Valle de San Lorenzo se destaca por ser una zona importante productora y exportadora de mango en el país, por lo que resulta relevante estudiar y evaluar diferentes enfoques de pronóstico para este cultivo en particular.

1.4 Objetivos de la Investigación

1.4.1 Objetivo General

Determinar la eficiencia de los modelos SARIMA y redes neuronales (LSTM) para el pronóstico mensual de la producción de mango en el Valle de San Lorenzo, Piura, 2024 – 2026.

1.4.2 Objetivos Específicos

- Efectuar un análisis descriptivo de la serie temporal de la producción mensual de mango en la parte media del Valle de San Lorenzo, entre enero del 2000 a agosto de 2024.
- Estimar el modelo SARIMA para el pronóstico de la producción mensual de mango en la parte media del Valle de San Lorenzo, con la metodología de Box Jenkins, basados en los datos históricos comprendidos entre enero del 2000 a agosto de 2024.

- Estimar los parámetros óptimos de redes neuronales (LSTM) para el pronóstico de la producción mensual de mango en la parte media del Valle de San Lorenzo, basados en los datos históricos comprendidos entre enero del 2000 a agosto de 2024.
- Contrastar ambos modelos utilizando medidas de precisión basadas en los errores del modelo y pronosticar la producción mensual de mango desde setiembre de 2024 hasta enero de 2026.

1.5 Delimitación de la Investigación

1.5.1 Delimitación Espacial

El estudio se realizó en la Región Piura, específicamente en el Valle de San Lorenzo, ateniéndose a los registros de producción de mango realizados por el MIDAGRI.

1.5.2 Delimitación Temporal

El presente estudio usó una base de datos en formato serie desde el mes de enero de 2000 al mes de agosto de 2024 para la producción de mango.

1.5.3 Delimitación Económica

El presente estudio fue autofinanciado por los propios investigadores.

CAPÍTULO II: MARCO TEORICO

2.1 Antecedentes de la Investigación

2.1.1 Antecedentes Internacionales

Patrick et al. (2023) publicaron el artículo científico titulado “Series temporales y ensemble models para pronosticar el rendimiento del cultivo de banano en Tanzania, considerando los efectos del cambio climático”, con el objetivo de facilitar información valiosa respecto a la interacción entre el clima y el rendimiento considerando los modelos de pronósticos. Dentro de los materiales y métodos se consideraron variables climáticas mensuales (Precipitación, temperatura Min °C, Max °C, Humedad relativa, la humedad del suelo que funcionan como variables exógenas) y la variable rendimiento del cultivo de plátano (t/ha). Por otro lado, dentro de la metodología el estudio presenta doble enfoque, donde el primero es el análisis de correlación (como las variables climáticas se relación con el rendimiento del cultivo), por otro lado, el segundo enfoque son los modelos de pronósticos SARIMAX, State Space (SS), Long short-term memory (LSTM), para luego realizar un ensemble models utilizando un enfoque promedio ponderado. Dentro de los resultados, se dividieron los datos en dos conjuntos 80% (entrenamiento), 20% (prueba), donde se rescata lo obtenido en el modelo LSTM quien registró un MSE de 0.5288, MAE de 0.689 y RMSE de 0.7272, con un R^2 de 0.9013, mientras que SARIMAX tuvo el desempeño más débil con un MSE de 4.3797, MAE de 1.4789 y RMSE de 2.0928, junto a un R -cuadrado de 0.1825 siendo el menos eficiente. Se concluye que la aplicación de técnicas como SARIMAX y LSTM en el análisis de series temporales permite identificar patrones clave, aunque el modelo State Space (SS) demostró un mejor desempeño; es notable destacar que los modelos LSTM presentaron un mejor R^2 y un menor error de pronóstico en comparación con el modelo SARIMAX, lo cual es relevante para la presente investigación.

Kumar y Rao (2023), publicaron el artículo en Springer “Predicción de los niveles de humedad del suelo en Andhra Loyola College- India, utilizando modelos SARIMA y LSTM”, donde el objetivo fue realizar una comparación entre los dos modelos en base a las métricas de error, por consiguiente, el estudio analizó los datos de humedad media mensual del suelo radicular, que abarcan desde 1981 hasta 2022, y para ello el análisis emplea dos enfoques distintos: el promedio móvil integrado autorregresivo estacional estadístico (SARIMA) y un aprendizaje profundo memoria larga a corto plazo (LSTM). Dentro del análisis de resultados

se hizo un entrenamiento (train) de ambos modelos que abarco el período de 1981 a 2021, adquiridos del sitio agrícola en Andhra Loyola College, posteriormente, los datos de 2021 a 2022 se reservan para fines de prueba (test). Finalmente, en el contexto de la predicción de la humedad del suelo superficial (radicular), el modelo LSTM demuestra un rendimiento superior en comparación con SARIMA, pues el LSTM logra un MAPE notablemente más bajo de 0.0615 en comparación con el 0.1541 de SARIMA, un MAE reducido de 0.0316 en comparación con 0.0871 y un RMSE disminuido de 0.0412 en comparación con 0.102. Este patrón de precisión mejorada persiste en las predicciones de humedad del suelo de raíces, concluyendo que las técnicas de aprendizaje profundo, como el modelo LSTM, pueden ser más efectivas para la predicción de humedad del suelo radicular en este caso específico en la agricultura, que los modelos tradicionales estadísticos como SARIMA.

Tian et al. (2021), realizaron el estudio “Una red neuronal LSTM para mejorar las estimaciones del rendimiento del trigo mediante la integración de datos de teledetección y datos meteorológicos en la llanura de Guanzhong, República Popular de China”, con el objetivo de implementar un modelo LSTM para estimar el rendimiento del trigo en la llanura de Guanzhong integrando datos meteorológicos y dos índices de detección remota. Los hallazgos revelaron que al emplear dos pasos de tiempo y combinar datos meteorológicos con índices de detección remota, la precisión en la estimación del rendimiento mejoró considerablemente ($RMSE = 357,77 \text{ kg/ha}$ y $R^2 = 0,83$). Se llevó a cabo una evaluación comparativa de la precisión entre el modelo LSTM óptimo y otros enfoques como la red neuronal de retropropagación (BPNN) y la máquina de vectores de soporte (SVM). El modelo LSTM demostró una superioridad significativa sobre BPNN ($R^2 = 0,42$ y $RMSE = 812,83 \text{ kg/ha}$) y SVM ($R^2 = 0,41$ y $RMSE = 867,70 \text{ kg/ha}$), debido a su capacidad inherente de incorporar relaciones no lineales entre múltiples entradas y el rendimiento, gracias a su estructura de red neuronal recurrente. Se concluyó que la comparación con otros enfoques como BPNN y SVM mostró una mejora significativa en la precisión del modelo LSTM, así mismo la validación en diferentes sitios de muestreo confirmó la robustez y adaptabilidad del modelo, incluso ante fluctuaciones climáticas interanuales.

2.1.2 Antecedentes Nacionales

Chilón (2023), presento su tesis titulada “Redes neuronales recurrentes y modelos ARIMA para el pronóstico de la inflación en el Perú”, donde el objetivo central fue determinar la eficiencia del modelo de redes neuronales recurrentes LSTM en el pronóstico de la inflación de Perú en comparación con el modelo ARIMA. Por consiguiente, el estudio

es de tipo aplicado, con un diseño comparativo observacional y longitudinal, tomándose datos mensuales desde el año 2000 a diciembre de 2023 de la variación porcentual de índice de precios al consumidor. Para llevar a cabo los resultados se hizo una partición de los datos pues se usó 264 para el entrenamiento y 12 (enero de 2022 diciembre de 2022) para la prueba, es así que se pudo encontrar varios modelos ARIMA y ARIMA estacionales específicamente 5 candidatos de los cuales se eligió al optimo entre ellos, pasando el modelo SARIMA $(1,0,1)(1,1,1)$ [12], dado que obtuvo un menor RMSE de 0.418, un MAE de 0.339, y un EMC de 0.174, en comparación de los otros modelos de sus mismo enfoque, posteriormente se procedió a elaboración de la red neuronal recurrente LSTM, donde también se usó la misma lógica de la división de datos, y se logró configurar 5 modelos candidatos con este enfoque donde se tomó en cuenta modelos de entre 50 y 100 epoch, y las capas ocultas variaron entre 4 y 5, además el número de neuronas por capas vario entre 50 a 200 neuronas, y el número de capas de salida en todos los modelos fueron de 12, dado que ese fue el horizonte temporal que se pronosticó, luego de obtener las medidas de error de los 5 modelos LSTM candidatos, se encontró que el cuarto modelo LSTM conformado por 100 epoch, 4 capas ocultas (80 neuronas para la primera y segunda capa, y 100 para la tercera, la última capa no cuenta con neuronas) así mismo la función de activación para este modelo optimo fue la relu teniendo 12 capas de salida. Finalmente, al realizar la comparación de ambos modelos encontrados SARIMA $(1,0,1)(1,1,1)$ [12] y LSTM (R4), se encontró que el modelo SARIMA presento un RMSE de 0.4177, un MAE de 0.3388, y un EMC, y el LSTM presento un RMSE de 0.3987, un MAE de 0.3168, y un EMC de 0.1589, concluyendo que la cuarta configuración de la red LSTM presenta menor error, siendo el mejor modelo para pronosticar la inflación el Perú.

Almeyda (2022), realizó la tesis titulada “Pronóstico de la demanda internacional del banano orgánico de Perú usando algoritmos de Machine Learning”; la finalidad principal de la investigación fue crear, entrenar, probar y evaluar diferentes modelos que utilizan el aprendizaje automático para estimar la demanda de banano orgánico en el Perú. El estudio utiliza registros oficiales de banano orgánico exportado desde Perú entre 2001 y 2020. Respecto a los resultados se realizó la configuración de los hiperparámetros de las cuatro arquitecturas de redes neuronales (MLP, RNN, LSTM y GRU) destacando que el look back es de 6 meses para MLP y 12 para los demás, cada uno tiene una capa oculta con variación en las unidades (32 para MLP, 64 para RNN, 128 para LSTM, 64 para GRU), entrenados con 25 epochs y un learning rate de 0.001; la función de activación es ReLu para MLP y tanH para los otros, con una función de pérdida MSE y el optimizador Adam, todos implementan

early stopping, no utilizan dropout, manejan un batch size de 8 y normalización. Finalmente, el modelo RNN tiene el menor MSE (0.00147), RMSE (0.0338) y MAE (0.02885), así como el MAPE más bajo (2.88516%), lo cual indica que tiene el mejor rendimiento en términos de precisión de predicción comparado con los otros modelos. El estudio concluye que se entrenaron un total de 60 modelos obteniéndose una optimización de parámetros única para cada modelo, destacando que el modelo basados en redes neuronales (RNN) indicó mejor adaptabilidad a la dependencia temporal y estacionalidad de la serie temporal (MSE=0.0014), finalmente la elección de neuronas por capa oculta influyó en el aprendizaje y la precisión del modelo, con una arquitectura estándar de RNN mostrando el mejor desempeño, con una capa oculta y 64 unidades recurrente.

2.1.3 Antecedentes Locales

Robles y Semillan (2023) realizó un estudio en Piura sobre “Creación de un modelo econométrico, para pronóstico de la producción de las principales frutas de la región Piura”, con el objetivo de identificar el modelo econométrico más adecuado para anticipar la producción de las principales frutas cultivadas en la región Piura entre ellas, banano orgánico, limón, mango y papaya, se aplicó una metodología cuantitativa, bajo un enfoque de investigación aplicada, longitudinal y no experimental. Los resultados del estudio indican que la producción de mango fue pronosticada mediante la metodología Box-Jenkins, identificándose al modelo SARIMA (1,1,0)(0,1,1)[12] como el más eficiente, este modelo evidenció un comportamiento cíclico en los años proyectados, considerando además que la cosecha de mango se concentra estacionalmente al inicio y al final del año.

Ramos (2020) realizó un estudio en Piura sobre “Pronóstico de los ingresos tributarios mensuales del Gobierno Central Peruano aplicando Redes Neuronales y Modelos SARIMA, en base a los años 2003 – 2018”. El objetivo general del estudio es evaluar la capacidad de pronóstico del modelo SARIMA y el modelo de red neuronal artificial (ANN) con base en los ingresos tributarios mensuales del país registrados desde enero de 2003 a diciembre de 2018, y hacer pronósticos para 2019-2020. Respecto a la parte metodológica el estudio presenta un enfoque cuantitativo, con un diseño no experimental, de nivel predictivo de tipo longitudinal. Respecto a los resultados, se utilizó el software estadístico gratuito R Studio, encontrándose un modelo SARIMA (2,1,0)(1,1,1) con un R-cuadrado de 0.9578192 y un MAE de 308.593, cumpliendo los supuestos de los residuos. Finalmente se obtuvo una red neuronal autorregresiva NNAR (2,1,5)[12], junto con un R-cuadrado de 0.951387, y un MAE =335.5066. El estudio concluyó que el modelo SARIMA (2,1,0)(1,1,1) presenta un mayor

R-cuadrado=0.957, proporcionando un mejor modelo de ingresos tributarios para el gobierno peruano y se utilizó para generar proyecciones para 2019-2020.

2.2 Bases Teóricas

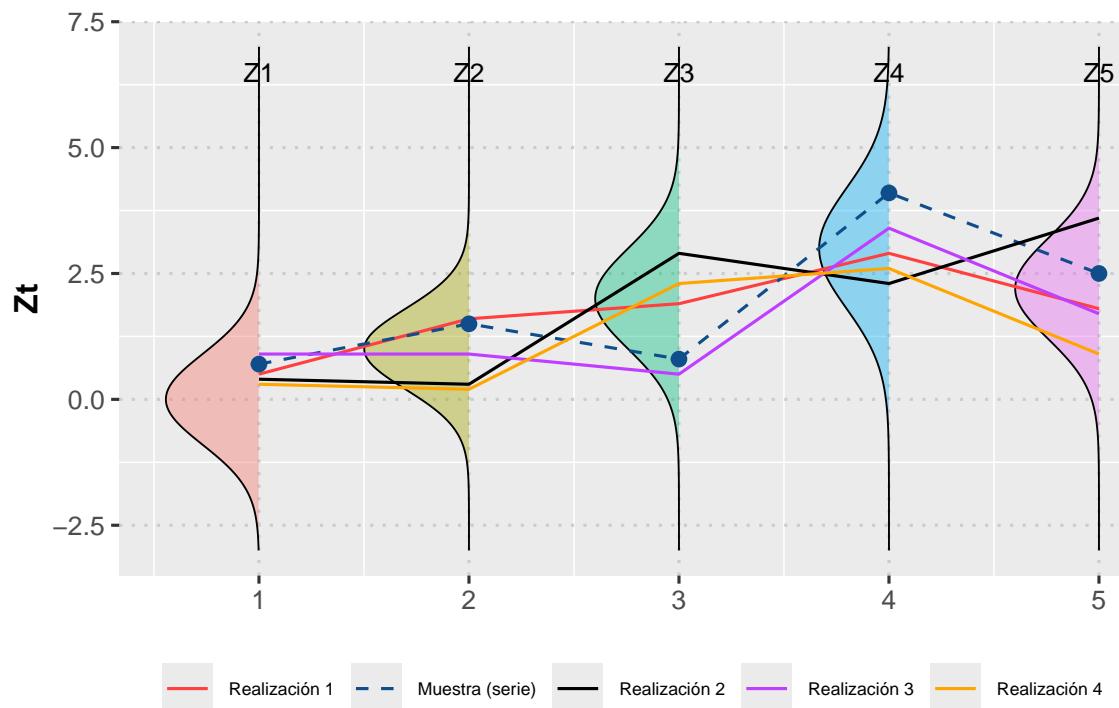
2.2.1 Series de tiempo

Definición

Desde el punto de vista teórico-estadístico Bowerman y Koehler (2007) afirman que una serie de tiempo es representada como $\{Z_1, Z_2, \dots, Z_n\}$, la cual se compone de observaciones secuenciales extraídas de un proceso estocástico, es decir es una muestra del proceso estocástico $Z_1 = z_1, Z_2 = z_2, \dots, Z_T = z_T$: Donde Z_T es una variable aleatoria y z_t es el valor específico que se obtiene dentro de la distribución en ese momento; en general, es una serie temporal en un periodo muestral, que tan solo es una parte del proceso estocástico del que procede dicha serie.

Figura 1

Realizaciones de un Proceso Estocástico y serie de tiempo



Nota: Elaboración propia, usando R

De la Figura 1, la serie de color azul puede considerarse una muestra de una realización en un proceso estocástico (donde pueden ver infinitas realizaciones) que toma los valores por ejemplo $Z_1 = z_1 = 0.7, Z_2 = z_2 = 1.5, Z_3 = z_3 = 0.8, Z_4 = z_4 = 4.1, Z_5 = z_5 = 2.5$ dentro de cada distribución. En la vida real las series de tiempo ya

registradas son consideradas muestras de una realización, donde los modelos de pronósticos estudian el proceso que las generó para poder estimar a futuro.

Por otro lado, según Bowerman y Koehler (2007), una serie de tiempo se define como una línea temporal que registra la evolución de un fenómeno o variable a lo largo del tiempo. Sin embargo, Guerrero (2009) argumenta que el término ‘serie de tiempo’ puede resultar insuficiente para describir completamente este conjunto de datos. Este autor señala que, en matemáticas, la palabra ‘serie’ se utiliza para denotar una suma infinita de valores variables, lo cual puede generar confusión. En su lugar, sugiere que un término más apropiado para referirse a estos grupos de datos sería ‘conjunto de datos de efectos en el tiempo’. A pesar de estas consideraciones, el término ‘serie temporal’ sigue siendo ampliamente utilizado debido a su popularidad y reconocimiento en el ámbito académico y profesional.

Así mismo Brockwell y Davis (2006) consideran una Serie de Tiempo como el proceso estocástico, donde las observaciones de un evento se toman en una secuencia a lo largo del tiempo con los mismos intervalos. Dicho de otro modo, el modelo de serie de tiempo hace referencia a un detalle de la distribución conjunta (fácilmente solo del promedio y covarianzas) de una continuación de variantes aleatorias Z_t para los cuales los z_t en cada t se afirma como una realización.

Objetivos de una Serie de Tiempo

Para Chatfield (2016), una serie de tiempo tiene como objetivos principales: descripción, explicación, predicción y control.

- **Descripción.** Como primer paso en el análisis se debe mostrar gráficamente las observaciones para obtener una medida descriptiva simple sobre las principales particularidades de la serie.
- **Explicación.** Cuando se toman observaciones de dos o más variables, puede ser posible utilizar la variación en una serie temporal para explicar la variación en otra serie, lo cual puede conducir a una comprensión más profunda del mecanismo que generó una serie temporal dada.
- **Predicción.** Podemos pronosticar los valores futuros que podría asumir la serie, es decir, estimar el valor futuro del fenómeno en estudio, en base a la conducta pasada de la serie.
- **Control.** Las series de tiempo a veces se recopilan o analizan para mejorar el control sobre algún sistema físico o económico. Por ejemplo, cuando se genera una serie de

tiempo que mide la calidad de un proceso de fabricación, el objetivo del análisis puede ser mantener el proceso funcionando a un nivel alto. Los problemas de control están estrechamente relacionados con la predicción en muchas situaciones. Por ejemplo, si se puede predecir que un proceso de fabricación se va a desviar del objetivo, entonces se pueden tomar las medidas correctivas adecuadas.

Componentes

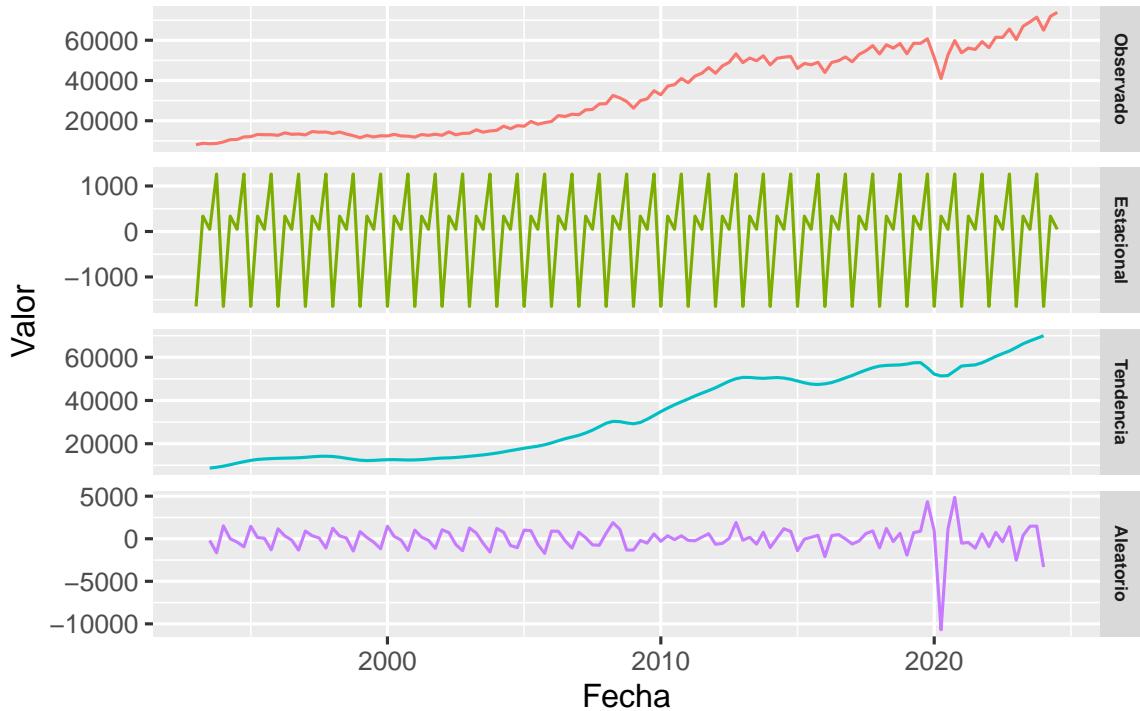
Para Chatfield (2016), las componentes de una serie de tiempo son: tendencia, estacionalidad, el componente irregular y componente cíclico.

Por otro lado, Villavicencio (2007) menciona tres componentes, Tendencia, Estacionalidad y aleatoriedad, fundamentando su análisis en el supuesto de que los valores asumidos por las variables observadas son el resultado de tres componentes que tienen un efecto común que conduce a los valores medidos:

- a. **Componente de tendencia:** Hace referencia al cambio de largo plazo que ocurre con respecto de la media. Las tendencias están determinadas por el suave movimiento de las series de largo plazo.
- b. **Componente estacional:** Muchas series temporales muestran una determinada periodicidad, es decir, fluctuaciones a lo largo de un determinado periodo de tiempo ya sea diario, mensual, etc. Por ejemplo, las ventas de la temporada escolar en el Perú. Estos efectos se comprenden fácilmente y pueden medirse claramente o incluso eliminarse del conjunto de datos; conocido como desestacionalización de la serie.
- c. **Componente aleatoria:** Este componente es la consecuencia de factores aleatorios o estocásticos que influyen a una serie temporal de forma aislada.

Figura 2

Componentes de una serie de tiempo



Nota: Elaboración propia usando R.

2.2.2 Procesos estocásticos estacionarios

Peña (2010) establece que los procesos estocásticos estacionarios son modelos esenciales para analizar series de tiempo. Estos procesos tienden a exhibir características estadísticas consistentes a lo largo del tiempo; es decir, con el tiempo, la estructura de media, varianza y correlación permanece constante. El análisis y pronóstico de series temporales se benefician enormemente de la estacionariedad, ya que permite reconocer patrones y tendencias en los datos sin requerir relaciones matemáticas complejas.

Un proceso aleatorio y_t es estacionario si es que los dos primeros momentos de la distribución conjunta (la media de las variables aleatorias y la varianza en el proceso aleatorio) no cambian con el tiempo; esto es:

$$E[y_t] = \mu < \infty \quad \text{para todo } t$$

$$E[(y_t - \mu)^2] < \infty \quad \text{para todo } t$$

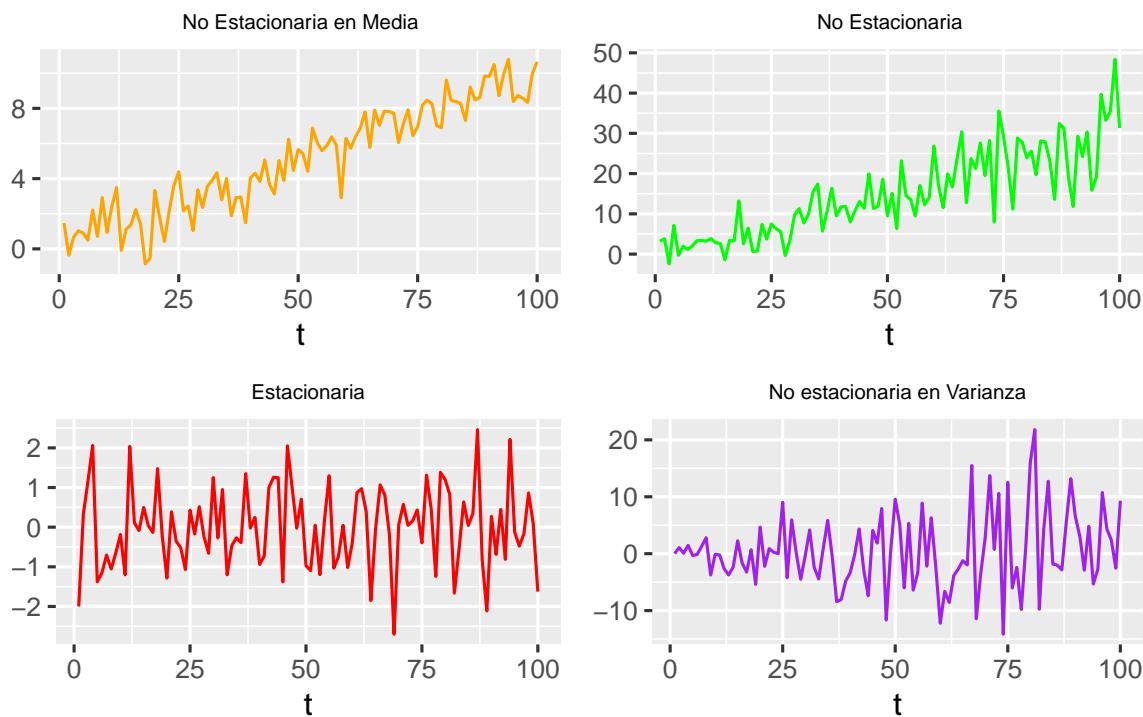
$$E[(y_t - \mu)(y_{t+k} - \mu)] = \gamma_k \quad \text{para todo } t, k$$

Por otro lado, un proceso estocástico es estrictamente estacionario, si es que la distribución multivariante de (y_t, \dots, y_{t+k}) es igual a (y_s, \dots, y_{s+k}) , t, s, k puesto que la

distribución normal está totalmente determinada por su primer y segundo momento.

Figura 3

Serie estacionaria vs no estacionaria.



Nota: Elaboración propia usando R.

Por lo tanto, las series temporales con tendencias o con estacionalidad no son estacionarias: la tendencia y la estacionalidad afectarán el valor de la serie temporal en diferentes momentos. Por otro lado, una serie de ruido blanco es estacionaria: no importa cuándo se observe, debería ser prácticamente igual en cualquier momento. Algunos casos pueden ser confusos: una serie temporal con comportamiento cíclico (pero sin tendencia ni estacionalidad) es estacionaria. Esto se debe a que los ciclos no tienen una duración fija, por lo que antes de observar la serie no podemos estar seguros de dónde se ubicarán los picos y valles de los ciclos. En general, una serie temporal estacionaria no presenta patrones predecibles a largo plazo (Hyndman y Athanasopoulos, 2021).

2.2.3 Test estadísticos de estacionariedad.

Existen pruebas estadísticas formales que permiten contrastar la hipótesis de estacionariedad de una serie de tiempo en la parte regular. Entre las más utilizadas se encuentran la prueba de Dickey-Fuller aumentada (ADF) y la prueba de Kwiatkowski Phillips Schmidt Shin (KPSS).

Test de Dickey-Fuller Aumentado (ADF)

Según Mahadeva y Robinson (2009), esta prueba se utiliza para evaluar la estacionariedad de una serie de tiempo, es decir, si sus propiedades estadísticas —como la media y la varianza— permanecen constantes a lo largo del tiempo. El test ADF es una extensión del test de Dickey-Fuller clásico, que incorpora términos rezagados de la variable dependiente con el fin de corregir por posibles correlaciones seriales en los errores.

El objetivo principal del test es comprobar la presencia de raíces unitarias en la serie. La hipótesis nula (H_0) establece que la serie tiene una raíz unitaria (no estacionaria), mientras que la hipótesis alternativa (H_1) plantea que la serie es estacionaria. Matemáticamente, la prueba se basa en la estimación de una regresión de la forma:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^p \delta_i \Delta y_{t-i} + \varepsilon_t$$

Test de Kwiatkowski-Phillips-Schmidt-Shin (KPSS)

A diferencia del test ADF, el test KPSS se utiliza para evaluar si una serie temporal es estacionaria alrededor de una media o una tendencia determinista. Este test parte de la hipótesis nula de que la serie es estacionaria, ya sea alrededor de una constante (nivel) o de una tendencia lineal. Según Kwiatkowski et al. (1991), si el estadístico de la prueba KPSS es mayor que el valor crítico, se rechaza la hipótesis nula, concluyéndose que la serie no es estacionaria. Por lo tanto, el test KPSS es útil como complemento del ADF, ya que permite verificar la estacionariedad desde un enfoque contrario, ayudando a fortalecer las conclusiones del análisis.

Para el modelo con tendencia determinista, el modelo base es:

$$y_t = \mu_t + \varepsilon_t, \quad \mu_t = \mu_{t-1} + \zeta_t$$

Donde:

- (y_t) : serie observada.
- (μ_t) : tendencia estocástica.
- (ε_t) : término estacionario con media cero.

- (ζ_t) : ruido blanco independiente.

La hipótesis nula del test establece que:

$$H_0 : \zeta_t = 0$$

Es decir, se asume que no hay tendencia estocástica, por lo tanto, la serie es estacionaria alrededor de una constante (si no hay término de tendencia) o de una tendencia determinista (si se incluye una tendencia lineal en la regresión).

Si el estadístico del test KPSS excede el valor crítico tabulado, se rechaza la hipótesis nula, concluyendo que la serie **no es estacionaria**. De esta manera, el test KPSS es especialmente útil cuando se utiliza junto al ADF, ya que ambos evalúan la estacionariedad desde enfoques opuestos, proporcionando un análisis más robusto.

Detección de diferencias mediante la librería ‘forecast’ en R

Hyndman y Athanasopoulos (2021) Creador de la librería `library(forecast)` y `library(fable)` para R, realizo dos funciones para detectar diferencias tanto en la parte regular, y la parte estacional, con ayuda de diferentes test; es importante mencionar que esta librería a la actualidad es una de las mas usadas en el mundo de los pronósticos tanto en el ámbito industrial y académico. A continuación se describe cada uno de ellos.

a) ndiffs: diferenciación no estacional

La función `ndiffs()` utiliza pruebas de raíz unitaria (como ADF, KPSS o PP) para determinar cuántas diferencias no estacionales se necesitan para hacer estacionaria una serie temporal univariada. Dependiendo del test seleccionado, se evalúa si la serie tiene o no raíz unitaria, y se devuelve el número mínimo de diferenciaciones necesarias para lograr la estacionariedad al nivel de significancia especificado.

b) nsdiffs: diferenciación estacional

La función `nsdiffs()` estima cuántas diferencias estacionales se requieren para hacer estacionaria una serie temporal univariada. Para ello, utiliza pruebas de raíz unitaria estacional como SEAS, OCSB, HEGY o CH. Dependiendo del test elegido, se evalúa si existe una raíz unitaria estacional y se devuelve el número mínimo de diferencias estacionales necesarias para alcanzar la estacionariedad. A continuación se detallan las pruebas que usa:

- **SEAS**: Según Wang et al. (2006) este evalúa la fuerza de la estacionalidad mediante una medida basada en la reducción del error de pronóstico; se diferencia si la estacionalidad es fuerte.
- **OCSB (Osborn-Chui-Smith-Birchenhall)**: Osborn et al. (1988) menciona que esta prueba se usa para detectar raíces unitarias estacionales, con hipótesis nula de que existe una raíz unitaria estacional.
- **HEGY (Hylleberg-Engle-Granger-Yoo)**: Hylleberg et al. (1990) realizaron el test basado en regresiones auxiliares para identificar raíces unitarias estacionales en series temporales.
- **CH (Canova-Hansen)**: Canova y Hansen (1995) enfatiza que este test evalúa la presencia de estacionalidad determinística versus estacionalidad estocástica.

2.2.4 Diferenciación

En la Figura 3 , se observa que la gráfica *amarilla, verde y morado* no son estacionarias, sin embargo existe un método llamado diferenciación que convierte una serie temporal no estacionaria en estacionaria.

La serie diferenciada es el cambio entre observaciones consecutivas en la serie original y se puede escribir como:

$$y'_t = y_t - y_{t-1}.$$

La serie diferenciada tendrá solo $T - 1$ valores, ya que no es posible calcular una diferencia y'_t para la primera observación. Cuando la serie diferenciada es ruido blanco, el modelo de la serie original se puede escribir como:

$$y_t - y_{t-1} = \varepsilon_t,$$

dónde ε_t , se denota ruido blanco. Reorganizando esto se obtiene el modelo de “caminata aleatoria”.

$$y_t = y_{t-1} + \varepsilon_t.$$

Los modelos de paseo aleatorio se utilizan ampliamente para datos no estacionarios, en particular datos financieros y económicos. Los paseos aleatorios suelen presentar:

- Largos períodos de tendencias aparentes al alza o a la baja
- cambios repentinos e impredecibles de dirección.

Los pronósticos de un modelo de paseo aleatorio son iguales a la última observación, ya que los movimientos futuros son impredecibles y tienen la misma probabilidad de ser al alza o a la baja (Hyndman y Athanasopoulos, 2021).

Diferenciación de segundo orden

Ocasionalmente, los datos diferenciados no parecerán ser estacionarios y puede ser necesario diferenciarlos una segunda vez para obtener una serie estacionaria:

$$\begin{aligned} y_t'' &= y_t' - y_{t-1}' \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2} \end{aligned}$$

En este caso, y_t' tendrá $t - 2$ Valores, sin embargo en la práctica, casi nunca es necesario ir más allá de las diferencias de segundo orden.

Diferenciación estacional

Una diferencia estacional es la diferencia entre una observación y la observación anterior de la misma temporada. Por lo tanto

$$y_t' = y_t - y_{t-m},$$

dónde m es el número de estaciones. Estas también se llaman “lag- m diferencias”, ya que restamos la observación después de un retraso de m periodos.

Si los datos con diferencias estacionales parecen ser ruido blanco, entonces es necesario un modelo apropiado para los datos originales.

$$y_t = y_{t-m} + \varepsilon_t.$$

Para distinguir las diferencias estacionales de las diferencias ordinarias, a veces nos referimos a las diferencias ordinarias como “primeras diferencias”, es decir, diferencias en el desfase 1.

Diferenciación estacional de segundo orden

Si $y'_t = y_t - y_{t-m}$, denota una serie diferenciada estacionalmente, entonces la serie diferenciada dos veces es:

$$\begin{aligned} y''_t &= y'_t - y'_{t-1} \\ &= (y_t - y_{t-m}) - (y_{t-1} - y_{t-m-1}) \\ &= y_t - y_{t-1} - y_{t-m} + y_{t-m-1} \end{aligned}$$

Al aplicar tanto la estacionalidad como las primeras diferencias, no importa cuál se aplique primero; el resultado será el mismo. Sin embargo, si los datos presentan un patrón estacional marcado, se recomienda aplicar primero la diferenciación estacional, ya que las series resultantes a veces serán estacionarias y no será necesario aplicar otra primera diferencia. Si se aplica primero la primera diferenciación, la estacionalidad persistirá (Hyndman y Athanasopoulos, 2021).

2.2.5 Notación de retroceso “Backshift notation”

El operador de desplazamiento hacia atrás B es un dispositivo de notación útil cuando se trabaja con rezagos de series temporales (Hyndman y Athanasopoulos, 2021):

$$By_t = y_{t-1} .$$

(Algunas referencias utilizan L , para “lag”, en lugar de B , para backshift), en otras palabras B operando en y_t , tiene un efecto de retroceder los datos un período. Dos aplicaciones de B para y_t , desplaza los datos dos períodos hacia atrás:

$$B(By_t) = B^2y_t = y_{t-2} .$$

Para los datos mensuales, si deseamos considerar “el mismo mes del año pasado”, la notación es $B^{12}y_t = y_{t-12}$. El operador de desplazamiento hacia atrás es conveniente para describir el proceso de diferenciación. Una primera diferencia puede escribirse como

$$y'_t = y_t - y_{t-1} = y_t - By_t = (1 - B)y_t .$$

Así que una primera diferencia puede representarse mediante $(1 - B)$. De manera similar, si se deben calcular diferencias de segundo orden, entonces:

$$y_t'' = y_t - 2y_{t-1} + y_{t-2} = (1 - 2B + B^2)y_t = (1 - B)^2 y_t.$$

En general, una diferencia de orden n se puede escribir como

$$(1 - B)^d y_t.$$

La notación de retroceso es particularmente útil al combinar diferencias, ya que el operador puede tratarse mediante reglas algebraicas ordinarias. En particular, los términos que involucran B se pueden multiplicar entre sí.

Por ejemplo, una diferencia estacional seguida de una primera diferencia se puede escribir como

$$(1 - B)(1 - B^m)y_t = (1 - B - B^m + B^{m+1})y_t = y_t - y_{t-1} - y_{t-m} + y_{t-m-1},$$

El mismo resultado que obtuvimos anteriormente.

2.2.6 Familia ARIMA y sus Extensiones

Dentro de los modelos que se basan en procesos estocásticos estacionarios se encuentran los modelos autorregresivos de media móvil de George Box y Gwilym Jenkins que incluyen los modelos ARMA, ARIMA y SARIMA, es importante mencionar que los dos últimos son extensiones de los modelos autorregresivos de media móvil, que mejoran los pronósticos cuando se incluye la componente estacional.

Modelos Autorregresivos (AR)

Hyndman y Athanasopoulos (2021), menciona que, en estos modelos, se puede representar cada observación como una regresión que depende de observaciones anteriores. El punto principal es que el valor actual de la serie se entienda o prediga en función de valores pasados y un término de error. La disposición del modelo indica el número de vistas anteriores consideradas en la relación. Por consiguiente. Así pues, un modelo autorregresivo del orden p se puede escribir como:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

dónde ε_t es ruido blanco. **Esto es como una regresión múltiple pero con valores rezagados como predictores de la misma variable.** A esto lo llamamos AR(p) modelo , un

modelo autorregresivo de orden p.

Los modelos autorregresivos son notablemente flexibles para manejar una amplia gama de patrones de series temporales.

Modelos de Medias Móviles (MA)

Peña (2010) sostiene que estos modelos son aquellos que intentan explicar el valor de una determinada variable en el momento t utilizando términos independientes y una serie de errores convenientes a períodos anteriores, es decir en lugar de utilizar valores pasados de la variable de pronóstico en una regresión, un modelo de promedio móvil utiliza errores de pronóstico pasados en un modelo similar a una regresión.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

dónde ε_t es ruido blanco. Lo llamamos $MA(q)$ modelo , un modelo de media móvil de orden q . Por supuesto, no observamos los valores de ε_t , por lo que no es realmente una regresión en el sentido habitual.

Tenga en cuenta que cada valor de y_t Puede considerarse como una media móvil ponderada de los últimos errores de pronóstico (aunque los coeficientes normalmente no suman uno). Sin embargo, los modelos de media móvil no deben confundirse con el suavizado de media móvil. Un modelo de media móvil se utiliza para pronosticar valores futuros, mientras que el suavizado de media móvil se utiliza para estimar el ciclo de tendencia de valores pasados.

ARIMA (Autoregressive Integrated Moving Average)

Para abordar las Series de Tiempo hay dos enfoques generales, uno de naturaleza determinista, como se observa en métodos como el suavizado exponencial, y otro de naturaleza estocástica, representado por los modelos que se presentarán a continuación. Xu y Chen (2021) sugiere que una extensión de los modelos AR(p) y MA(q) donde ARIMA es el acrónimo de AutoRegressive Integrated Moving Average (en este contexto, «integración» es lo contrario de la diferenciación). El modelo completo puede escribirse como:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

dónde y_t es la serie diferenciada (puede haberse diferenciado más de una vez). Los “predictores” del lado derecho incluyen ambos valores rezagados de y_t y errores rezagados.

A esto lo llamamos $ARIMA(p, d, q)$ modelo , donde:

- p = orden de la parte autorregresiva;
- d = grado de primera diferenciación involucrada;
- q = orden de la parte de media móvil.

Las mismas condiciones de estacionariedad e invertibilidad que se utilizan para los modelos autorregresivos y de promedio móvil también se aplican a un modelo ARIMA.

Una vez que empezamos a combinar componentes de esta manera para formar modelos más complejos, es mucho más fácil trabajar con la notación de retroceso. Por ejemplo, **la ecuación anterior** puede escribirse en notación de retroceso como (Hyndman y Athanasopoulos, 2021).

$$\underbrace{(1 - \phi_1 B - \cdots - \phi_p B^p)}_{AR(p)} \underbrace{(1 - B)^d}_{\text{Diferenciación } d} y_t = c + \underbrace{(1 + \theta_1 B + \cdots + \theta_q B^q)}_{MA(q)} \varepsilon_t$$

El fin principal de un modelo ARIMA es hacer predicciones utilizando información obtenida de datos pasados, estos modelos tienen en cuenta la relación temporal entre observaciones, lo que significa que cada punto de datos depende del valor anterior. Box et al. (2015) recomienda que una serie temporal contenga más de 50 observaciones, y ambos conjuntos de datos analizados en este estudio cumplen con este criterio.

SARIMA (Seasonal Autoregressive Integrated Moving Average)

Los modelos ARIMA también son capaces de modelar una amplia gama de datos estacionales, Un modelo ARIMA estacional se forma incluyendo términos estacionales adicionales en los modelos ARIMA vistos hasta ahora. Se escribe de la siguiente manera (Hyndman y Athanasopoulos, 2021):

$$ARIMA \underbrace{(p, d, q)}_{\text{No estacional}} \underbrace{(P, D, Q)_m}_{\text{Estacional}}$$

Representación en forma extendida:

$$\begin{aligned}
y'_t &= c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} \\
&\quad + \Phi_1 y'_{t-s} + \cdots + \Phi_P y'_{t-Ps} \\
&\quad + \phi_1 \Phi_1 y'_{t-s-1} + \cdots \\
&\quad + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \\
&\quad + \Theta_1 \varepsilon_{t-s} + \cdots + \Theta_Q \varepsilon_{t-Qs} \\
&\quad + \theta_1 \Theta_1 \varepsilon_{t-s-1} + \cdots + \varepsilon_t
\end{aligned}$$

o la misma en operador de desplazamiento hacia atrás

$$\begin{aligned}
&\underbrace{(1 - \phi_1 B)}_{\text{AR no estacional}} \underbrace{(1 - \Phi_1 B^s)}_{\text{AR estacional}} \underbrace{(1 - B)^d}_{\text{Diferenciación}} \underbrace{(1 - B^s)^D}_{\text{Diferenciación estacional}} y_t = \underbrace{(1 + \theta_1 B)}_{\text{MA no estacional}} \underbrace{(1 + \Theta_1 B^s)}_{\text{MA estacional}} \varepsilon_t \\
&\underbrace{(1 - \phi_1 B)}_{\phi_p(B)} \underbrace{(1 - \Phi_1 B^s)}_{\Phi_P(B^s)} \underbrace{(1 - B)^d}_{(1-B)^d} \underbrace{(1 - B^s)^D}_{(1-B^s)^D} y_t = \underbrace{(1 + \theta_1 B)}_{\theta_q(B)} \underbrace{(1 + \Theta_1 B^s)}_{\Theta_Q(B^s)} \varepsilon_t \\
&\phi_p(B) \Phi_P(B^s) (1 - B)^d (1 - B^s)^D y_t = \theta_q(B) \Theta_Q(B^s) \varepsilon_t
\end{aligned}$$

Donde:

- $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$: componente autorregresiva no estacional (AR)
- $\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$: componente autorregresiva estacional (SAR)
- $(1 - B)^d$: diferenciación ordinaria de orden d
- $(1 - B^s)^D$: diferenciación estacional de orden D con frecuencia s
- $\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$: media móvil no estacional (MA)
- $\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}$: media móvil estacional (SMA)
- ε_t : término de error aleatorio (ruido blanco)

dónde m es período estacional (p. ej., número de observaciones por año). Utilizamos notación en mayúsculas para las partes estacionales del modelo y en minúsculas para las no estacionales.

2.2.7 Metodología de Box y Jenkins

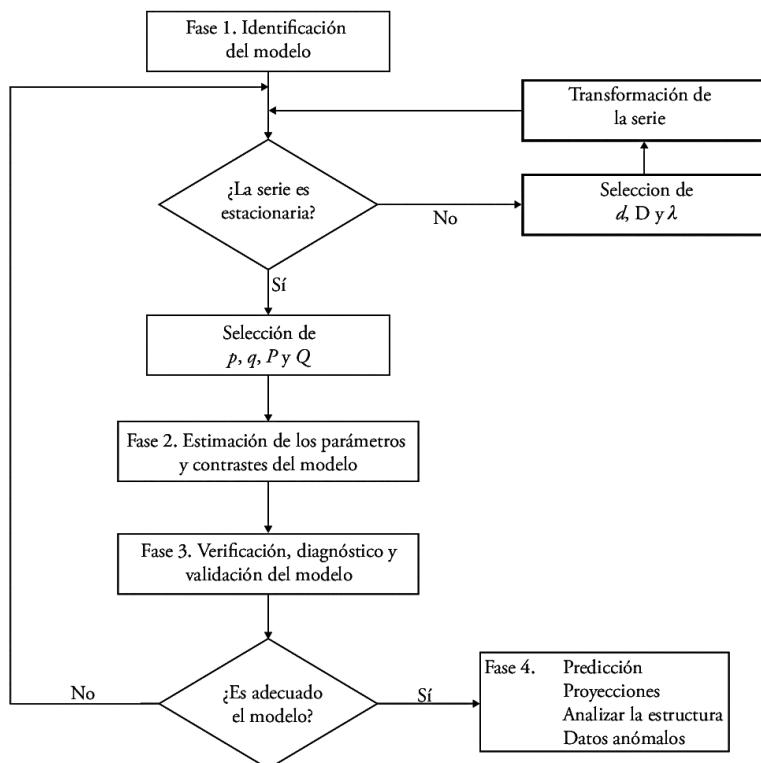
Hyndman y Athanasopoulos (2021), manifiestan que el método Box-Jenkins, lo propuso George Box y Gwilym Jenkins en la década de 1970, y es ampliamente empleado para el análisis de series de tiempo. Este método es un instrumento eficaz para analizar

y predecir datos de series temporales, ofreciendo un método organizado y metódico para el modelado de este tipo de información aplicándose específicamente a modelos como el Autorregresivo de Media Móvil (ARMA), el Modelo Autorregresivo Integrado de Media Móvil (ARIMA), y el Modelo Autorregresivo Integrado de Media Móvil estacional (SARIMA). En ARIMA, ‘AR’ representa un proceso autorregresivo, I denota el orden de integración de la serie de tiempo de acuerdo con su estructura real, y ‘MA’ se refiere al componente de promedio móvil, por otro lado, SARIMA que puede capturar patrones estacionales (S).

El método Box-Jenkins se centra en el ajuste de modelos autorregresivos integrados de media móvil (ARIMA) y sus extensiones estacionales (SARIMA) a series temporales univariadas.

Figura 4

Diagrama de flujo de la metodología Box-Jenkins



Nota: Tomado del artículo elaborado por Rojas et. al (2019).

Su enfoque iterativo permite construir un modelo parsimonioso que capture adecuadamente la dinámica de la serie. Esta metodología se detalla a continuación.

1. Identificación

Primero se debe verificar si la serie es estacionaria en sus dos primeros momentos (media y varianza constante), lo cual se puede comprobar visualmente mediante gráficos de la

serie temporal y validarse con pruebas estadísticas como el test de Dickey-Fuller Aumentado (ADF), la prueba KPSS o la prueba de Phillips-Perron. De acuerdo con Box et al. (2015), la estacionariedad es una condición clave para aplicar modelos ARIMA, ya que garantiza que las propiedades de la serie sean constantes en el tiempo, permitiendo una modelación confiable.

Si la serie no es estacionaria, se aplican diferenciaciones sucesivas hasta lograr la estacionariedad en la media (diferenciación regular) y en la varianza (transformaciones como logaritmos si es necesario). En el caso de que existan patrones estacionales evidentes, también es necesario aplicar diferenciación estacional. Hyndman y Athanasopoulos (2021) recomiendan el uso de funciones como `ndiffs()` y `nsdiffs()` en R para estimar automáticamente el número de diferenciaciones necesarias.

Una vez que se ha logrado la estacionariedad, se procede con la identificación de uno o más modelos candidatos mediante el análisis de los correlogramas de la Función de Autocorrelación (ACF) y la Función de Autocorrelación Parcial (PACF). Estos correlogramas permiten determinar el posible orden de los términos autorregresivos (AR) y de medias móviles (MA), así como sus componentes estacionales (SAR y SMA). A continuación, se detalla cada uno de estos elementos.

Función de autocorrelación simple

Esta función (FAS o ACF) proporciona información sobre su dependencia en términos de linealidad. En otras palabras, la FAS de un proceso estocástico $\{Y_t\}$ consiste en una serie de coeficientes de autocorrelación lineal simple calculados entre la variable aleatoria Y_t en el momento t y la misma variable en un momento anterior $t - j$ (Vásquez, 2023).

$$\rho_{t,t+k} = \frac{\gamma_{t,t+k}}{\text{Var}(Y_t)\text{Var}(Y_{t+k})}, \quad 1, 2, 3, \dots, \quad -1 \leq \rho_{t,t+k} \leq 1$$

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}$$

Las medidas muestrales de esta función de una serie de tiempo es la siguiente:

$$\hat{\rho}_j = \frac{\sum_{i=1}^t (Y_i - \bar{Y})(Y_{i-j} - \bar{Y})}{\sum_{i=1}^t (Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^t Y_i Y_{i-j}}{\sum_{i=1}^t Y_i^2}$$

Función de autocorrelación parcial

La función (PACF) de un proceso estocástico $\{Y_t\}$ es la colección de los coeficientes de autocorrelación parcial de la variable aleatoria Y_t , medido entre en el momento t y el momento $t - j$, luego de eliminar el efecto que tienen los rezagos intermedios $Y_1, Y_2, \dots, Y_{t-j-1}$ sobre Y_t . Las medidas muestrales de la función de autocorrelación parcial, ϕ_{jj} se obtienen del coeficiente asociado a la variable $t - j$ en una regresión de y_t contra sus rezagos hasta y_{t-j} donde la variable aleatoria y_t se mide en desviaciones con respecto a su media, esto es, $y_t = Y_t - \bar{Y}$ (Vásquez, 2023).

$$\phi_{jj} = \text{Corr}(y_t, y_{t-k} | y_{t-1}, \dots, y_{t-k+1}) = \frac{\text{Cov}((y_t - \hat{y}_t)(y_{t-k} - \hat{y}_{t-k}))}{\sqrt{\text{Var}(y_t - \hat{y}_t) \cdot \text{Var}(y_{t-k} - \hat{y}_{t-k})}}$$

Esta expresión formaliza el concepto de autocorrelación parcial como una correlación condicional: mide la relación directa entre y_t y su rezago y_{t-k} , descontando los efectos lineales de los rezagos intermedios $y_{t-1}, \dots, y_{t-k+1}$. En este sentido, \hat{y}_t y \hat{y}_{t-k} representan las partes explicadas de y_t y y_{t-k} a partir de los rezagos intermedios, por lo que la covarianza entre los residuos captura exclusivamente la dependencia neta entre los dos extremos del rezago. Esta propiedad hace que la PACF sea una herramienta fundamental para la identificación del orden p en modelos $AR(p)$, ya que los valores de ϕ_{jj} tienden a ser significativamente distintos de cero hasta el rezago p , y cercanos a cero a partir de allí.

Correlograma

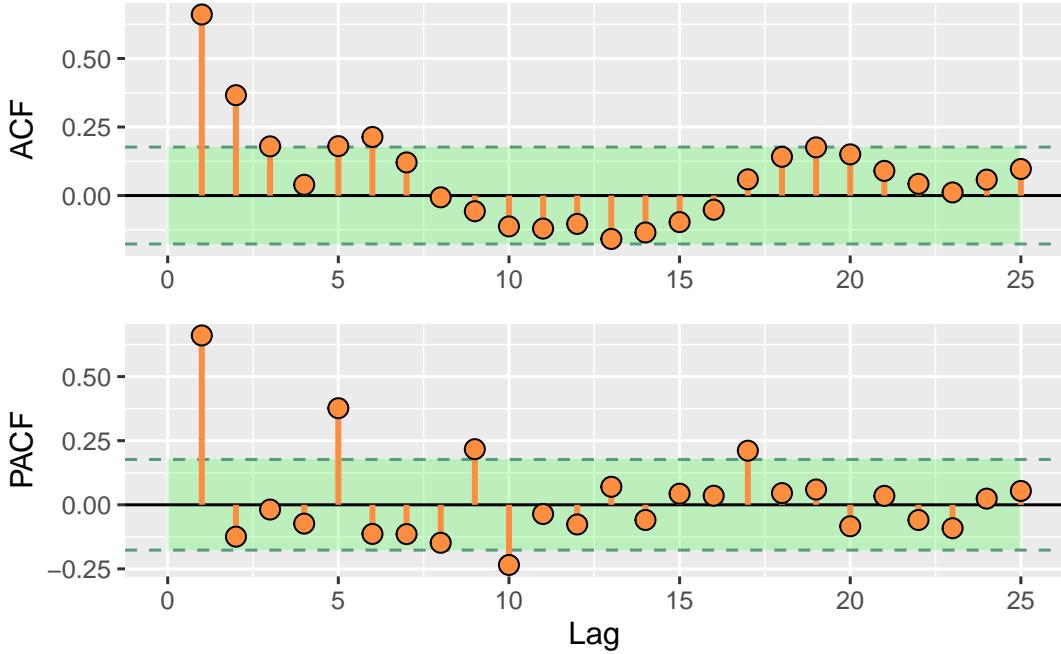
Según Amat Rodrigo y Escobar Ortiz (2025) un correlograma es una representación gráfica de las correlaciones entre una observación de una serie temporal y sus distintos rezagos. En el contexto del análisis de series temporales, estos gráficos corresponden a las funciones de autocorrelación simple (ACF) y autocorrelación parcial (PACF), las cuales son herramientas fundamentales en la fase de identificación de la metodología Box-Jenkins.

A través de los correogramas es posible detectar la presencia o ausencia de patrones aleatorios en los datos, evaluar si la serie presenta ruido blanco o evidencia de estructura dependiente, y sobre todo, proponer uno o más modelos candidatos ARIMA.

De este modo, la observación conjunta de las funciones ACF y PACF permite al investigador formular hipótesis sobre la estructura del modelo (valores de p, d y q), que serán luego evaluadas en la fase de estimación y diagnóstico.

Figura 5

Ejemplo de correlograma ACF y PACF



Nota: Elaboración propia usando R.

2. Estimación

Una vez seleccionados los posibles órdenes del modelo mediante la fase de identificación, se procede a estimar los parámetros del modelo ARIMA. Según Hyndman y Athanasopoulos (2021), esta estimación puede realizarse a través de **máxima verosimilitud (MLE)** o **mínimos cuadrados condicionales (CLS)**, dependiendo del software y del contexto.

Los parámetros que se estiman incluyen:

- (ϕ_1, \dots, ϕ_p) : coeficientes autorregresivos.
- $(\theta_1, \dots, \theta_q)$: coeficientes de medias móviles.
- $(\Phi_1, \dots, \Phi_P, \Theta_1, \dots, \Theta_Q)$: componentes estacionales (si aplica).
- (μ) o c : constante del modelo (si es incluida).

Maxima verosimilitud

La estimación por máxima verosimilitud busca encontrar los parámetros que maximicen la probabilidad de que el modelo haya generado la serie observada. Dado un conjunto de datos (y_1, y_2, \dots, y_T) , se definen los residuales (ε_t) en función de los parámetros, y se maximiza:

$$L(\theta) = \prod_{t=1}^T f(\varepsilon_t | \theta)$$

Donde (θ) representa el vector de parámetros. Bajo supuestos de normalidad, esto equivale a minimizar la suma de los errores cuadráticos:

$$\sum_{t=1}^T \varepsilon_t^2$$

Este enfoque tiene la ventaja de aprovechar toda la estructura del modelo e incorporar correlaciones entre observaciones, siendo además asintóticamente eficiente.

Mínimos Cuadrados Condicionales (CLS)

El método de mínimos cuadrados condicionales (CLS, por sus siglas en inglés) se basa en minimizar la suma de los errores al cuadrado, condicionados a los primeros valores observados de una serie temporal. Es particularmente útil en la estimación de modelos autorregresivos (AR), de media móvil (MA) o mixtos (ARMA, ARIMA, y SARIMA), ya que permite obtener estimadores sin necesidad de suposiciones completas sobre la distribución de los errores. Aunque en términos asintóticos suele ser menos eficiente que el método de máxima verosimilitud (MLE), el CLS destaca por su simplicidad computacional y su estabilidad numérica, especialmente cuando se requiere una estimación inicial para algoritmos iterativos.

Además, el método CLS ofrece consistencia bajo condiciones generales, y sus estimadores tienden a ser insesgados en muestras grandes. En modelos ARMA, este método estima los parámetros condicionando la función de verosimilitud a los primeros valores conocidos, lo cual evita problemas de optimización no lineal que surgen al utilizar métodos como MLE desde el inicio. Por esta razón, CLS se emplea frecuentemente como paso preliminar antes de realizar una estimación más refinada mediante MLE o métodos bayesianos.

3. Diagnóstico

Según Villavicencio (2007), el diagnóstico de modelos de Box-Jenkins, se esperan residuos de innovación cumplir con ciertas propiedades:

1. Los residuos de innovación deben ser independientes entre sí. Si se observa alguna

correlación entre ellos, esto sugiere que aún queda información relevante en los residuos que podría aprovecharse para mejorar los pronósticos.

2. La media de los residuos de innovación debe ser cero. Si la media es diferente a cero, las predicciones estarán sesgadas.

En caso de que alguna de estas dos propiedades no se cumpla, es necesario ajustar el enfoque de pronóstico para lograr resultados más precisos. Hyndman y Athanasopoulos (2021) enfatizan que aparte de estas propiedades fundamentales, es beneficioso (aunque no imprescindible) que los residuos presenten también las siguientes dos características:

- a. La varianza de los residuos de innovación debe ser constante, lo que se denomina “homocedasticidad”.
- b. Los residuos de innovación deben seguir una distribución normal.

Estas dos características simplifican el proceso de cálculo de los intervalos de predicción

4. Predicción

En esta última etapa, se utiliza el modelo ajustado para realizar predicciones de valores futuros. No solo se deben generar estimaciones puntuales, sino también calcular intervalos de confianza que cuantifiquen la incertidumbre asociada a las predicciones, puesto que una correcta interpretación de los intervalos de predicción permite tomar decisiones más informadas y gestionar mejor los riesgos asociados a la variabilidad futura.

Adicionalmente se debe tener cuidado con los supuestos de los residuos del modelo, pues según Hyndman y Athanasopoulos (2021) manifiestan que si la suposición de una distribución normal de los residuos es poco razonable, una alternativa es el uso del método **Bootstrap**, este método no requiere que los residuos sigan una distribución normal, sino que solo asume que los residuos no están correlacionados y tienen varianza constante, Bootstrap es una técnica de remuestreo que permite generar intervalos de predicción y estimaciones de incertidumbre sin depender de suposiciones estrictas sobre la distribución de los residuos.

Como recomendaciones adicionales, una vez validado que el modelo se ajusta adecuadamente a los datos —a través del análisis de residuos y pruebas de independencia—, se procede a realizar las predicciones, posteriormente, se evalúa la calidad de estas predicciones utilizando métricas de error como MAE, RMSE, MASE, MAPE, etc; además, como describe Hyndman y Athanasopoulos (2021) es recomendable emplear validación

fuerza de muestra para medir la capacidad predictiva real del modelo, permitiendo la identificación de posibles limitaciones y fortalezas.

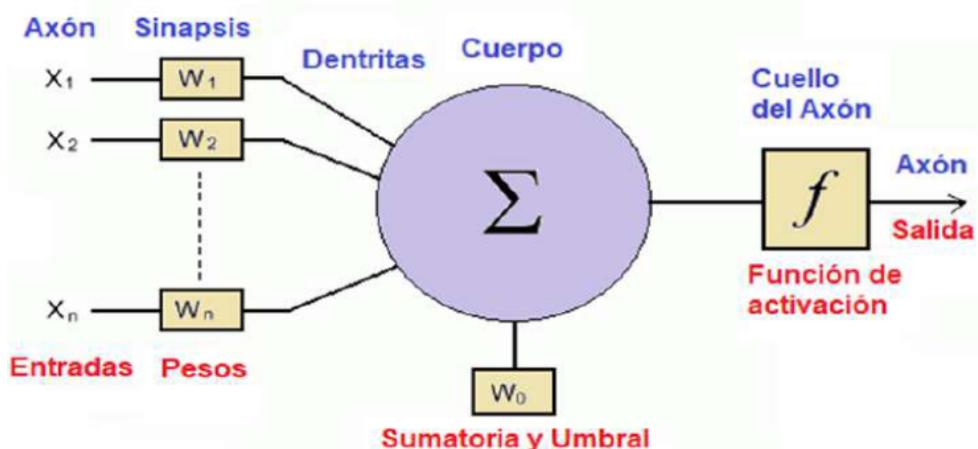
2.2.8 Redes Neuronales Artificiales

Según Kelleher (2019), las (RNA) representan la faceta esencial de la IA, cuyos comienzos se remontan a la década de los años 40, cuando McCulloch y Pitts sentaron las bases de los primeros modelos de este tipo de redes. Estas estructuras computacionales, se desarrollaron basándose en las funciones del cerebro humano y desde entonces han ido evolucionado significativamente, dando lugar a una extensa gama de aplicaciones en campos tan variados como la visión por ordenador o el análisis del lenguaje humano y la robótica.

Rojas (2022) manifiesta que las neuronas son células altamente especializadas del sistema nervioso que reciben, procesan y transmiten información mediante señales eléctricas y químicas. Por tanto, su función principal es recibir estimulación y conducir impulsos nerviosos llamados potenciales de acción, a través de las conexiones llamadas sinapsis. La estructura de una neurona comprende varias partes distintivas. Las dendritas son responsables de recibir las señales provenientes de otras neuronas, mientras que el cuerpo celular, o soma, integra estas señales para generar la respuesta neuronal. Esta respuesta se manifiesta como una señal eléctrica, el potencial de acción, que además coordina las actividades metabólicas de la célula. Por otro lado, el axón es la prolongación alarga de la neurona que conduce los potenciales de acción a distancias considerablemente largas. Finalmente, las sinapsis son los puntos de conexión entre neuronas, permitiendo la transmisión de señales de una célula a otra.

Figura 6

Neurona Artificial genérica y su correspondencia biológica.



Nota: Tomado del artículo de Castañeda et al. (2023)

La neurona artificial desempeña un papel crucial en el procesamiento al actuar como un mecanismo de activación que genera una salida basada en la información recibida. Esta función de activación se encuentra en todas las capas de las redes neuronales artificiales, y su tarea es fundamental para la transformación de las entradas en resultados significativos.

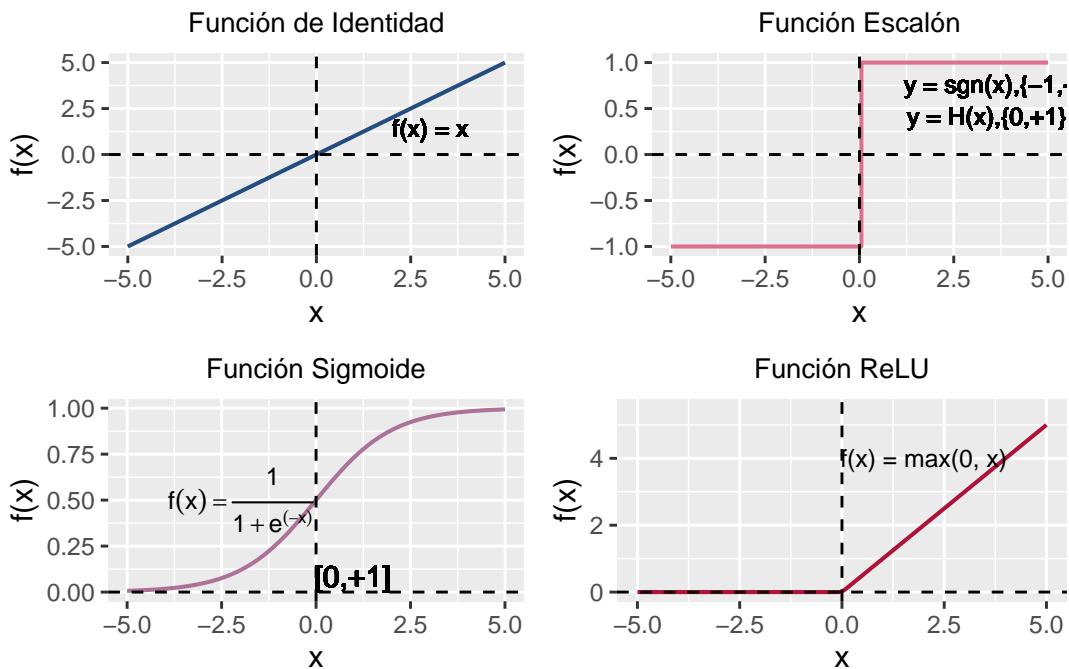
En la misma línea que se exhibe en la Figura 6, la analogía entre la neurona biológica y la artificial, subrayando su estrecha similitud en función. Tal como señala Fernández Salguero (2021), “esta similitud radica en la emulación de la habilidad de aprendizaje humana, lo que permite a las RNA automatizar la adquisición de múltiples reglas de aplicación”.

A continuación, Castañeda et al. (2007) describen cada una de ellas:

- Entradas: Representan los datos o señales que ingresan a la neurona artificial, ya sea información original o salidas de otras neuronas de la red.
- Pesos: Cada entrada está vinculada a un peso que determina la relevancia o influencia de esa entrada en la neurona. Estos pesos son análogos a la eficacia sináptica en las neuronas biológicas.
- Sumatoria y Umbral: Dentro de la neurona, se suman las entradas ponderadas por sus respectivos pesos. A esta suma se le resta un valor umbral propio de la neurona. Esta operación modela la activación de la neurona, similar al potencial postsináptico en las neuronas biológicas.
- Función de activación: La señal resultante de la sumatoria y umbral se procesa mediante una función de activación, también llamada función de transferencia. Esta función determina la salida de la neurona y puede introducir no linealidades en el modelo, permitiendo a la red neuronal capturar patrones complejos en los datos de entrada. Es importante recordar que existen varias funciones de activación, a continuación, se presentan algunas de ellas:

Figura 7

Principales funciones de activación para redes neuronales artificiales.



Nota: Elaboración propia en R.

Finalmente, Maya (2022) manifiesta que a los modelos basados en RNA se les considera modelos de caja negra, dado que su funcionamiento se fundamenta en la observación de datos y la generación de respuestas sin una comprensión completa de cómo se llega a esas respuestas. Como se mencionó anteriormente, las redes neuronales artificiales son capaces de capturar patrones complejos en los datos de entrada mediante funciones de activación que pueden ser no lineales, sin embargo, la forma exacta en que estas redes procesan la información y generan resultados puede ser difícil de interpretar, ya que no se puede trazar una línea directa entre las entradas y las salidas.

Este enfoque de caja negra puede ser tanto una ventaja como una limitación. Por un lado, permite a las redes neuronales adaptarse y comprender de manera flexible a partir de los datos disponibles, lo que las hace extremadamente útiles en una amplia gama de aplicaciones. Por otro lado, la opacidad en el proceso de decisión puede plantear desafíos en términos de interpretación.

A pesar de estas limitaciones, los modelos de caja negra basados en redes neuronales artificiales siguen siendo herramientas valiosas en campos como el reconocimiento de patrones, la predicción y la clasificación, donde la habilidad para comprender relaciones complejas en los datos es esencial.

2.2.9 Algunas RNA'S Especializadas en Series de Tiempo

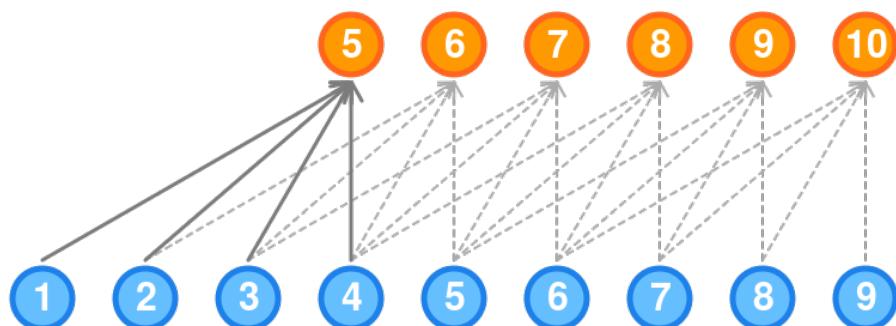
En el ámbito de las redes neuronales, hay una amplia gama de tipos, pero solo algunas se centran en el análisis de datos secuenciales, como las series temporales. A continuación, mencionaremos las más destacadas:

Redes neuronales autorregresivas (NNAR)

En el entorno del análisis de series temporales, la clave para utilizar este tipo de redes radica en considerar los datos secuenciales de la serie temporal como entradas de la red, de manera similar a como se utilizan los valores retrasados en un modelo de autorregresión lineal, de esta manera, Hyndman y Athanasopoulos (2021) manifiestan que nos encontramos con los modelos autoregresivos para la red, conocidos como NNAR (Redes Neuronales Autorregresivas). Estos modelos son unidireccionales y constan de una única capa oculta. Por ejemplo, un modelo NNAR (9,5) se refiere a una red neuronal que utiliza las últimas nueve observaciones como entradas para predecir la salida, y cuenta con cinco neuronas en la capa oculta, además, es trascendental destacar que un modelo NNAR (p,q) es similar a un modelo ARIMA (p,q), pero sin las restricciones en los parámetros que aseguran la estacionariedad. En otras palabras, los modelos NNAR tienen la ventaja de la no linealidad, lo que les permite capturar relaciones más complejas entre las variables en comparación con los modelos ARIMA tradicionales.

Figura 8

Arquitectura de una NNAR básica.



Nota: Nota. Tomado de Mañas (2019)

Las redes más simples no contienen capas ocultas y son equivalentes a las regresiones lineales. La Figura 8 muestra la versión de red neuronal de una regresión lineal con cuatro predictores. Los coeficientes asociados a estos predictores se denominan “ponderaciones”. Los pronósticos se obtienen mediante una combinación lineal de las entradas. Las ponderaciones se seleccionan en el marco de la red neuronal mediante un “algoritmo de aprendizaje” que minimiza una “función de coste” como el MSE. Por supuesto, en este

sencillo ejemplo, podemos utilizar la regresión lineal, que es un método mucho más eficiente para entrenar el modelo, sin embargo si añadimos una capa intermedia con neuronas ocultas, la red neuronal se vuelve no lineal.

Aunque los modelos NNAR son relativamente simples en su arquitectura (una sola capa oculta), sientan las bases para arquitecturas más complejas como las redes LSTM o GRU, que también trabajan con secuencias temporales pero están diseñadas para manejar dependencias de largo plazo.

Redes Neuronales Recurrentes (RNN)

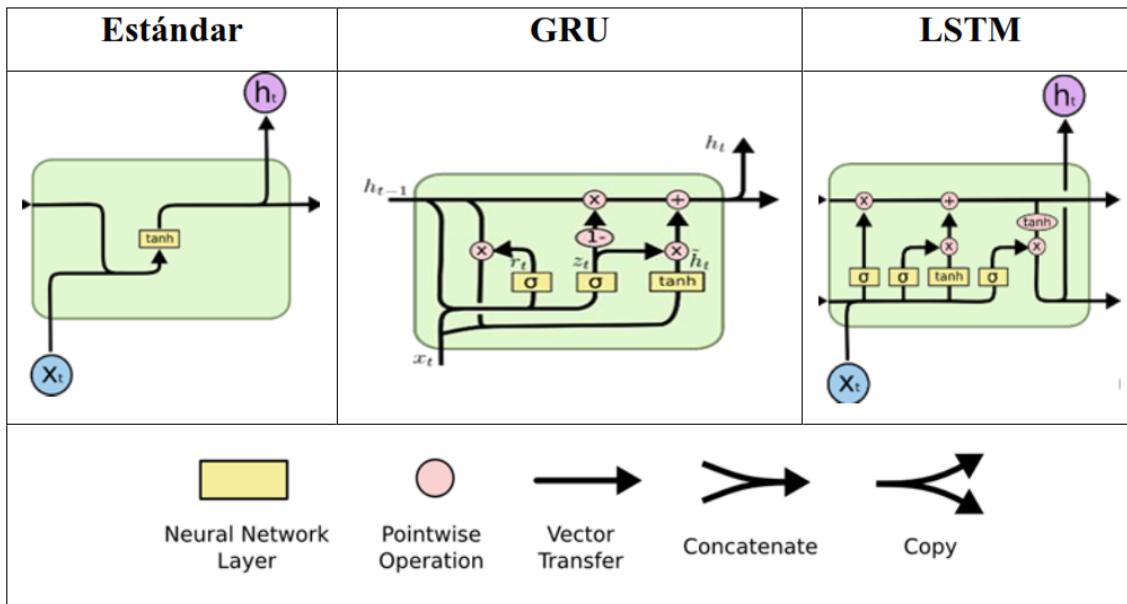
Según Olah (2015), las Redes Neuronales Recurrentes (RNN) son un tipo de red neuronal que pueden modelar secuencias de datos utilizando conexiones recurrentes en su arquitectura. Estas conexiones les permiten mantener y actualizar un estado interno que codifica información sobre la historia pasada de la secuencia. Sin embargo, el RNN tradicional presenta dificultades para procesar secuencias largas debido al desvanecimiento del gradiente o problemas de explosión.

Por consiguiente Hesaraki (2023) afirma que para tratar este problema, se introdujeron variantes de las RNN, como las Unidades Recurrentes con Compuertas (GRU) y las Memorias a Corto y Largo Plazo (LSTM). Las GRU son una versión simplificada de las LSTM que combinan algunas de las compuertas de las LSTM en una sola estructura, lo que las hace más eficientes computacionalmente. Sin embargo, las GRU pueden tener dificultades para capturar dependencias a largo plazo en secuencias muy largas debido a su diseño más simple.

Por otro lado, LSTM es una extensión más compleja de RNN que resuelve el problema del desvanecimiento del gradiente introduciendo una estructura de memoria más compleja. Los LSTM utilizan tres puertas diferentes (entradas, salida y olvido) para comprobar el flujo de información en la red y pueden mantener estados de memoria a largo plazo, lo que les permite capturar dependencias de secuencias a largo plazo.

Figura 9

Estructura de los 3 tipos de RNN's



Nota: Adaptado de Olah (2015).

Dada la configuración más sofisticada y los resultados óptimos que suelen proporcionar las LSTM, nos centraremos en explorar esta opción más detalladamente en la siguiente sección.

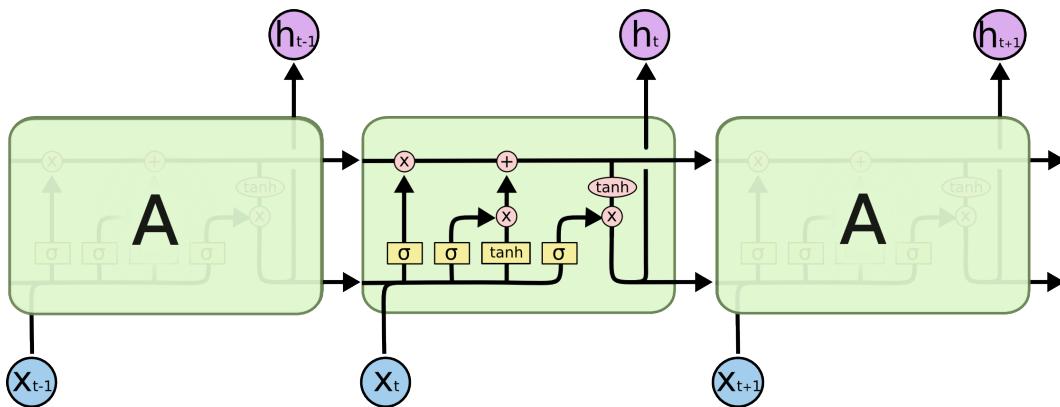
2.2.10 Long-Short Term Memory (LSTM)

Las redes de almacenamiento de información a largo plazo, conocidas como LSTM, constituyen una variante especializada de las redes neuronales recurrentes (RNN), elaboradas para identificar y asimilar dependencias temporales prolongadas en conjuntos de datos secuenciales. Introducidas inicialmente por Hochreiter y Schmidhuber (1997), estas redes han sido objeto de refinamiento y difusión por parte de numerosos investigadores en trabajos posteriores relacionadas a series de tiempo.

Por otro lado, Olah (2015) manifiesta que los LSTM siguen una disposición similar en forma de cadena, pero se distinguen por la estructura de su módulo repetido. En contraste con las RNN estándar que emplean una única capa neuronal, los LSTM presentan una configuración más compleja, constando de cuatro capas que interactúan de manera peculiar y especial.

Figura 10

Módulo repetido en un LSTM con cuatro capas en interacción



Nota: Adaptado de Olah (2015).

De la Figura 10, Olah (2015), manifiesta que cada conexión en la red neuronal transporta consigo un vector completo, que se envía desde la salida de un nodo hacia las entradas de otros nodos. Los nodos, representados como círculos de tono rosado, llevan a cabo operaciones elementales como la suma de vectores, mientras que los cuadros amarillos denotan las capas de redes neuronales que han sido aprendidas. Las líneas que conectan los nodos indican concatenación, lo que significa que dos o más vectores se unen para formar uno solo. Además, una línea que se ramifica señala que su contenido es duplicado, con las copias dirigidas a diferentes destinos.

Por consiguiente, Soria et al. (2022) manifiesta que el concepto central se encuentra en el estado de la celda, representado a través de la línea horizontal que cruza la parte superior del diagrama en la figura 14, esta entidad, equivalente a una cinta transportadora, que atraviesa toda la cadena con solo algunas interacciones lineales menores. De esta manera, se facilita el flujo de información sin alteraciones significativas, lo que permite que la información se mantenga intacta a lo largo del tiempo y de las distintas etapas de procesamiento.

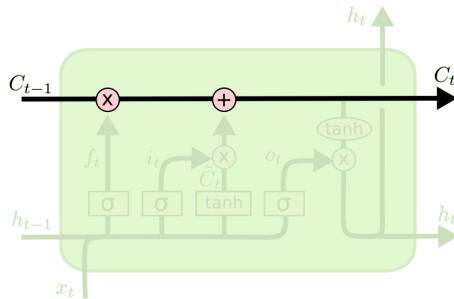
Así mismo Kelleher (2019) en su libro de “Deep learning” describe que el LSTM presenta la habilidad de controlar cuidadosamente la adición o eliminación de información en el estado de la célula, mediante la regulación de estructuras conocidas como puertas. Estas puertas actúan como mecanismos para permitir opcionalmente el paso de información. Están compuestas por una capa de red neuronal sigmoidea, que determina qué información debe ser olvidada o mantenida, y una operación de multiplicación puntual, que regula cómo la información filtrada se integra con el estado de la célula.

La idea central detrás de los LSTM

La clave de los LSTM es el estado de la celda, la línea horizontal que recorre la parte superior del diagrama .El estado celular es como una cinta transportadora. Recorre toda la cadena en línea recta, con solo algunas interacciones lineales menores. Es muy fácil que la información fluya por ella sin cambios.

Figura 11

Cinta transportadora

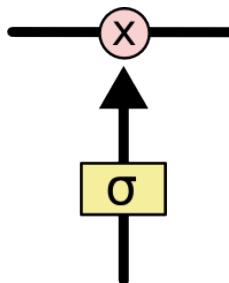


Nota: Adaptado de Olah (2015).

El LSTM tiene la capacidad de eliminar o agregar información al estado celular, cuidadosamente regulada por estructuras llamadas puertas.Las puertas permiten el paso opcional de información. Se componen de una capa de red neuronal sigmoidea y una operación de multiplicación puntual (Olah, 2015).

Figura 12

Compuertas en la LSTM



Nota: Adaptado de Olah (2015).

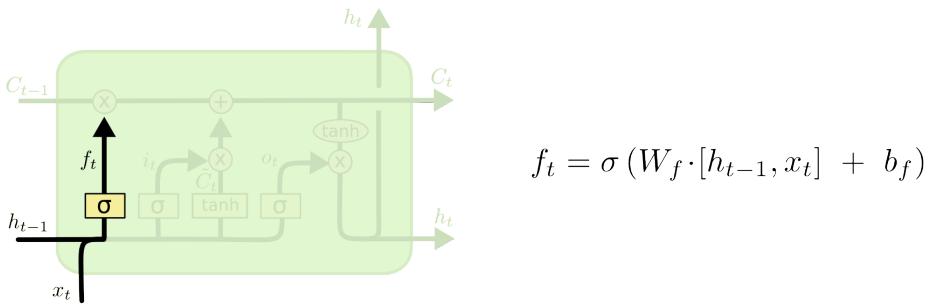
La capa sigmoidea produce valores en el rango de cero a uno, los cuales indican la proporción de cada componente que se permite pasar. Tal como lo describe Olah (2015) en la Figura 12, manifestando que un valor de cero denota la ausencia de paso de cualquier cosa, mientras que un valor de uno indica que todos los componentes pueden pasar. En un LSTM, se emplean tres de estas puertas para salvaguardar y regular el estado de la célula, garantizando así un control preciso sobre la información que fluye dentro de la red y contribuyendo a su capacidad para capturar y aprender dependencias temporales a largo plazo en los datos.

2.2.11 Recorrido general que realiza una LSTM

La red neuronal recurrente de tipo LSTM, normalmente realiza un recorrido que sigue los siguientes pasos: Paso 1: Olah (2015) menciona que en el LSTM, la capa de puerta de olvido, representada por una capa sigmoidea, determina qué información se descarta del estado de la celda anterior. Esta capa genera un valor entre 0 y 1 para cada componente en el estado de la celda anterior, donde 1 indica que se mantiene completamente la información y 0 indica que se elimina por completo. En un modelo de lenguaje, por ejemplo, el estado de la celda puede contener información sobre el género del sujeto actual, y la capa de puerta de olvido permite olvidar este género cuando se encuentra un nuevo sujeto.

Figura 13

Primer recorrido de la LSTM estándar

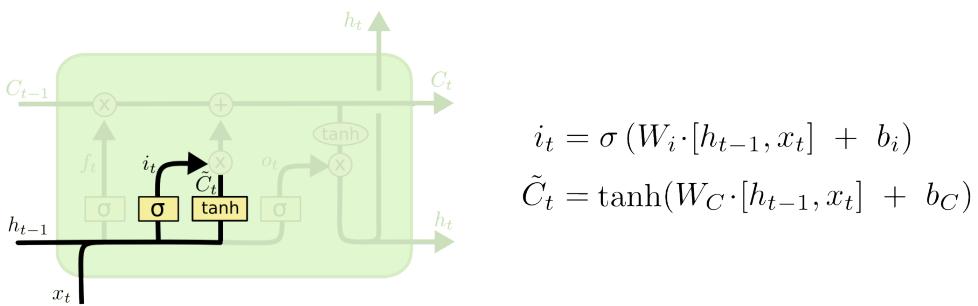


Nota: Adaptado de Olah (2015).

Paso 2: En el proceso siguiente del LSTM, se determina qué información adicional se añadirá al estado de la celda. En esta etapa Olah (2015) manifiesta que consta de dos partes: en primer lugar, una capa sigmoidea denominada “capa de puerta de entrada” o update gate, decide qué valores se actualizarán; luego, una capa “tanh” genera un vector de nuevos valores candidatos \tilde{C}_t , que podría sumarse al estado actual. Posteriormente, combinamos estos dos componentes para crear una actualización del estado.

Figura 14

Segundo recorrido de la LSTM estándar



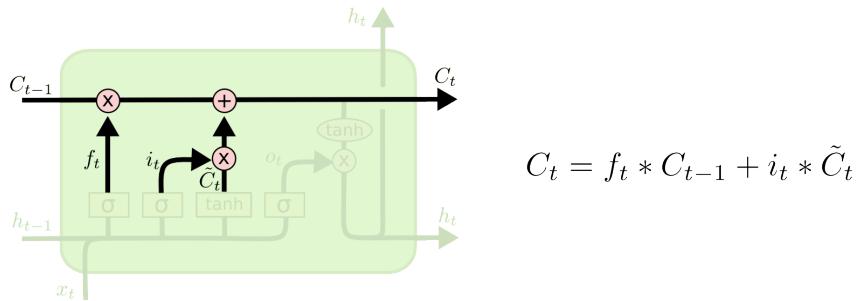
Nota: Adaptado de Olah (2015).

Paso 3: En este paso, actualizamos el estado anterior de la celda, C_{t-1} , al nuevo

estado celular C_t . Los pasos previos ya han determinado qué acciones tomar, por lo que solo debemos ejecutarlas. Para ello, Olah (2015) manifiesta que se multiplica el estado anterior por f_t , que representa la decisión de olvidar la información previa. Posteriormente, sumamos $i_t * \tilde{C}_t$, que son los nuevos valores candidatos, ajustados según la importancia asignada a cada componente del estado.

Figura 15

Tercer recorrido de la LSTM estándar

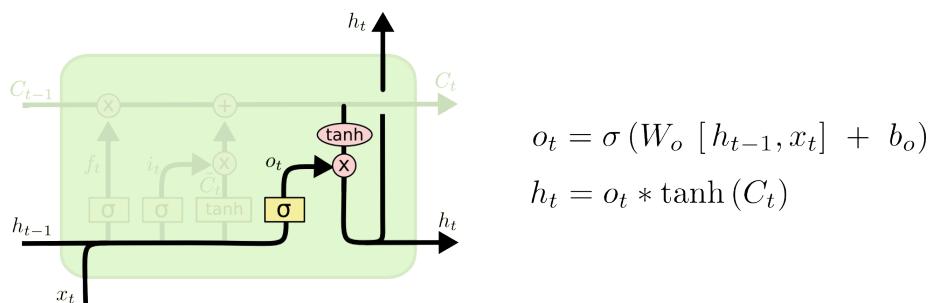


Nota: Adaptado de Olah (2015).

Paso 4: Finalmente, en el proceso de generación, se determina qué información generar utilizando como base el estado de la celda, aunque de manera filtrada. Por consiguiente tal como menciona Olah (2015), primero se aplica una capa sigmoidea para determinar qué partes del estado de la celda serán empleadas en la generación. Luego, se pasa el estado de la celda por una función tanh (para restringir los valores entre -1 y 1) y se multiplica a través de la salida de la compuerta sigmoide, lo que permite generar únicamente las partes seleccionadas.

Figura 16

Cuarto recorrido de la LSTM estándar



Nota: Adaptado de Olah (2015).

Variantes en la estructura de la LSTM

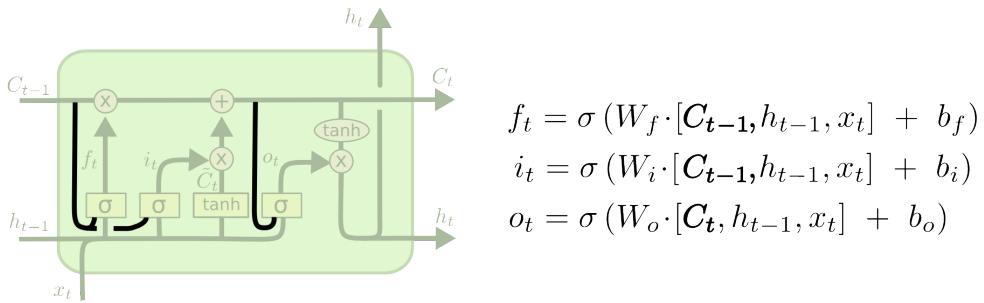
Lo descrito hasta ahora representa un LSTM estándar. Sin embargo, es importante destacar que no todos los LSTM son idénticos entre sí. De hecho, la mayoría de los trabajos relacionados con LSTM emplean versiones ligeramente modificadas. Aunque las diferencias

pueden ser sutiles, es relevante mencionar algunas de ellas.

Variante 1: Olah (2015), resalta que una variante popular de LSTM, implica la inclusión de “conexiones de mirilla”. Esto implica que las capas de las puertas tienen acceso para observar el estado actual de la celda.

Figura 17

Primera variante de la LSTM

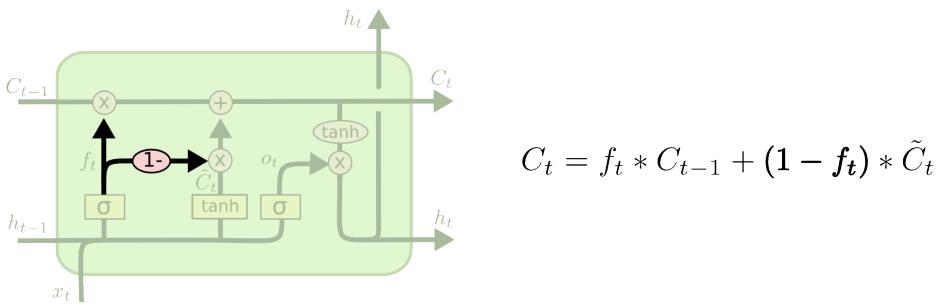


Nota: Adaptado de Olah (2015).

Variante 2: Una variación adicional implica la utilización de puertas de entrada y olvido acopladas. Tal como expone Olah (2015), en lugar de tomar decisiones separadas sobre qué olvidar y qué agregar, estas decisiones se realizan de manera conjunta. Es decir, el proceso de olvido ocurre únicamente cuando se va a introducir nueva información en su lugar, y la actualización del estado ocurre solamente cuando se elimina información más antigua.

Figura 18

Segunda variante de la LSTM

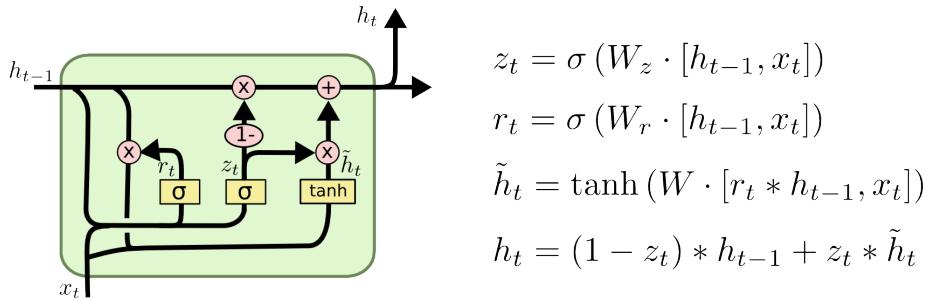


Nota: Adaptado de Olah (2015).

Variante 3: Una variación notable del LSTM es la Unidad Recurrente Cerrada (GRU), introducida por Cho et al. (2014). En este modelo, las puertas de entrada y olvido se combinan en una única “puerta de actualización”. Así mismo, se fusionan el estado de la celda y el estado oculto, y se realizan algunos ajustes adicionales. Como resultado, el modelo GRU es más simple que los LSTM estándar y ha ganado popularidad en diversos campos.

Figura 19

Tercera variante de la LSTM



Nota: Adaptado de Olah (2015).

2.2.12 Alternativas de pronósticos usando Redes LSTM

Según los datos disponibles y los objetivos de la investigación, las redes neuronales LSTM, constan con 3 principales modos de pronósticos, los cuales están en base al horizonte que se requiere pronosticar, y las variables que se deseen incluir al modelo, por ello dentro de las LSTM tenemos:

Modelos univariados unistep

Según Sotaquirá (2023) son el enfoque más básico para hacer predicciones utilizando redes LSTM. En esta configuración, empleamos una sola variable como entrada del modelo y obtenemos una única variable como salida. La predicción se realiza únicamente para un único paso en el futuro dentro de la serie temporal. Además, es común que la variable de entrada sea la misma que se desea predecir.

Modelos multivariados unistep

Sotaquirá (2023) manifiesta que para este caso se introducen múltiples variables en la red neuronal, y la predicción se realiza para una única variable en particular, tan solo para un momento futuro dentro de la serie temporal.

Modelos univariados multistep

En los modelos univariados-multistep, Sotaquirá (2023) resalta que se introduce una sola variable en el modelo (normalmente la misma variable que se desea predecir) y buscamos pronosticar el comportamiento de esa única variable durante varios momentos en el tiempo

Modelos multivariados multistep

Sotaquirá (2023) subraya que, en esta configuración, se introduce diversas variables

en el modelo, con el objetivo de predecir una única variable específica durante varios instantes en el futuro. Las configuraciones mencionadas previamente son las preferidas y ampliamente empleadas para realizar predicciones en series de tiempo. Sin embargo, existen opciones menos convencionales que reciben menos atención, como la inclusión de múltiples variables de salida (tales como pronósticos de temperatura, humedad y presión) o la predicción del comportamiento futuro de estas variables múltiples a lo largo de uno o varios períodos de tiempo.

Finalmente Hesaraki (2023), menciona los desafíos y problemas comunes con los LSTM:

- a) Complejidad: Los LSTM son más complejos que las RNN tradicionales, lo que puede resultar en tiempos de entrenamiento más largos y mayores requisitos computacionales.
- b) Dificultad con Secuencias Extremadamente Largas: Aunque los LSTM son mejores para capturar dependencias a largo plazo que las RNN tradicionales, aún pueden tener dificultades con secuencias extremadamente largas.
- c) Sobreajuste: Como otras redes neuronales, los LSTM son susceptibles al sobreajuste, especialmente cuando se emplean conjuntos de datos reducidos.
- d) Dificultad con Conjuntos de Datos Pequeños: Los LSTM pueden requerir grandes cantidades de datos de entrenamiento para funcionar correctamente.
- e) Desafíos de Entrenamiento: Entrenar LSTM's puede ser desafiante, especialmente al optimizar hiperparámetros o lidiar con problemas de gradiente desvaneciente.

Implementación de redes neuronales recurrentes LSTM

Existen diversas formas de implementar una red neuronal LSTM, sin embargo según el artículo consultado de Pang (2021), donde se realiza un pronóstico incluyendo más variables (Modelo multivariados multistep) junto con otras investigaciones, generalmente su implementación conlleva lo siguiente:

- a) Carga y Visualización de Datos: Lectura del conjunto de datos que contiene variables independientes (x_1, x_2) y la variable dependiente (y), además se realiza una visualización de los primeros lotes de datos para comprender la estructura y distribución.
- b) Preprocesamiento de Datos: Se efectúa una normalización de los datos, así mismo se realiza una reorganización de los datos en secuencias de entrada y salida apropiadas

para el modelo LSTM.

- c) Configuración del Modelo LSTM: Se define la arquitectura del modelo. Por otro lado, se agrupa el modelo con un optimizador Adam (en un software) y la función de pérdida adecuada.
- d) Entrenamiento del Modelo: Se fraccionan los datos en conjuntos de entrenamiento, validación y prueba, posteriormente se realiza un entrenamiento del modelo utilizando los datos de entrenamiento y validación en el conjunto de prueba.
- e) Preparación y Evaluación de Datos de Prueba: Se preparan los datos de prueba y se evalúa el desempeño del modelo utilizando métricas como el Error Cuadrático Medio de la Raíz (RMSE) y el Error Absoluto Medio (MAE).
- f) Visualización de Resultados: Finalmente se trazan los resultados predichos frente a los valores reales para evaluar la precisión del modelo.

2.2.13 Producción de mango

Según Martinez et al. (2020), el mango ha sido cultivado por el hombre durante aproximadamente cuatro mil años y sus orígenes se remontan a la región de Hindo-Berma, que abarca Desde las regiones orientales de la India hasta el sur de China y las áreas del sudeste asiático. No fue hasta alrededor del año 300 o 400 d.C. que las semillas de mango fueron llevadas por viajeros desde Asia al Medio Oriente, África Oriental y América del Sur, un proceso que tomó unos 3000 años. El comercio de especias, particularmente en el siglo XV, desempeñó un papel crucial en la difusión y cultivo del mango, con los portugueses estableciendo un comercio de mangos en la India, por consiguiente, en el siglo XVII, los exploradores españoles introdujeron la fruta en América del Sur y México.

Dentro de las especificaciones técnicas de este importante producto, el MIDAGRI (2010), manifiesta que el mango, reconocido por su nombre científico *Mangifera* indica, pertenece a la familia botánica *Anacardiaceae*, por otro lado, dentro de las diversas variedades de mango, se destacan algunas como Haden, Kent, Tommy Atkins y Criollo, cada una con características únicas en cuanto a sabor, tamaño y resistencia a enfermedades. Por consiguiente, resalta que el proceso de injerto del mango es crucial para garantizar la calidad y productividad del árbol, siendo comúnmente necesario esperar alrededor de cinco años para obtener las primeras cosechas luego de este procedimiento.

En la actualidad mango es una variedad de fruta que se encuentra típicamente en climas tropicales cultivada globalmente por su valor nutricional y su versatilidad culinaria,

además de cumplir las necesidades de los que lo producen y consumen, constituye una base económica vital para los entes participantes, por consiguiente, la producción involucra una interdependencia entre agricultores, comerciantes y consumidores, lo cual es esencial para garantizar la calidad y disponibilidad del producto en el mercado.

Modalidad de propagación

Por lo general el mango presenta 3 fases de manejo en el proceso de su propagación cuando el interviniente es el hombre.

Por su parte Michel et al. (2000) manifiesta que en el proceso de propagación del mango, se comienza con la etapa de semillero. Se aconseja emplear sustratos livianos y con alto contenido de materia orgánica, colocada en camas o eras de ancho de 1 metro y 15 centímetros de altura. Las semillas de mango deben ser sembradas inmediatamente después de la cosecha en un sustrato suelto, al cual se le puede adicionar pulpa de fruta o puño de hojas descompuestas. Si se elimina la corteza de la semilla, la germinación puede acelerarse, aunque se debe mantener la membrana que recubre los cotiledonales. En caso de semillas poliembriónicas, se recomienda eliminar las plántulas no deseadas para promover el vigor de las plantas deseables. Es importante sembrar las semillas con la parte aguda hacia arriba para facilitar un crecimiento vertical adecuado.

Así mismo Michel et al. (2000) describe que una vez que las plántulas alcanzan los 15 centímetros de altura, son trasladadas a un vivero, donde son sembradas en bolsas plásticas y se cuidan hasta alcanzar el porte adecuado para ser trasplantadas. Tanto el sustrato del semillero como del vivero deben ser desinfectados y así evitar enfermedades.

Finalmente Michel et al. (2000) expone que cuando los arbolitos están listos para ser injertados, se observa un cambio de color en el tallo y alcanzan un grosor mínimo. La oportunidad más adecuada para cortar el material vegetativo es cuando el árbol madre presenta brotes con hojas maduras y la corteza de la rama está en condiciones favorables. El injerto más efectivo es el de escudete o enchape lateral, utilizando material vegetativo que esté en pleno crecimiento y que tenga una yema terminal bien desarrollada. Es fundamental que tanto el patrón como el escudete sean de tamaño y madurez similares para mejorar las probabilidades de éxito del injerto.

2.2.14 Fases fenológicas

El mango atraviesa diversas fases fenológicas a lo largo de su ciclo de vida, las cuales son cruciales para su desarrollo y producción óptima. Según MIDAGRI (2010), en el Perú menciona que las principales fases incluyen:

- Brote: El proceso de brotación en el mango comienza con la aparición de yemas, que presentan una leve hinchazón y una tonalidad verde tenue, posteriormente, las yemas terminales se estiran y empiezan a aparecer los primeros botones de la flor con aspectos de espinas. Los primordios son alargados y las hojas son de color marrón rojizo. Finalmente, el tallo alcanza su tamaño final y la hoja crece por completo.
- Floración: Comienza cuando los botones comienzan a abrirse, permitiendo que las piezas florales iniciales se revelen. La inflorescencia se estira hasta alcanzar cerca de la mitad de su tamaño final y culmina con la separación y apertura total de las flores.
- Cuajado: En este caso se presentan 3 estados, en el primero los pétalos se han marchitado, y cubren en parte el ovario, el cual tiene un diámetro de 1 a 2 cm, mientras que el estilo seco aún es perceptible. Posteriormente, se experimenta una caída de frutas que continúa hasta la fase de maduración completa. En el último estado, los frutos en desarrollo se ubican separados individualmente, y el pedúnculo floral se ha alargado y fortalecido.
- Fructificación: Se le conoce como llenado del fruto, se produce un crecimiento continuo y progresivo de los frutos. Esta fase es decisiva en el ciclo de vida del mango y comienza después del proceso de cuajado, cuando las flores fecundadas han dado lugar a frutos incipientes. Durante la fructificación, los frutos experimentan un aumento en tamaño y peso, a medida que se desarrollan las estructuras internas y se acumulan nutrientes esenciales para su maduración.
- Maduración: Durante el proceso de maduración, los mangos alcanzan las características distintivas de su variedad, como su tamaño, color y sabor característicos. Sin embargo, es importante resaltar que la maduración es un proceso fisiológico complejo, que implica cambios en la composición química y física de la fruta.

En el contexto del manejo postcosecha y la comercialización, los mangos suelen ser cosechados en un estado de madurez fisiológica, identificada por la formación de hombros en la fruta, indicando que aún están en proceso de maduración. Este enfoque permite extender la vida útil de los mangos y garantizar su calidad durante el almacenamiento y el transporte. Por otro lado, estas fases se dan en temporadas o meses específicos del año, dependiendo en gran

medida de factores agroclimáticos, quienes influirán en la cantidad y calidad de producción, así mismo es importante precisar que estas fases pueden variar dependiendo del lugar y el contexto climático de cada región en el mundo.

Figura 20

Fenología general del mango peruano según los meses del año

MAR	ABR	MAY	JUN	JUL	AGO	SET	OCT	NOV	DIC	ENE	FEB
Brotamiento - Maduración de brotes				Floración - Cuajado			Crecimiento y maduración de frutos				
											

Nota: Recuperado del Boletín agroclimático realizado por Convenio Específico Interinstitucional suscrito entre SENAMHI, SENASA y ADEX (2023).

Según lo representado en la Figura 20, se aprecia que la fase de preparación de la planta para la producción, basado en su fenología, comienza en el mes de marzo con el brotamiento. La producción efectiva de frutos, por su parte, se inicia desde octubre hasta febrero (durante el crecimiento y la maduración), e incluso en algunos casos se extiende hasta marzo, dependiendo de las condiciones climáticas particulares de cada año. Es crucial destacar que cada fase fenológica requiere condiciones climáticas apropiadas, lo que significa que la producción está ampliamente influenciada por estos factores agroclimáticos, los cuales pueden tanto adelantar como retrasar la producción.

2.2.15 Los factores participes en la producción de mango

La producción de mango está influenciada por una diversidad de factores que afectan su crecimiento, desarrollo y rendimiento. Entre ellos se incluyen: Condiciones agroclimáticas.

La expansión del cultivo del mango se limita a regiones con climas tropicales y subtropicales, principalmente debido a su sensibilidad al frío. Para los climas tropicales según Villalobos (2020), implica la presencia de veranos y otoños cálidos seguidos de inviernos moderadamente fríos. Los momentos de temperaturas elevadas durante el verano de 30-34°C promueven una maduración adecuada de la fruta, por otro lado, las temperaturas cálidas durante el otoño promueven un crecimiento vegetativo óptimo y saludable después de la cosecha. Un crecimiento óptimo de las plantas se encuentra dentro de un rango de temperatura de 24-27°C, y se ve afectado negativamente por temperaturas por debajo de los 15°C.

Villalobos (2020) manifiesta que la temperatura desempeña un papel crucial en otros momentos importantes del ciclo del mango, tales como la inducción floral y el cuajado del fruto. Una inducción adecuada de la floración requiere un invierno moderadamente frío, con temperaturas mínimas en el rango de 10-15°C. Por otro lado, para un cuajado exitoso del fruto, es fundamental contar con una primavera cálida, evitando temperaturas inferiores a los 15°C. Por otro lado el MIDAGRI (2010) menciona que las áreas con una temperatura promedio anual entre 22 y 33 °C son propicias para el óptimo crecimiento del mango, aunque se observan variaciones según la región de origen de las distintas variedades. La amplitud térmica entre el día y la noche desempeña un papel crucial en el proceso de la floración, especialmente en variedades que tienen su origen en regiones subtropicales. La temperatura influye en la viabilidad del polen; temperaturas extremas, tanto bajas (menos de 10°C) como altas (más de 33°C), pueden afectar la fertilización, lo que posiblemente contribuye al bajo rendimiento de frutas en algunas variedades subtropicales comerciales. Las temperaturas cálidas nocturnas (28-32 °C) favorecen la dulzura y maduración adecuada de la fruta, mientras que los días calurosos con noches frescas (12 a 20 °C) parecen contribuir a un color más atractivo en la fruta.

Finalmente García (2010) destaca que en regiones con altitudes inferiores a 500 m.s.n.m, las temperaturas varían entre 24 y 28 °C, manteniéndose bastante estables a lo largo de todo el año. Aunque no se observan fluctuaciones significativas en la temperatura promedio, se pueden notar diferencias considerables entre las temperaturas máximas y mínimas durante las estaciones más secas.

Figura 21

Rangos generales de temperaturas para la producción de mango



Nota: Extraído de “4 condiciones ambientales claves para el cultivo del mango” de Villalobos (2020).

Es fundamental comprender que las temperaturas durante cada fase fenológica del mango pueden considerarse beneficiosas o perjudiciales según las temperaturas máximas y mínimas, por consiguiente, esta variación es ligeramente diferente entre las diversas variedades de mango, lo que da lugar a diferentes cifras reportadas en la literatura. Sin embargo, en su mayoría, estas cifras tienden a situarse en los rangos mencionados anteriormente. Suelos

El crecimiento saludable del mango se sostiene en gran parte del suelo y sus características en el que se cultiva. Según Villalobos (2020) recomienda suelos con buen drenaje y alta materia orgánica, como limo-arenosos con humus y suelos ligeramente ácidos. Incluso suelos alcalinos tratados con fertilizantes pueden ser adecuados, así mismo menciona que el pH ideal del suelo para el mango es entre 5.5 y 6.0, aunque puede tolerar valores de hasta 7.5, fuera de este rango se puede observar un crecimiento deficiente. Además, se aconseja plantar mangos en suelos con una profundidad de 80-100 cm, preferiblemente en suelos que sean ligeros para que faciliten la expansión de las raíces. Un estudio sobre el suelo en un área de cultivo exitoso de mango reveló niveles de nutrientes como 1.2% de calcio, 1.18% de magnesio, 2.73% de potasio, 0.15% de anhídrido fosfórico y 0.105% de nitrógeno.

Riego y precipitaciones

Según Arce et al. (2019), en la costa peruana, el cultivo de mango se realiza principalmente bajo sistemas de riego por gravedad, por consiguiente se estima que el requerimiento hídrico puede variar entre 10000 y 15000m³ por hectárea. Sin embargo, durante las épocas de mayor cosecha, especialmente en verano, los excesos de lluvia en la costa norte representan un riesgo climático significativo. Estas precipitaciones excesivas pueden propiciar la aparición de enfermedades como la antracnosis (*Colletotrichum gloeosporioides*), que afecta o influye en la calidad de los frutos y representa una amenaza para la producción de mango.

2.2.16 Zonas clave de producción en el Perú

Según León (2023), de la agencia agraria de noticias agraria.pe, durante el año 2023, se evidenció que las regiones productoras de mango principales en Perú fueron Piura, Lambayeque y Áncash. Se resaltó que Piura lidera la producción, acaparando alrededor del 75% del total nacional, seguida por Lambayeque y Áncash.

Por otro lado el SENAMHI (2024), en su boletín agroclimático, manifiesta que la producción nacional de dicho fruto está principalmente centrada en la costa peruana, siendo Piura la región con la mayor producción y extensión cultivada, abarcando 19,867 hectáreas (equivalentes al 64.6% del total). Esta región ha experimentado un crecimiento constante en su producción, con un patrón de crecimiento cíclico, además se ha observado que aproximadamente cada tres años de crecimiento sostenido es seguido por un año de recesión en la producción, como es el caso de la presente temporada (campaña 2023/2024), cuya disminución se atribuye a factores climáticos.

Figura 22

Provincias y distritos de Piura que producen mango según VBP.



Nota: Extraído del sistema agrícola (SISAGRI) implementado por el MIDAGRI.

Según los datos presentados en la Figura 22, para el año 2019, se observa que,

en el departamento de Piura, la provincia de Piura representa el 93.6% del Valor Bruto de Producción (VBP) en mango, superando significativamente a Morropón y Sullana. Es importante señalar que estos datos corresponden al año 2019, ya que el sistema agrario del MIDAGRI aún no ha sido actualizado hasta la fecha. Sin embargo, diversos actores del sector agrario afirman que actualmente la provincia de Piura sigue ocupando el primer lugar en producción de mango a nivel nacional. Además, dentro de los distritos de la provincia de Piura, Tambo Grande lidera en VBP con un 80.9%, seguido de Las Lomas con un 16.9%. Cabe destacar que estos dos distritos se ubican en la parte media del valle de San Lorenzo, siendo los únicos productores de mango en esta zona.

Por otro lado en Piura a noviembre de la última campaña del 2023, según León (2023), como menciona en la agencia agraria de noticias, Ángel Gamarra Condori, presidente de la Asociación de Productores de Mango del Perú (Promango), presentó un informe sobre la temporada 2023/2024, en el que pronostica que la producción de mango en el Perú alcanzará las 100.000 a 120.000 toneladas. Esto supondría una disminución significativa del 80% al 83,5% respecto al volumen de producción de la temporada anterior (2022/2023) de 600.000 toneladas. Gamarra Condori atribuyó esta disminución en la producción a las altas temperaturas registradas durante todo el año, las cuales afectaron el proceso de floración de las plantas y distorsionaron la fisiología del cultivo, afectando la economía regional de todos los entes participantes en este sector agrario

2.2.17 Producción de mango en el Valle de San Lorenzo

La producción masiva de mango en los distritos de Tambo Grande y Las Lomas, ubicados en la parte media del Valle de San Lorenzo, en la provincia de Piura, tiene sus raíces en el proyecto de irrigación y colonización San Lorenzo, concebido como un programa piloto de desarrollo agrícola por el Banco Mundial en América Latina. Este proyecto transformó una zona que prácticamente era imposible producir (desértica) en un valle altamente productivo, convirtiéndose en el programa de expansión de frontera agrícola más destacado del Perú.

La Junta de Usuarios del Sector Hidráulico Menor San Lorenzo (JUSHSAL, 2019), menciona que la ejecución del proyecto San Lorenzo se llevó a cabo en tres etapas entre 1948 y 1965, incluyendo la construcción de infraestructura vital como canales de riego, reservorios, y distribución de tierras a colonos. JUSHSAL (2019) manifiesta que esta iniciativa proporcionó a los agricultores las condiciones necesarias para el cultivo, como un suministro constante de agua y viviendas con servicios básicos. La provincia de Piura, específicamente los distritos de Tambo Grande y Las Lomas se beneficiaron

significativamente de este proyecto, convirtiéndose en los principales productores de mango a nivel nacional.

Actualmente, el valle de San Lorenzo enfrenta una crisis en la producción de mango, como señala un informe de León (2023) en la agencia agraria de noticias, quien cita las declaraciones del presidente de la Asociación Peruana de Productores de Mango (Promango), Ángel Gamarra Condori. Según Gamarra, se ha observado una reducción del 80% en la producción de mango en este valle debido al gran cambio climático lo cual afecta parte del proceso fenológico y por ello su optima producción.

2.3 Glosario de Términos Básicos

- **Celda de estado:** Se almacenan entradas relevantes e ignoran otras entradas mediante las compuertas LSTM.
- **Estacionalidad:** variación periódica en la producción y la disponibilidad de mangos a lo largo del año, influenciada por factores climáticos, como las estaciones y el clima.
- **Forget gate:** Es una compuerta que decide la información que se va a descartar, y que por tanto no pasará a la celda de estado.
- **Función de activación.** Este término se refiere a una función matemática de la forma $f(x)$, que se añade a las redes neuronales artificiales para ayudar a la red a aprender patrones de datos complejos.
- **LSTM (Long Short-Term Memory):** Un tipo de unidad de red neuronal recurrente diseñada para abordar problemas de desvanecimiento de gradiente y capturar dependencias de secuencias a largo plazo.
- **Modelos LSTM multivariados-multistep:** son variantes de redes neuronales recurrentes para el pronóstico de series temporales, donde se utilizan múltiples variables como entrada y producen predicciones a varios pasos del futuro.
- **Output gate:** o también llamada compuerta de salida, es la compuerta que calcula el nuevo estado oculto.
- **Perceptrón:** es la unidad de procesamiento básica de las redes neuronales artificiales que acepta múltiples entradas, las pondera y las suma y luego utiliza una función de activación para obtener la salida.
- **Producción de Mango:** La actividad agrícola que implica el cultivo y la cosecha de mangos, una fruta tropical ampliamente cultivada en regiones de clima cálido en todo el mundo.
- **Redes Neuronales Recurrentes (RNN):** un tipo de red neuronal que puede procesar

datos secuenciales y mantener la memoria de estados anteriores.

- **SARIMA (Seasonal Autoregressive Integrated Moving Average):** modelo estadístico empleado para analizar y pronosticar series temporales que contienen componentes estacionales, de tendencia y de estacionariedad.
- **Update gate:** En esta compuerta se puede actualizar la memoria de la celda LSTM, y se indica que ahora el sujeto es singular.

2.4 Hipótesis

2.4.1 Hipótesis General

Existen diferencias significativas en la precisión de los pronósticos de la producción mensual de mango de la parte media del Valle de San Lorenzo-Piura, obtenidos a partir del modelo SARIMA y de las Redes Neuronales (LSTM).

2.5 Operacionalización de Variables

- Producción de mango: Aquel que está conformado por el peso en miles de toneladas de mango sin considerar las variedades.

Tabla 1

Operacionalización de la variable

Variable	Definición conceptual	Definición operacional	Unidades	Escala de medición
Producción de Mango	Cantidad total de la producción mensual de mango en la región Piura (Miles de toneladas).	Cantidad total de la producción mensual de mango obtenida mediante solicitud al MIDAGRI.	Producción de mango (Miles de Toneladas)	Cuantitativa continua

Nota: Elaboración propia

CAPÍTULO III: MARCO METODOLÓGICO

3.1 Enfoque

El enfoque de esta investigación es cuantitativo, para ello, Hernández et al. (2010) argumentaron que utilizan la recopilación de datos para probar hipótesis basadas en mediciones numéricas y análisis estadísticos para identificar patrones de comportamiento y probar teorías, cada etapa precede a la siguiente y no podemos eludir pasos.

3.2 Diseño

El diseño correspondiente es no experimental, tal como lo afirma Hernández et al. (2010), que se ejecuta sin manipulación deliberada de variables y sólo en este caso, su observación de los fenómenos naturales conduce al análisis.

3.3 Nivel

El nivel que se indica en el estudio según los objetivos es predictivo, donde según Whitney (1990) el enfoque implica predecir escenarios futuros, teniendo en cuenta estudios detallados de la evolución dinámica de los acontecimientos, su correlación con el contexto, los factores deliberados de las partes involucradas y la probabilidad de que algunos de estos escenarios sucedan.

3.4 Tipo

El estudio es de tipo aplicada, donde Castro et al. (2023) menciona que la investigación aplicada recibe el nombre de “investigación práctica o empírica”, que se caracteriza porque busca la aplicación o utilización de los conocimientos adquiridos, a la vez que se adquieren otros, después de implementar y sistematizar la práctica basada en investigación. El uso del conocimiento y los resultados de investigación que da como resultado una forma rigurosa, organizada y sistemática de conocer la realidad.

3.5 Sujetos de Investigación

3.5.1 Población

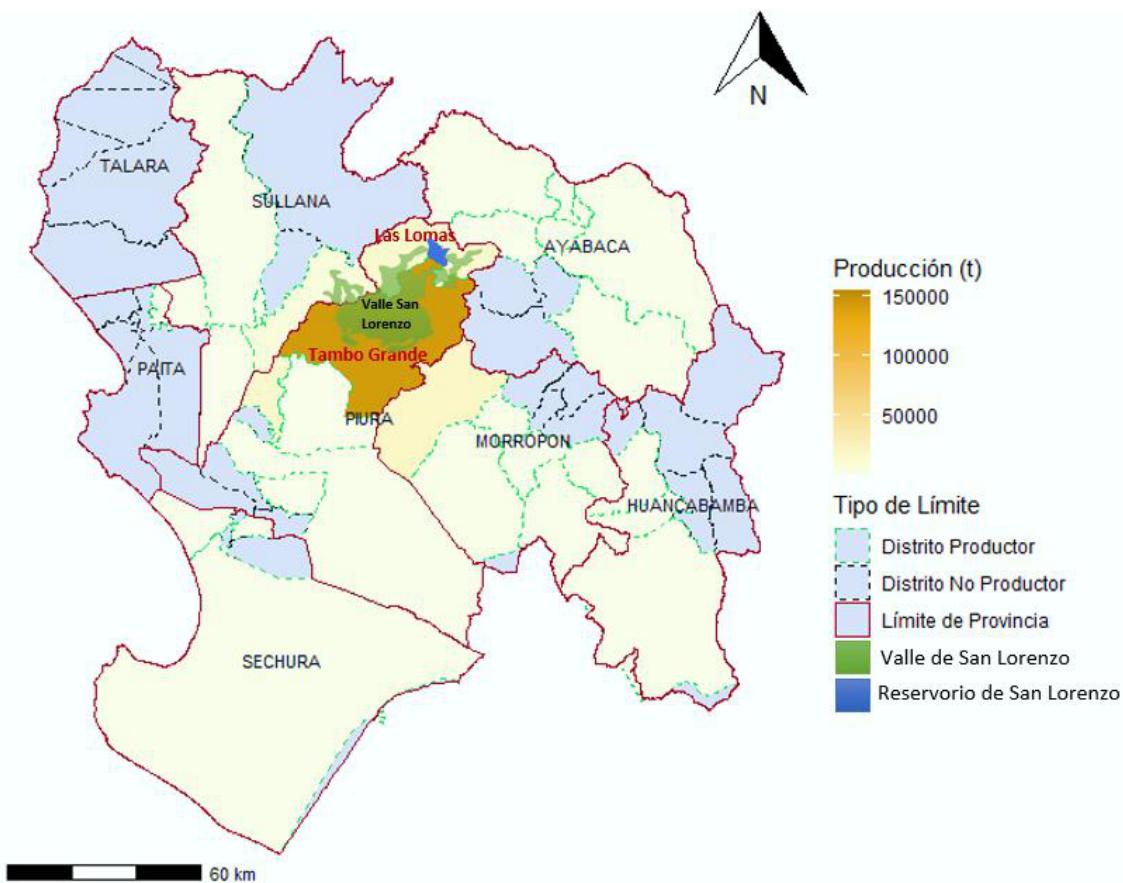
La población de la investigación está conformada por el total de los datos de valor bruto de producción de mango mensual en el departamento de Piura, desde el año que se empezó a producir, hasta la actualidad.

3.5.2 Muestra

Así mismo, la muestra fue conformada por el total de datos de la producción de mango en la zona media del valle de San Lorenzo entre los meses de enero de 2000 a agosto de 2024, por lo tanto, correspondió a 296 valores.

Figura 23

Producción de mango en el Valle de San Lorenzo, y Provincias de Piura



Nota: Elaboración propia en R, basada en datos geoespaciales del INEI, y datos de producción por parte del MIDAGRI en el 2023

3.6 Métodos y Procedimientos

Con base en los objetivos propuestos, se procedió a recolectar los datos necesarios para el estudio mediante una solicitud al MIDAGRI, a fin de obtener la información mensual sobre la producción de mango. Los datos se descargaron inicialmente en una hoja de cálculo de Excel y posteriormente fueron exportados al lenguaje de programación estadística R, utilizando el entorno de desarrollo integrado RStudio. A partir de ello, se realizó un análisis exploratorio para comprender el comportamiento de la serie temporal y detectar posibles valores faltantes o atípicos. Luego, se verificó la estacionariedad de la serie mediante pruebas estadísticas específicas, y se procedió a estimar modelos de series temporales, empleando

tanto el modelo SARIMA como el modelo de red neuronal recurrente LSTM, con el fin de comparar su desempeño utilizando medidas de eficiencia basadas en los errores de predicción. Finalmente, se efectuó el pronóstico de la producción de mango utilizando el modelo con mejor ajuste.

Por otro lado, es importante destacar que se uso de herramientas de software libre como R y Quarto, estas herramientas permiten integrar el análisis estadístico con la redacción automatizada del informe final.

3.7 Técnicas e Instrumentos

La técnica utilizada en el estudio correspondió a la búsqueda electrónica en las bases del MIDAGRI. El instrumento utilizado es la hoja de cálculo de Excel, donde se descargó la información para su análisis.

3.8 Aspectos Éticos

El presente estudio cumple con diversos principios éticos fundamentales en la investigación científica. En primer lugar, se respetó el principio de **objetividad**, garantizando que los datos utilizados fueron obtenidos de manera legítima mediante una **solicitud formal al Ministerio de Desarrollo Agrario y Riego (MIDAGRI)**. La interpretación de los resultados se realizó con rigurosidad técnica y se presenta de forma clara y accesible para cualquier lector interesado, promoviendo la **transparencia** y la **comprensión pública** de los hallazgos.

Asimismo, se ha procurado mantener la **transparencia metodológica**, describiendo detalladamente el proceso seguido para llegar a los resultados de pronóstico, desde el tratamiento de los datos hasta la selección y validación de los modelos utilizados. Se han respetado en todo momento los **derechos de autor**, citando debidamente todas las fuentes conforme a las normas de la **APA, séptima edición**.

En concordancia con principios éticos relacionados con la **legalidad, la equidad y la transparencia**, se optó por utilizar exclusivamente **software libre** en el desarrollo de esta tesis, como **Quarto** para la redacción del documento en PDF y **R** para los análisis estadísticos evitando el uso de software comercial con licencias alteradas, contribuyendo a la **democratización del conocimiento científico**.

CAPÍTULO IV: RESULTADOS Y DISCUSIÓN

4.1 Resultados

Descripción de la serie temporal de la producción mensual de mango (x 1000 toneladas) en la parte media del Valle de San Lorenzo, entre enero del 2000 a agosto de 2024.

Tabla 2

Medidas descriptivas de la variable Producción de Mango

Estadístico	Valor
Obs.	296
Mín.	0
1st qu.	0
Med.	235
Media	15,148
3rd qu.	10,119
Máx.	148,800
D.E.	32,368.2
Asim.	2.574
Curt.	5.845
Moda	0

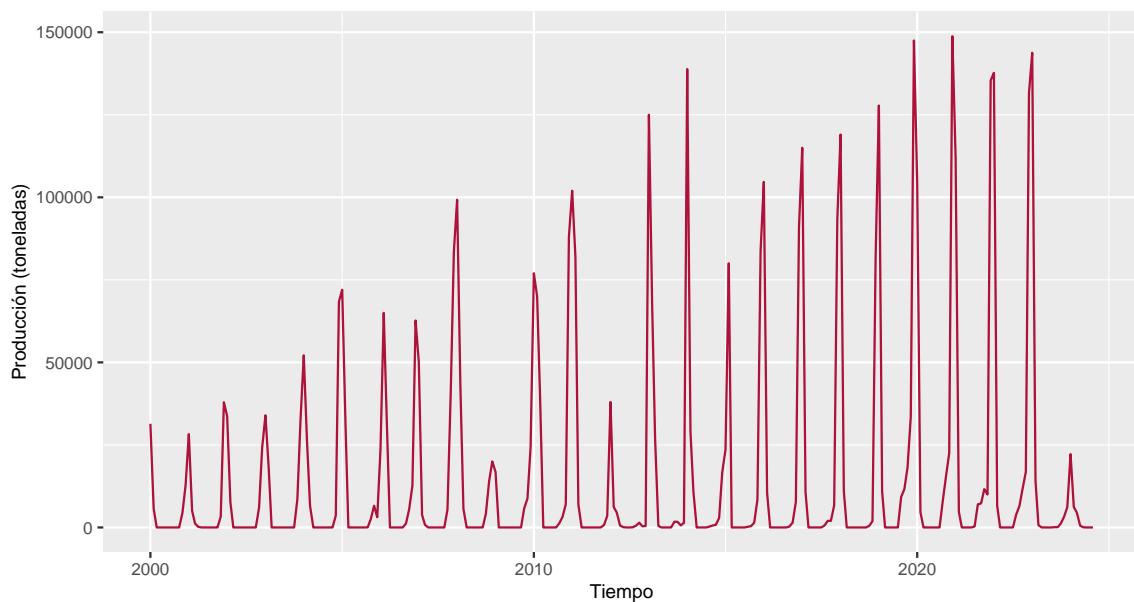
Nota: Estadísticas obtenidas de los datos de MIDAGRI.

En la Tabla 2 se presentan los estadísticos descriptivos de la serie temporal de la producción de mango en el Valle de San Lorenzo, desde enero de 2000 hasta agosto de 2024. La serie muestra que hubo periodos sin producción (Mínimo =0, primer cuartil=0 y moda=0), que corresponden a períodos fuera de temporada. La mediana alcanzó el valor de 235 (x1000 toneladas), lo que implica que el 50% de los periodos alcanzaron una producción igual o inferior a este valor, mientras que en el resto de los periodos la producción superó a dicha cifra. La media de 15,148 (x1000 toneladas) mucho más alta que la mediana, sugiere que la presencia de valores extremadamente altos en la serie, lo cual está corroborado por el tercer cuartil, que establece que en el 75% de los periodos se produjo 10,119 (x1000 toneladas) o menos, mientras que en el 25% de dichos periodos, la producción fue superior. El valor máximo alcanzó las 148,800 (x1000 toneladas), lo que evidencia una considerable variabilidad en los datos. La desviación estándar de 32,368.2 (x1000 toneladas) destaca una

alta dispersión alrededor de la media. La asimetría de 2.574 y la curtosis de 5.845 indican una distribución con una asimetría y curtosis marcada, lo que evidencia falta de normalidad en los datos generada por la presencia de valores extremos.

Figura 24

Serie Temporal de la producción de mango (t)

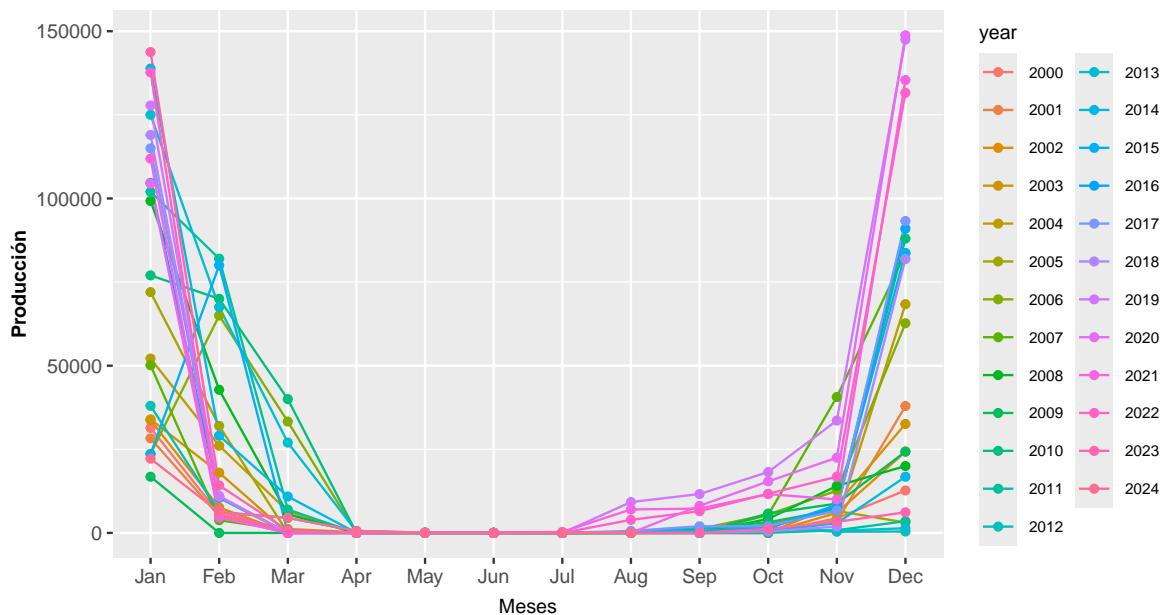


Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 24, se muestra la producción de mango en el valle de San Lorenzo, 2024-2026. Se puede observar que la serie presenta una fuerte estacionalidad, acompañada de meses sin producción, la que también decae considerablemente en el año 2023. Así mismo, la figura muestra que los valores de la serie a partir del año 2008 se muestran más variables, lo que es un indicio de que la varianza de la serie no es constante, lo cual es un requisito de las series estacionarias.

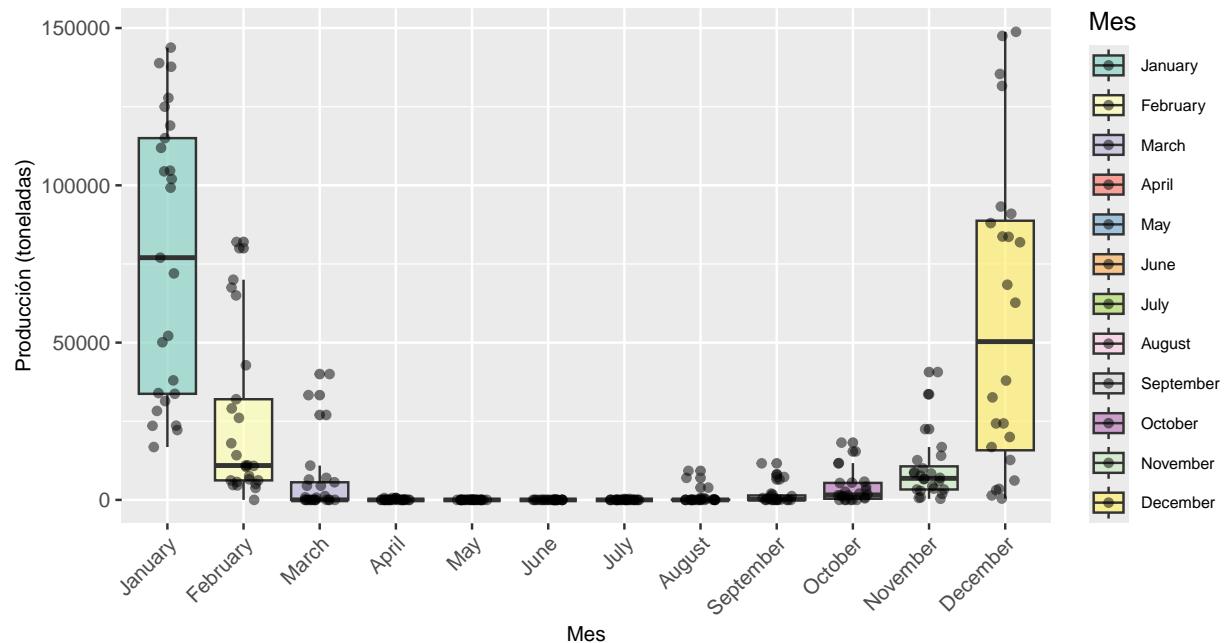
Figura 25

Estacionalidad mensual de la producción de mango (t)



Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 25 se presenta la gráfica estacional de la producción de mango en el Valle de San Lorenzo para el período 2000-2024. Se ilustra detalladamente la evolución mensual de la producción a lo largo del tiempo especificado. El análisis estacional revela que, en la mayoría de los años, no se registra producción de mango entre los meses de abril y julio. Por ende, la producción se concentra principalmente en dos períodos: de enero a abril y de julio a diciembre. En particular, destacan los meses de diciembre y enero por presentar los niveles más altos de producción, alcanzando los valores más altos en este último mes.

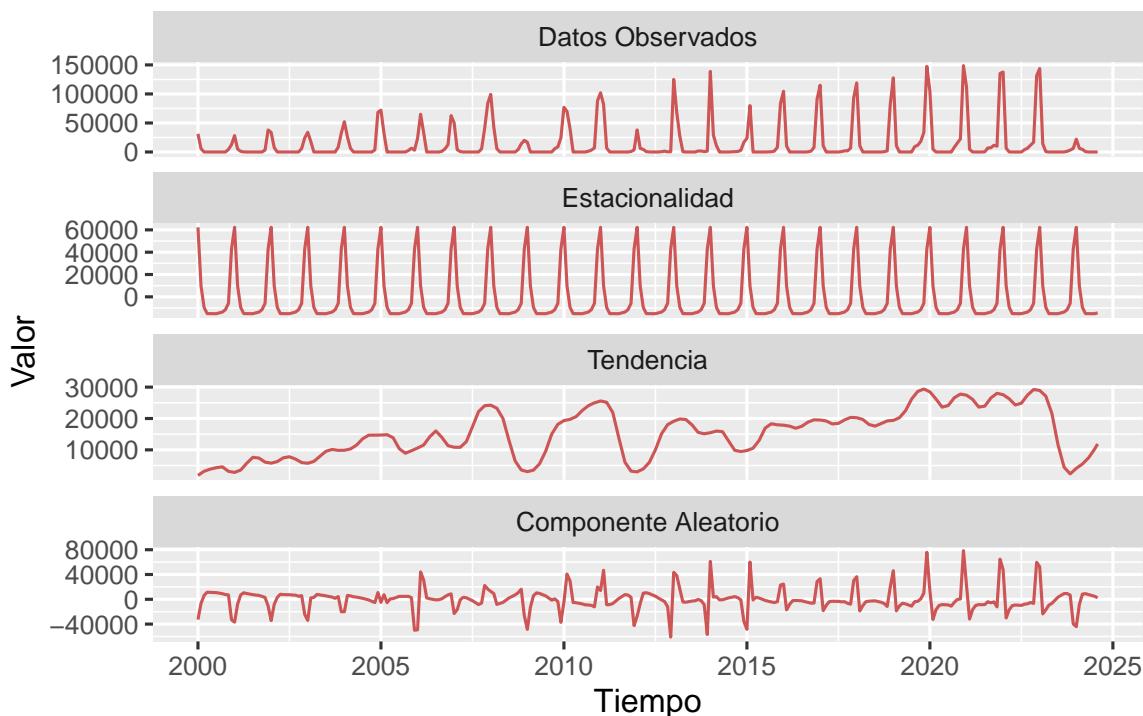
Figura 26*Distribución mensual de la producción de mango (t)*

Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 26, se observa que en los meses de mayor producción (diciembre, enero y febrero), las medianas se encuentran cerca del centro de la caja de bigotes, lo que indica que el valor central de la producción en estos meses está relativamente equilibrado en comparación con el rango intercuartílico, mientras que las medias tienden a estar influenciadas por valores atípicos, los que también contribuyen a una alta dispersión. Por otro lado, en los meses de marzo a septiembre, las medianas son cercanas a cero, prácticamente nulas debido a la ausencia de producción. Este comportamiento proporciona indicios de que la serie no es estacionaria, dado que el promedio fluctúa considerablemente a lo largo del año.

Figura 27

Componentes de la serie de producción de mango (t)



Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 27 se presentan las componentes de la serie de tiempo de la producción de mango en el valle de San Lorenzo, 2000-2024, donde corrobora la presencia de estacionalidad (seasonal), y una ligera tendencia (trend), la cual, no se percibe claramente debido a los períodos de ausencia de producción; en esta figura también se observa que la tendencia decrece en el último año debido a la baja producción histórica que vivió el valle de San Lorenzo en el 2023. La figura muestra también el comportamiento de los residuos (remanider) que, al inicio de la serie, son bastante estables, pero que, a partir del año 2008, se muestran más variables, dejando en evidencia un comportamiento más errático en estos períodos.

Estimación del modelo SARIMA para el pronóstico de la producción mensual de mango en la parte media del Valle de San Lorenzo, con la metodología de Box Jenkins, basados en los datos históricos comprendidos entre enero del 2000 a agosto de 2024.

Verificación de la estacionariedad de la serie

- **Hipótesis Nula (H_0):** La serie temporal tiene una raíz unitaria, es decir, no es estacionaria.
- **Hipótesis Alternativa (H_1):** La serie temporal no tiene una raíz unitaria, es decir, es estacionaria.

Tabla 3

Test de Dickey-Fuller Aumentada (ADF)

Test	t.Statistic
Augmented Dickey-Fuller (ADF)	-10.04784
Critical Value for: 1%	-2.58000
5%	-1.95000
10%	-1.62000

Nota: Estadísticas obtenidas de los datos de MIDAGRI.

En la Tabla 3 se presenta el test de Dickey-Fuller Aumentado (ADF) para la serie de producción de mango en el Valle de San Lorenzo (2000-2024), donde el valor del estadístico de prueba ADF es -10.04842, que es significativamente más negativo que los valores críticos para los niveles de significancia del 1% (-2.580000), 5% (-1.950000) y 10% (-1.620000). Por lo tanto, podemos rechazar la hipótesis nula de que la serie temporal tiene una raíz unitaria, concluyendo que la serie de producción de mango es estacionaria.

- **Hipótesis Nula (H_0):** La serie temporal es estacionaria en torno a una media (o tendencia determinística).
- **Hipótesis Alternativa (H_1):** La serie temporal no es estacionaria (tiene una raíz unitaria).

Tabla 4*Test de Kwiatkowski Phillips Schmidt Shin (KPSS)*

Test	Statistic
Kwiatkowski Phillips Schmidt Shin (KPSS)	0.0249726
Critical Value for: 10%	0.1190000
5%	0.1460000
2.5%	0.1760000
1%	0.2160000

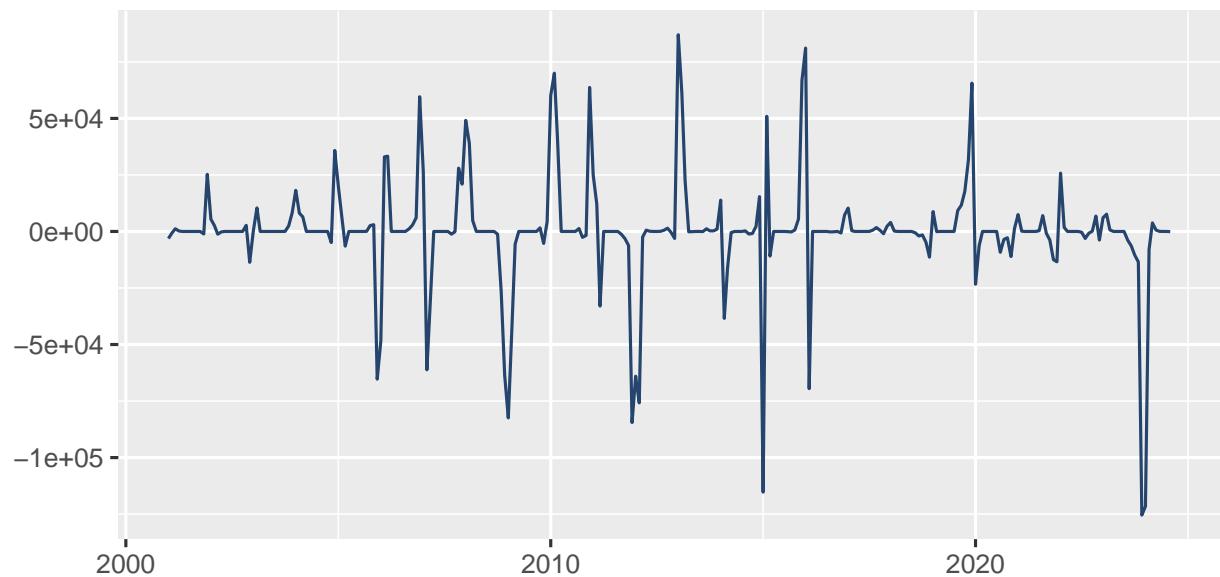
Nota: Estadísticas obtenidas de los datos de MIDAGRI.

En la Tabla 4 se muestra el test de Kwiatkowski–Phillips–Schmidt–Shin (KPSS) para la serie de producción de mango en el Valle de San Lorenzo (2000-2024), donde el valor del estadístico de prueba (0.0249726) es significativamente menor que todos los valores críticos proporcionados para los distintos niveles de significancia, esto significa que no hay suficiente evidencia estadística para rechazar la hipótesis nula en ninguno de los niveles de significancia usuales (10%, 5%, 2.5%, 1%), por lo tanto, la serie de tiempo es estacionaria alrededor de la tendencia.

Si bien es cierto ambos test concluyeron que la serie es estacionaria, y en teoría no necesita diferenciación alguna, es importante tener en cuenta que estos test solo evalúan la parte no estacional, por lo tanto, no permiten detectar la presencia de raíces unitarias en la componente estacional. Adicional a ello considerando los valores extremos y la fuerte estacionalidad, se tomo la decisión de aplicar una diferencia estacional $D = 1$, esta decisión también fue respaldada por la función `nsdiffs()` de la librería `forecast` la cual usa “SEAS” (predeterminado), donde Hyndman y Athanasopoulos (2021) manifiesta que es una medida de intensidad estacional, donde se selecciona la diferenciación si la intensidad estacional supera 0,64 (basado en la minimización de MASE al realizar pronósticos).

Figura 28

Diferencia estacional ($D=1$) de la producción mensual de mango



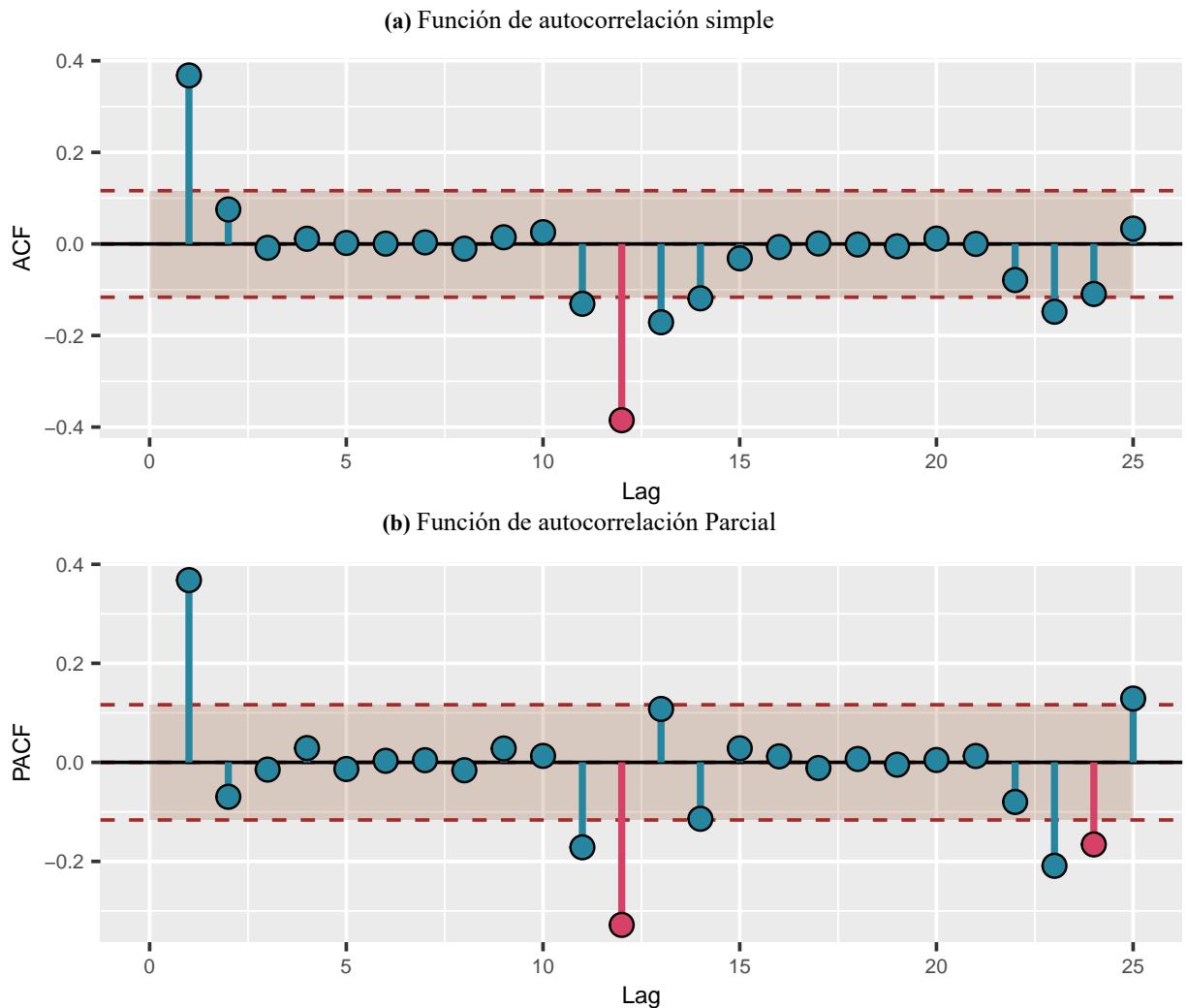
Nota: Estadísticas obtenidas de los datos de MIDAGRI.

De la Figura 28, se aprecia los resultados de la primera diferencia estacional para la producción de mango en el Valle de San Lorenzo (2000-2024), donde se percibe una media que persiste en 0, y una varianza visualmente más constante.

1. Identificación de los parámetros del modelo ARIMA.

Figura 29

Correlogramas de la serie con diferencia estacional



Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 29, se muestra las funciones de autocorrelación simple (a) y autocorrelación parcial (b) para la primera diferencia de la serie de la producción de mango en el Valle de San Lorenzo (2000-2024). En la gráfica del apartado b) se visualiza un claro componente autorregresivo AR(1), por otro lado, en a), se evidencia un componente de media móvil MA(1). Analizando la parte estacional, se observa que en b), presenta una barra por encima del intervalo, justo en el retardo 12, lo que demuestra que también existe un componente AR(1), en la parte estacional, de la misma manera sucede en a), por lo tanto, el modelo encontrado es un SARIMA (1,0,1)(2,1,1)[12], sin embargo, dada la posibilidad de obtener otro modelo mejor, se plantean los siguientes modelos candidatos:

- SARIMA(1,0,1)(2,1,0)[12]
- SARIMA(1,0,1)(2,1,1)[12]

- SARIMA(1,0,0)(2,1,0)[12]
- SARIMA(0,0,1)(2,1,0)[12]
- SARIMA(1,0,0)(0,1,1)[12]
- SARIMA(0,0,1)(2,1,1)[12]
- SARIMA(0,0,1)(0,1,1)[12]

2. Estimación.

Tabla 5

Modelos Estimados para la Serie de Producción de Mango con Diferencia Estacional

Estadísticas de los modelos					
Modelo	Parámetro	Estimación	Error estándar	Estadístico	Valor p
SARIMA(1,0,1)(2,1,0)[12]	ar1	0.120	0.170	0.703	0.483
	ma1	0.199	0.166	1.197	0.232
	sar1	-0.611	0.061	-9.948	0.000
	sar2	-0.371	0.059	-6.305	0.000
SARIMA(1,0,1)(2,1,1)[12]	ar1	0.111	0.154	0.719	0.473
	ma1	0.261	0.149	1.750	0.081
	sar1	-0.214	0.157	-1.362	0.174
	sar2	-0.190	0.103	-1.840	0.067
	sma1	-0.467	0.154	-3.027	0.003
SARIMA(1,0,0)(2,1,0)[12]	ar1	0.298	0.058	5.175	0.000
	sar1	-0.612	0.062	-9.934	0.000
	sar2	-0.372	0.059	-6.305	0.000
SARIMA(0,0,1)(2,1,0)[12]	ma1	0.304	0.055	5.573	0.000
	sar1	-0.616	0.061	-10.063	0.000
	sar2	-0.376	0.059	-6.400	0.000
SARIMA(1,0,0)(0,1,1)[12]	ar1	0.350	0.055	6.314	0.000
	sma1	-0.671	0.049	-13.570	0.000
SARIMA(0,0,1)(2,1,1)[12]	ma1	0.355	0.055	6.403	0.000
	sar1	-0.220	0.157	-1.401	0.162
	sar2	-0.194	0.103	-1.874	0.062
	sma1	-0.467	0.154	-3.020	0.003
SARIMA(0,0,1)(0,1,1)[12]	ma1	0.369	0.052	7.064	0.000
	sma1	-0.687	0.049	-14.152	0.000

Nota: Creada con R a partir de los datos del MIDAGRI.

En la Tabla 5 se presentan los modelos estimados mediante máxima verosimilitud para la serie de producción de mango en el valle de San Lorenzo, correspondiente al período 2000-2024. El primer modelo, SARIMA(1,0,1)(2,1,0)[12], muestra la presencia de dos parámetros no significativos, lo que determina que no se considere en el diagnóstico. De

manera similar, el modelo SARIMA(1,0,1)(2,1,1)[12] presenta únicamente un parámetro significativo, lo que limita su validez. Por último, el modelo SARIMA(0,0,1)(2,1,1)[12] también exhibe dos parámetros no significativos, lo que sugiere que su capacidad explicativa es limitada.

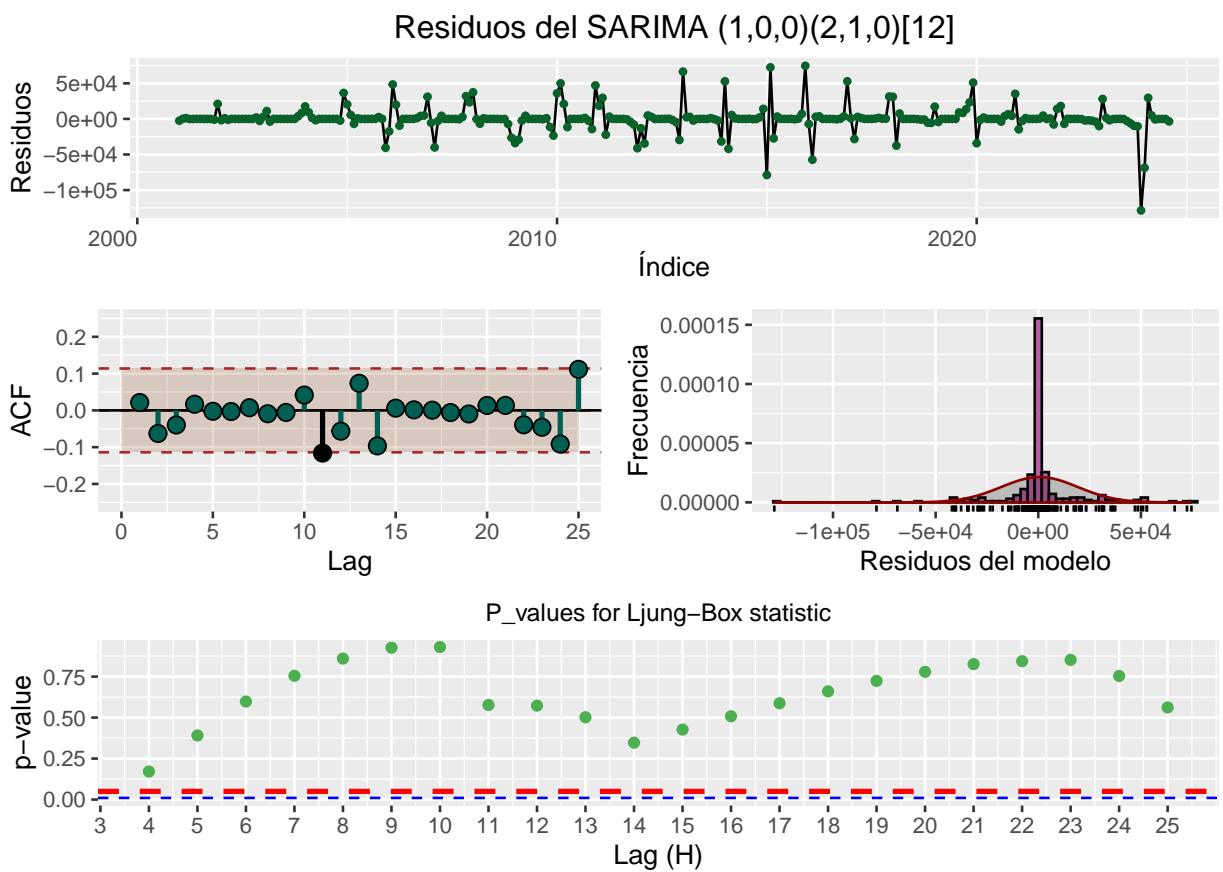
3. Diagnóstico

Los modelos que pasan a diagnóstico son los siguientes:

- MOD3: SARIMA (1,0,0)(2,1,0)[12]
- MOD4: SARIMA(0,0,1)(2,1,0)[12]
- MOD5: SARIMA(1,0,0)(0,1,1)[12]
- MOD7: SARIMA(0,0,1)(0,1,1)[12]

Figura 30

Diagnóstico visual del modelo SARIMA (1,0,0)(2,1,0)[12]



Nota: Creada con R a partir de los datos del MIDAGRI.

Tabla 6*Test's de Diagnóstico del modelo SARIMA (1,0,0)(2,1,0)[12]*

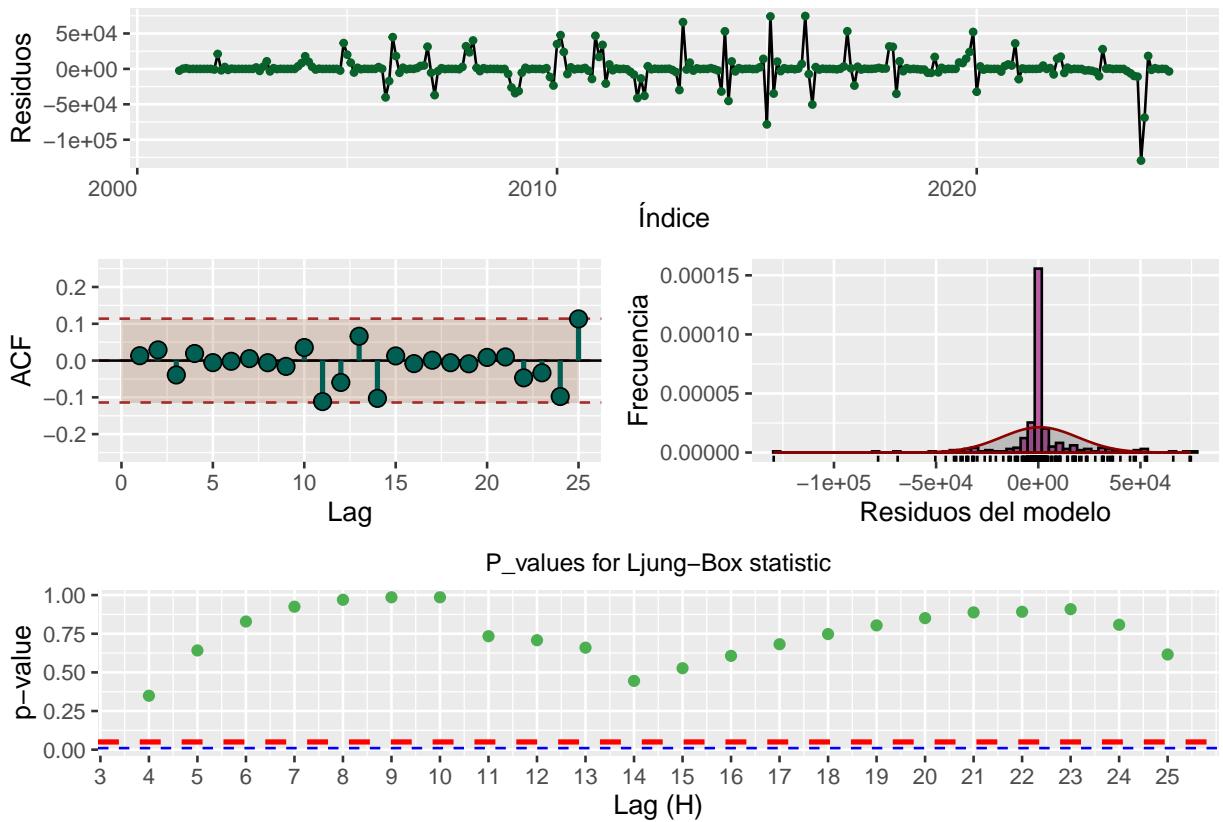
Supuesto	Prueba	Estadístico	P valor	Decisión
Independencia	Ljung-Box	16.281	0.878	Independientes
Normalidad	Jarque-Bera	1666.447	0.000	No Normal
Homocedasticidad	Levene	1.024	0.435	Homocedasticidad
Media 0	t de Media	0.544	0.587	No difiere de 0
Estacionariedad	KPSS	0.132	0.100	Stationary

Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 30 se presenta el diagnóstico visual del modelo SARIMA(1,0,0)(2,1,0), donde los residuos, siguen un comportamiento constante en torno a 0, así mismo, todos los retardos del AFC se encuentran dentro de la banda de confianza, excepto el 11, sin embargo, este no es pronunciado, por consiguiente la figura del test de Ljung-Box para cada retardo se encuentran muy por encima del 0.05 lo que demuestra que los residuos son independientes, finalmente según el histograma, no se cumple el supuesto de normalidad. El análisis anterior se refuerza en la Tabla 6, donde se corrobora que los residuos son independientes (Ljung-Box conjunto $p = 0.878$), homocedásticos (Levene $p = 0.435$) y con media cercana a 0 (t de media $p = 0.587$). No obstante, los residuos no son normales (Jarque-Bera $p = 0.000$). Finalmente, el proceso es estacionario (KPSS $p = 0.100$) y por lo tanto se deja evidencia de que los residuos del modelo son ruido blanco.

Figura 31*Diagnóstico visual del modelo SARIMA (0,0,1)(2,1,0)[12]*

Residuos del SARIMA (0,0,1)(2,1,0)[12]

*Nota:* Creada con R a partir de los datos del MIDAGRI.**Tabla 7***Test's de Diagnóstico del modelo SARIMA (0,0,1)(2,1,0)[12]*

Supuesto	Prueba	Estadístico	P valor	Decisión
Independencia	Ljung-Box	15.294	0.912	Independientes
Normalidad	Jarque-Bera	1738.859	0.000	No Normal
Homocedasticidad	Levene	1.109	0.335	Homocedasticidad
Media 0	t de Media	0.600	0.549	No difiere de 0
Estacionariedad	KPSS	0.140	0.100	Stationary

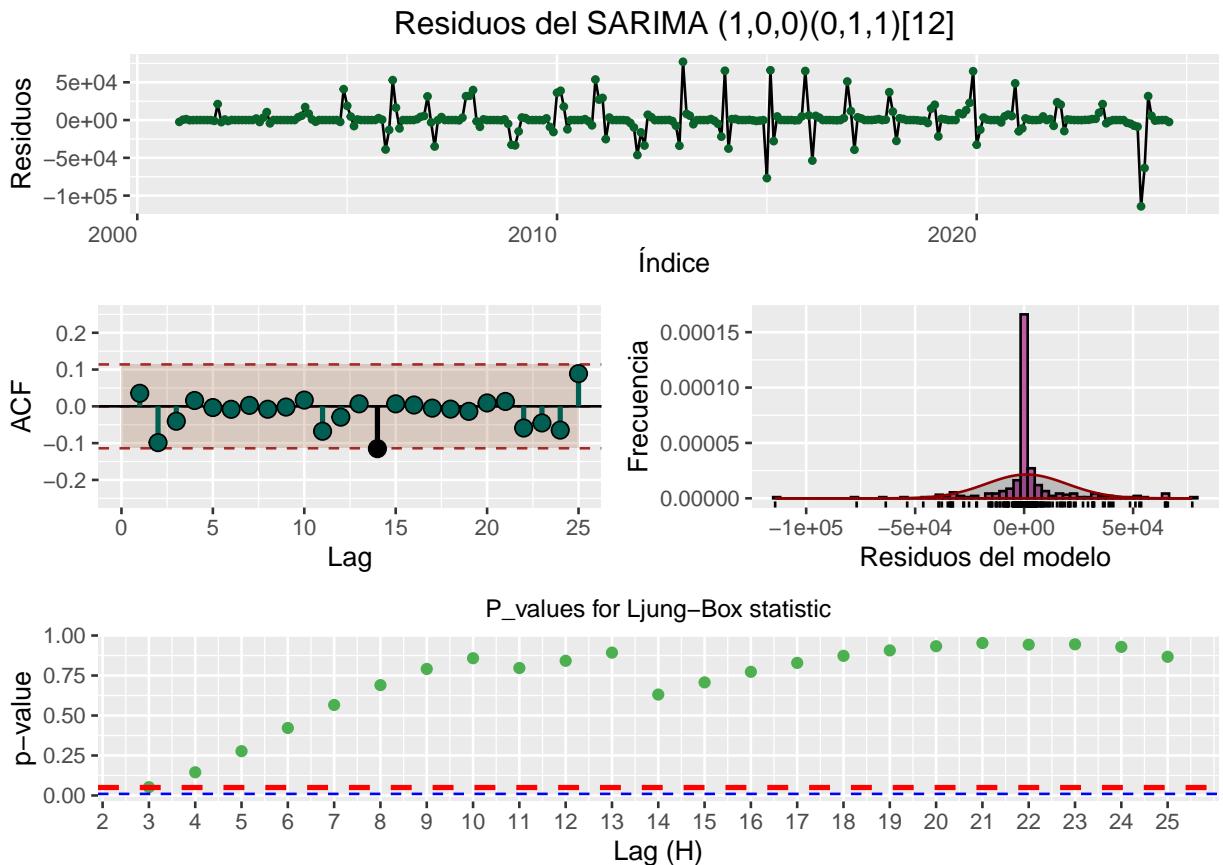
Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 31 se muestra el diagnóstico visual del modelo SARIMA(0,0,1)(2,1,0)[12], donde los residuos fluctúan en torno a 0. Todos los retardos del AFC se mantienen dentro de la banda de confianza, lo cual sugiere independencia de los residuos, corroborada por los p-valores del test de Ljung-Box que están por encima de 0.05, indicando que no hay autocorrelación significativa. Sin embargo, el histograma revela que los residuos no siguen una distribución normal. Estos resultados visuales coinciden con los de la Tabla 7, que confirman la independencia de los residuos (Ljung-Box $p = 0.912$), homocedasticidad

(Levene $p = 0.335$) y que la media de los residuos no difiere significativamente de 0 ($p = 0.549$). No obstante, los residuos no son normales (Jarque-Bera $p = 0.000$). En general, el proceso es estacionario (KPSS $p = 0.100$), lo que indica que los residuos del modelo pueden considerarse ruido blanco.

Figura 32

Diagnóstico visual del modelo SARIMA (1,0,0)(0,1,1)[12]



Nota: Creada con R a partir de los datos del MIDAGRI.

Tabla 8

Test's de Diagnóstico del modelo SARIMA (1,0,0)(0,1,1)[12]

Supuesto	Prueba	Estadístico	P valor	Decisión
Independencia	Ljung-Box	13.141	0.964	Independientes
Normalidad	Jarque-Bera	1010.205	0.000	No Normal
Homocedasticidad	Levene	0.894	0.607	Homocedasticidad
Media 0	t de Media	0.959	0.338	No difiere de 0
Estacionariedad	KPSS	0.090	0.100	Stationary

Nota: Creada con R a partir de los datos del MIDAGRI.

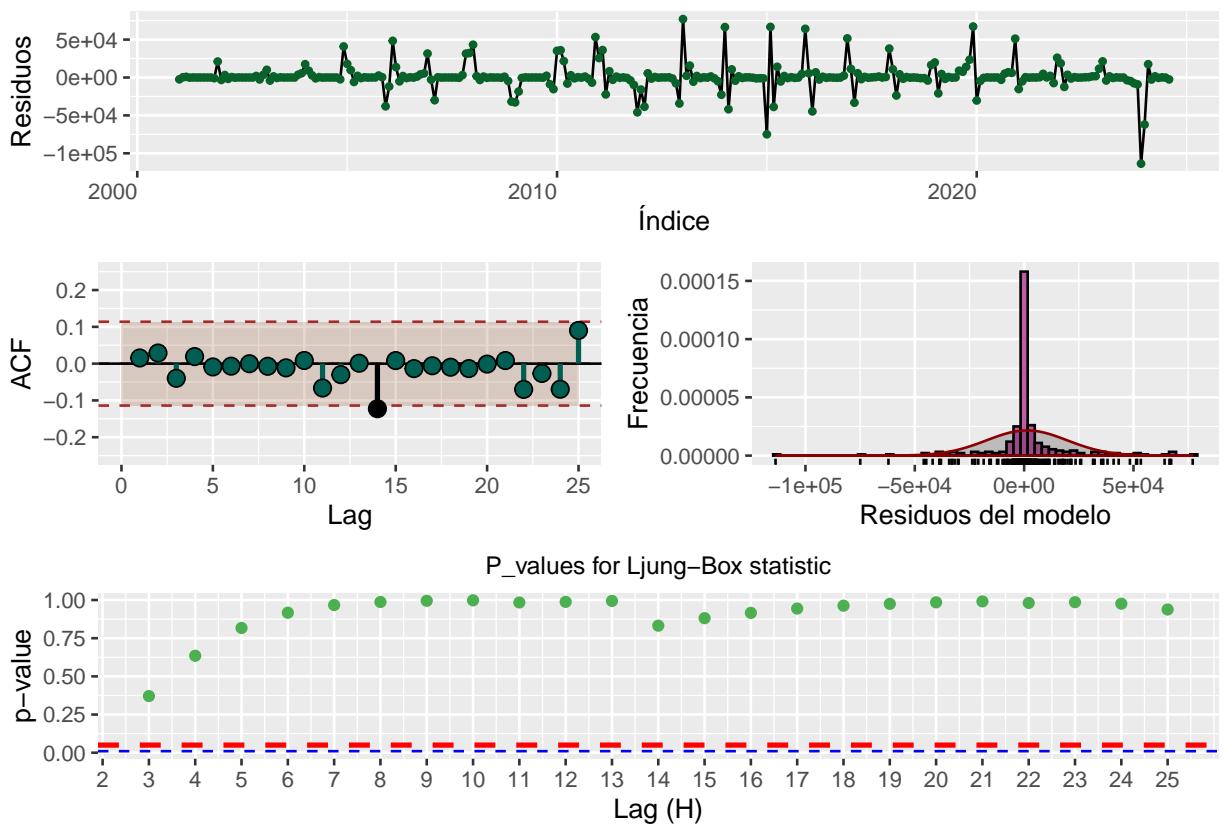
En la Figura 32 se aprecia el diagnóstico visual del modelo SARIMA (1,0,0)(2,1,0)[12], donde los residuos, siguen un comportamiento constante en torno a 0, así mismo todos los retardos del AFC se encuentran dentro de la banda de confianza, excepto el lag 14,

pero de manera leve, por consiguiente la figura del test de Ljung-Box para cada lag se encuentran muy por encima del 0.05 lo que demuestra que los residuos son independientes, finalmente según el histograma, no se cumple el supuesto de normalidad. Estos hallazgos visuales se refuerzan en la Tabla 8, mediante test de los supuestos, donde los residuos son independientes (Ljung-Box conjunto $p = 0.964$), homocedásticos (Levene $p = 0.607$) y con media cercana a 0 (t de media $p = 0.338$). Sin embargo, los residuos no siguen una distribución normal (Jarque-Bera $p = 0.000$). En general el proceso es estacionario (KPSS $p = 0.100$) y por lo tanto se deja evidencia de que los residuos del modelo son ruido blanco.

Figura 33

Diagnóstico visual del modelo SARIMA(0,0,1)(0,1,1)[12]

Residuos del SARIMA(0,0,1)(0,1,1)[12]



Nota: Creada con R a partir de los datos del MIDAGRI.

Tabla 9*Test's de Diagnóstico del modelo SARIMA(0,0,1)(0,1,1)[12]*

Supuesto	Prueba	Estadístico	P valor	Decisión
Independencia	Ljung-Box	10.941	0.989	Independientes
Normalidad	Jarque-Bera	1018.546	0.000	No Normal
Homocedasticidad	Levene	1.030	0.428	Homocedasticidad
Media 0	t de Media	1.148	0.252	No difiere de 0
Estacionariedad	KPSS	0.093	0.100	Stationary

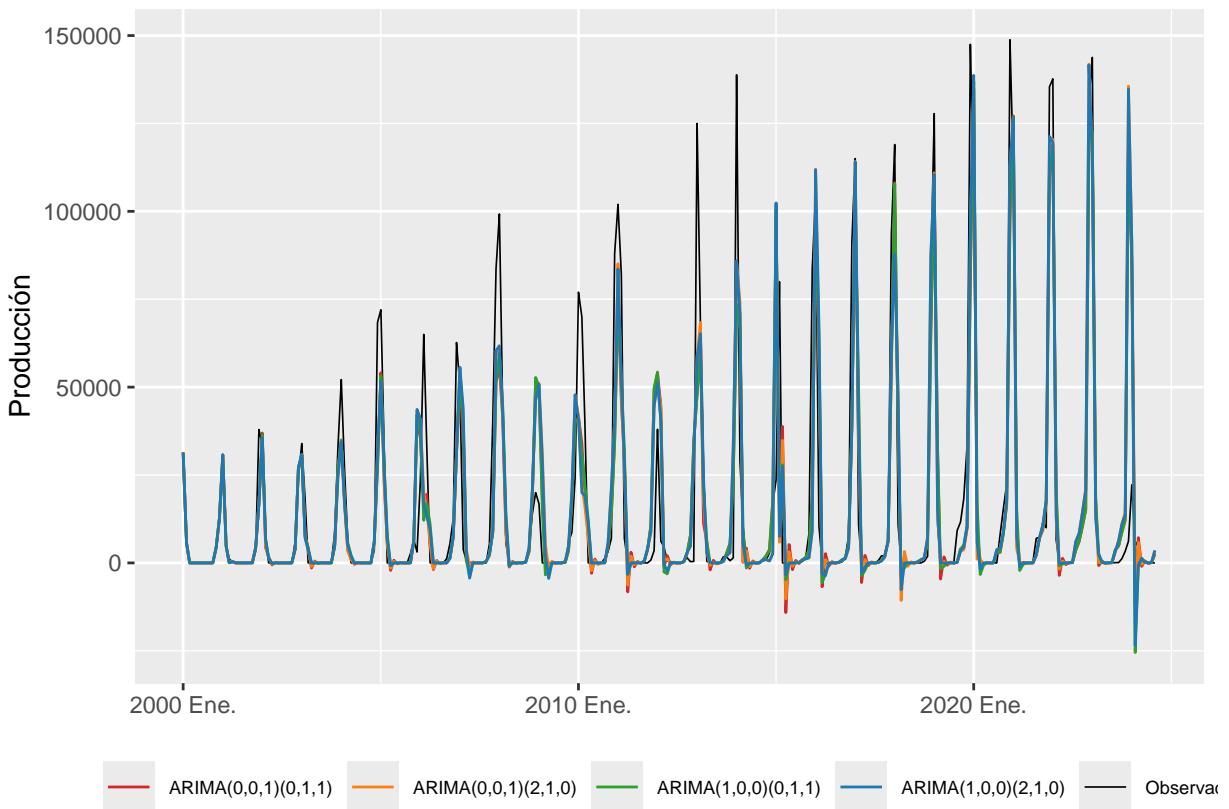
Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 33 se presenta el diagnóstico visual del modelo SARIMA(0,0,1)(0,1,1)[12]. Los residuos exhiben un comportamiento constante alrededor de 0, y todos los retardos del ACF se encuentran dentro de la banda de confianza, a excepción del lag 14, que sobresale levemente. Además, los resultados del test de Ljung-Box para cada lag se sitúan muy por encima de 0.05, lo que indica que los residuos son independientes. Sin embargo, el análisis del histograma sugiere que no se cumple el supuesto de normalidad. Estos hallazgos visuales son respaldados por la tabla Tabla 9, que presenta los resultados de los tests de los supuestos. En ella se evidencia que los residuos son independientes (Ljung-Box conjunto p = 0.989), homocedásticos (Levene p = 0.428) y tienen una media cercana a 0 (t de media p = 0.252). No obstante, los residuos no siguen una distribución normal (Jarque-Bera p = 0.000). En general, el proceso es estacionario (KPSS p = 0.100), lo que sugiere que los residuos del modelo se comportan como ruido blanco.

Comparación entre modelos SARIMA

Figura 34

Ajuste de modelos SARIMA preseleccionados sobre producción de mango (t)



Nota: Creada con R a partir de los datos del MIDAGRI.

De la Figura 34, se puede apreciar que por lo general los 4 modelos candidatos, siguen muy de cerca los valores observados de la producción de mango en el valle de San Lorenzo 2000-2024, sin embargo, no se evidencia con exactitud cuál de ellos es el mejor, por lo tanto, se considera un análisis comparativo en bases a los errores de cada modelo.

Tabla 10

Comparación entre modelos SARIMA

Modelos	ME	RMSE	MAE	MASE	RMSSE	MDAE	SMAPE
ARIMA(1,0,0)(2,1,0)	577.452	18248.91	8218.088	0.833	0.773	911.690	1.358
ARIMA(0,0,1)(2,1,0)	636.639	18219.94	8299.674	0.841	0.772	994.325	1.363
ARIMA(1,0,0)(0,1,1)	1012.636	18164.38	8358.835	0.847	0.770	1001.029	1.352
ARIMA(0,0,1)(0,1,1)	1205.543	18069.88	8441.995	0.856	0.766	1355.431	1.356

Nota: Creada con R a partir de los datos del MIDAGRI.

La Tabla 10 compara varios modelos SARIMA, destacando al SARIMA(1,0,0)(2,1,0) como el más adecuado entre modelos SARIMA para analizar la producción mensual de mango en miles de toneladas. Esta serie, caracterizada por valores extremos, producción nula por estacionalidad y variaciones abruptas, requiere un modelo que capture tanto patrones

regulares como irregularidades significativas.

En la comparación de métricas, el modelo SARIMA(1,0,0)(2,1,0)[12] muestra un desempeño competitivo y equilibrado, pues su RMSSE (0.773) es ligeramente superior al de otros modelos como el modelo SARIMA(0,0,1)(0,1,1)[12] (0.766), pero sigue siendo destacable por su capacidad de manejar variaciones extremas, además, el MAE (8218.088) y el MASE (0.833) de este modelo son los más bajos entre las opciones, lo que indica un menor error absoluto y relativo en comparación con los modelos alternativos, por ende estas últimas métricas son particularmente relevantes porque la serie incluye valores extremos y producción nula, donde es crucial minimizar los errores en términos absolutos y relativos. El SMAPE (1.358) del modelo SARIMA(1,0,0)(2,1,0)[12] es comparable al de los otros modelos, pero su bajo valor sigue siendo una ventaja en series con valores cercanos a cero, ya que esta métrica pondera los errores relativos; en cuanto al MDAE (911.690), su menor sensibilidad a valores extremos refuerza la estabilidad del modelo frente a posibles outliers, finalmente, el ME (577.452), más bajo que el de otros modelos, evidencia un sesgo reducido, lo que lo convierte en una opción confiable para capturar tendencias sin sobreestimar ni subestimar consistentemente.

El modelo SARIMA(1,0,0)(2,1,0)[12] se puede expresar de la siguiente manera:

$$(1 - \phi_1 B) (1 - \Phi_1 B^{12} - \Phi_2 B^{24}) (1 - B^{12}) y_t = \varepsilon_t$$
$$(1 - 0.3 B) (1 + 0.61 B^{12} + 0.37 B^{24}) (1 - B^{12}) y_t = \varepsilon_t$$

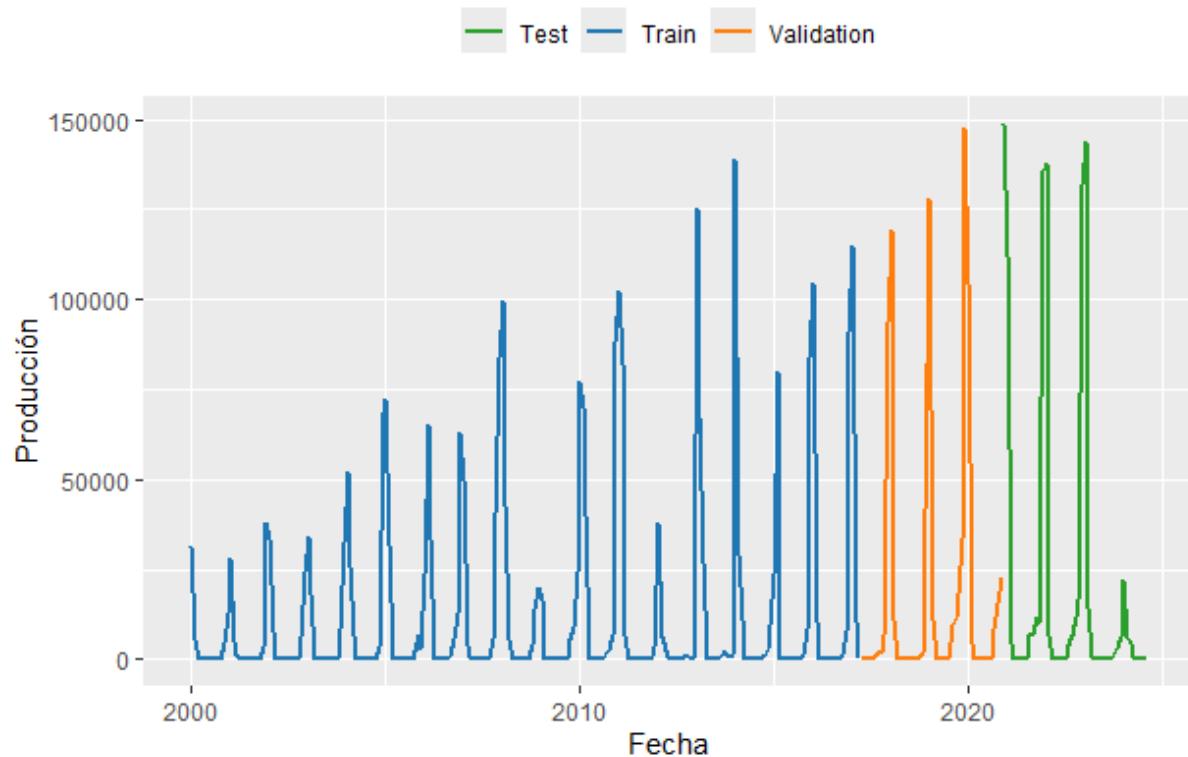
Estimación de los parámetros óptimos de redes neuronales (LSTM) para el pronóstico de la producción mensual de mango en la parte media del Valle de San Lorenzo, basados en los datos históricos comprendidos entre enero del 2000 a agosto de 2024.

En esta sección se presentan los resultados obtenidos a través del modelo RNN-LSTM Many-to-Many, el cual sigue un enfoque de predicción multi-step con salidas múltiples; los resultados incluyen la división de datos para poder entrenar, validar y probar la red LSTM, posteriormente la selección del mejor modelo LSTM se realiza mediante ajustes de parámetros e hiperparámetros acompañados de la evaluación del desempeño del modelo en términos de métricas clave, como el MAE, así como una comparación gráfica entre las predicciones y los valores observados en el conjunto de prueba. Para todo el modelado LSTM se usaron librerías de R como keras3.

Fase I: Pre-Procesamiento de los datos.

Figura 35

Partición del conjunto de datos para encontrar el mejor LSTM.



Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 35, se presenta la serie mensual de la producción de mango en el valle de San Lorenzo, la cual está compuesta por 296 observaciones, correspondientes a 23 años y 8 meses; la muestra se divide en tres subconjuntos siguiendo una proporción del 70%

para entrenamiento, 15% para validación y 15% para prueba, esto resulta en 207 datos para entrenamiento, 44 para validación y 45 para prueba, asegurando una adecuada partición para la construcción y evaluación del mejor modelo LSTM, que posteriormente se confronta con el mejor modelo SARIMA. Es importante mencionar que, para este procedimiento se estandarizan los datos después de la división, de tal manera que estos garanticen que el modelo LSTM procese la información de manera eficiente, evitando problemas de escala y mejorando la estabilidad, precisión del entrenamiento y una evaluación optima.

Fase II: Creación, entrenamiento y evaluación de modelos.

Tabla 11

Configuración de los modelos candidatos LSTM.

Parámetro/Hiperparámetro	LSTM Univariados Multi-Step			
	A	B	C	D
Longitud de la secuencia de entrada	24	12	12	12
Longitud de la secuencia de salida	17	17	17	17
Unidades de la primera capa LSTM	256	64	32	356
Unidades de la segunda capa LSTM	128	128	64	228
Función de activación	Linear	tanH	tanH	Linear
Tasa de abandono (Dropout Rate)	SI	SI	SI	SI
Número de épocas (Epochs)	50	60	50	50
Optimizador	Adam	Adam	Adam	Adam
Función de pérdida (MAE - Error absoluto medio)	MAE	MAE	MAE	MAE
Callback de parada anticipada (Early Stopping)	SI	SI	SI	SI

Nota: Elaboración propia en base a previos experimentos.

La Tabla 11 presenta configuraciones de modelos LSTM univariados a múltiples pasos con diferencias en la longitud de la secuencia de entrada (12 o 24 pasos), arquitecturas de capas (desde 32 a 356 unidades), y funciones de activación (linear o tanH), lo que refleja distintas capacidades para capturar patrones temporales complejos. Todos los modelos predicen 17 pasos futuros, utilizan regularización con dropout y optimización con Adam, mientras que el Early Stopping previene sobreajuste.

Tabla 12

Resumen de la Arquitectura del Modelo LSTM (A) creado en keras3

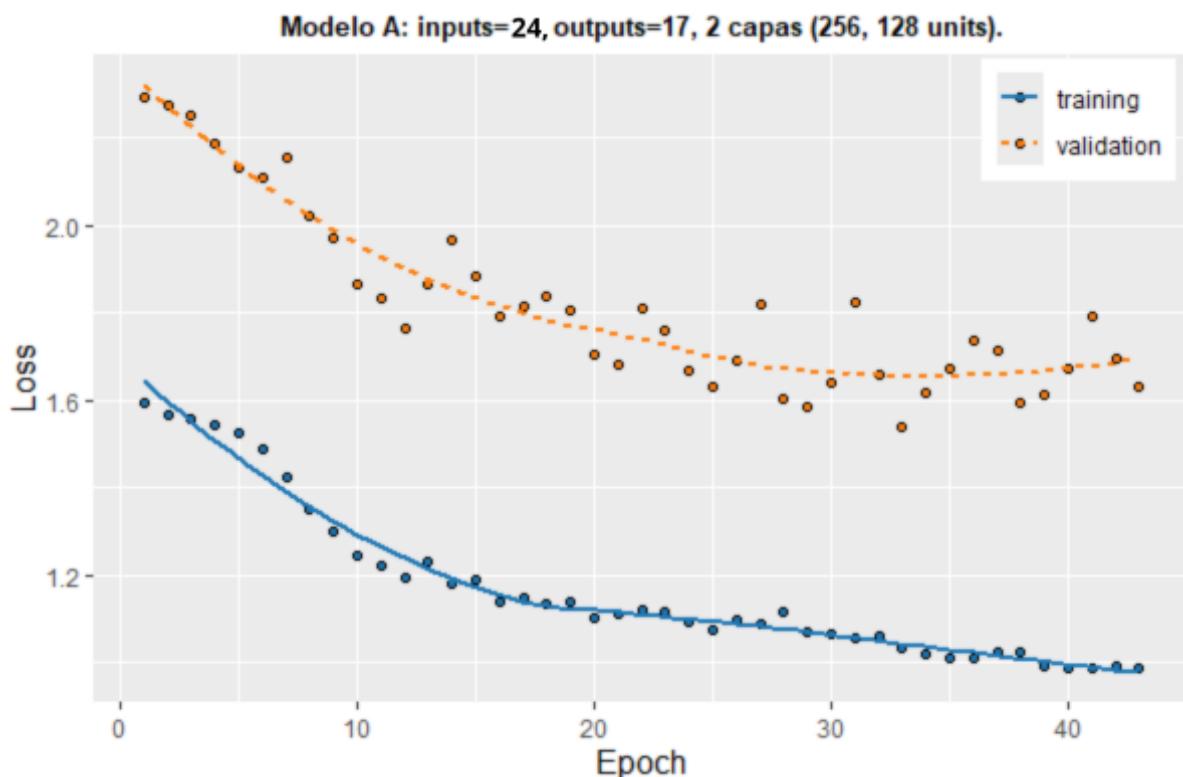
Layer (type)	Output Shape	Param Num
lstm (LSTM)	(None, 24, 256)	264,192
dropout (Dropout)	(None, 24, 256)	0
lstm_1 (LSTM)	(None, 128)	197,120
dense (Dense)	(None, 17)	2,193

Nota: Total de parámetros: 1,390,517 (5.30 MB). Parámetros entrenables: 463,505 (1.77 MB). Parámetros no entrenables: 0 (0.00 B). Parámetros del optimizador: 927,012 (3.54 MB).

La Tabla 12 muestra la arquitectura del modelo LSTM (A) obtenida en Rstudio con la librería Keras3, la que está compuesta por dos capas LSTM (256 y 128 unidades) seguidas de una capa densa para generar las 17 predicciones; así mismo, el modelo tiene un total de 1,390,517 parámetros, de los cuales 463,505 son entrenables, y 927,012 corresponden a parámetros del optimizador. Este diseño balancea capacidad de modelado y regularización para evitar el sobreajuste.

Figura 36

Entrenamiento y validación de modelo LSTM (A).

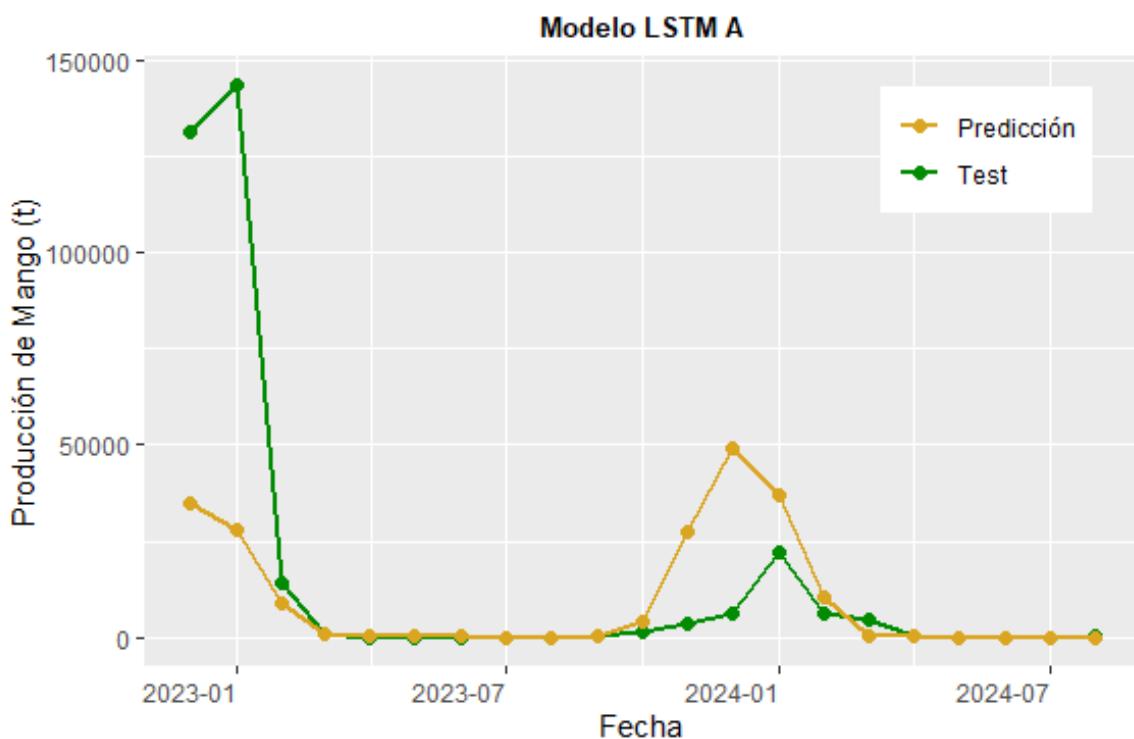


Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 36 el modelo LSTM muestra un entrenamiento adecuado, con una disminución sostenida de la pérdida en entrenamiento y una pérdida de validación relativamente estable tras unas 25 épocas, sin evidencia clara de sobreajuste. Las fluctuaciones en la curva de validación son leves y podrían deberse a la variabilidad de los datos, indicando que el modelo generaliza razonablemente bien.

Figura 37

Prueba del modelo LSTM A con 24 inputs.



Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 37 se presenta la prueba del modelo, donde se utilizaron 24 entradas, lo que implica que los primeros 24 datos de prueba no se predicen, sino que la red predice el valor inmediatamente posterior, alcanzando a predecir los últimos 21 valores de la serie de producción de mango en miles de toneladas. Se observa que la red LSTM logra ajustar adecuadamente los valores nulos, correspondientes a la ausencia de producción, y sigue el comportamiento estacional de la serie; sin embargo, no logra ajustar con precisión los valores extremos de los meses con mayor producción de mango.

Tabla 13

Resumen de la Arquitectura del Modelo LSTM (B) creado en Keras3.

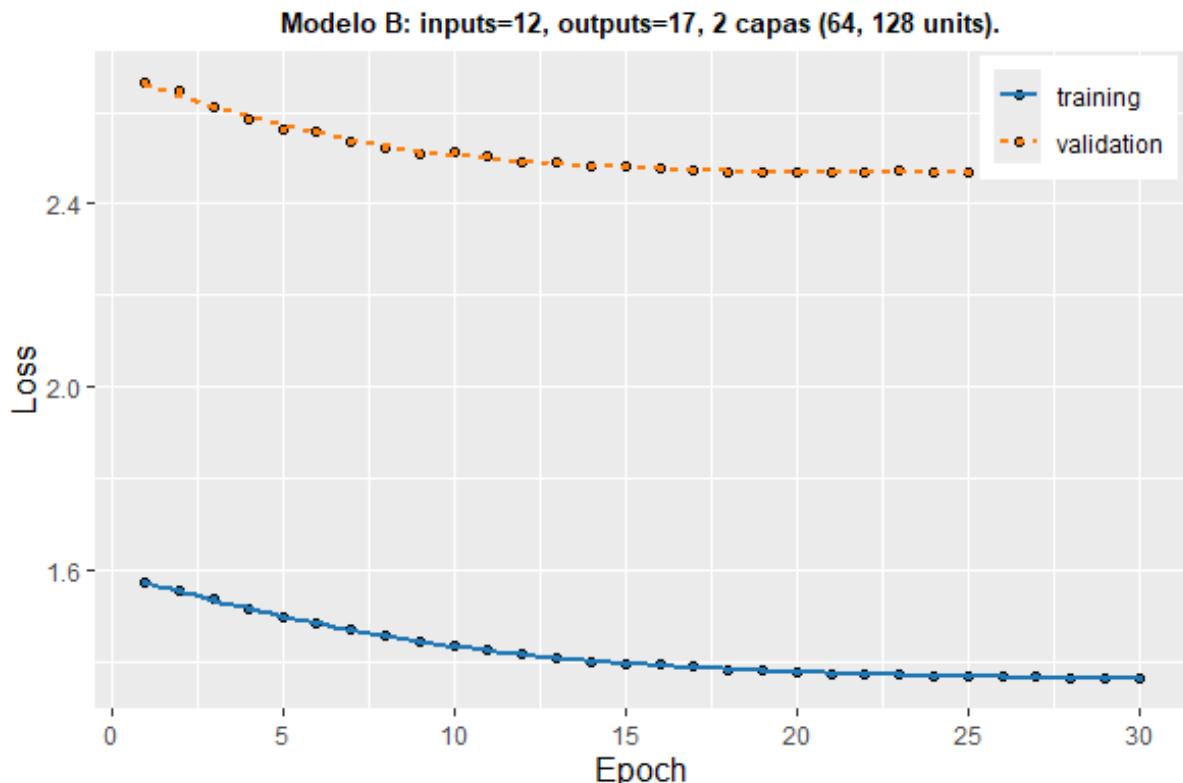
Layer (type)	Output Shape	Param Num
lstm_6 (LSTM)	(None, 12, 64)	16,896
dropout_3 (Dropout)	(None, 12, 64)	0
lstm_7 (LSTM)	(None, 128)	98,816
dense_3 (Dense)	(None, 17)	2,193

Nota: Total de parámetros: 353,717 (1.35 MB). Parámetros entrenables: 117,905 (460.57 KB). Parámetros no entrenables: 0 (0.00 B). Parámetros del optimizador: 235,812 (921.14 KB).

La Tabla 13 muestra la arquitectura del modelo LSTM (B), que consta de dos capas LSTM, una con 64 unidades y otra con 128 unidades, seguidas por una capa densa para generar las 17 predicciones. Este modelo tiene un total de 353,717 parámetros, de los cuales 117,905 son entrenables, y 235,812 corresponden a parámetros del optimizador, con un diseño más compacto en comparación con el modelo anterior. Esta configuración busca un equilibrio entre eficiencia en la capacidad de modelado y la prevención del sobreajuste mediante la regularización.

Figura 38

Entrenamiento y validación de modelo LSTM (B).

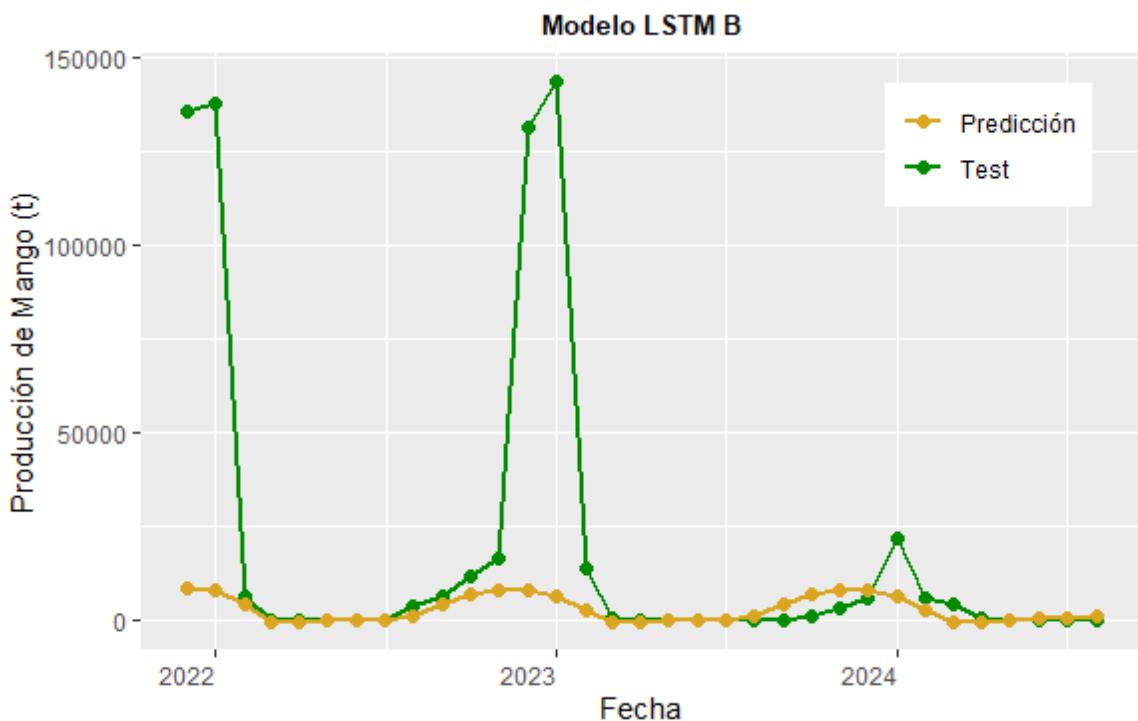


Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 38 se muestra el modelo LSTM (B), evidenciándose un comportamiento diferente: la pérdida de entrenamiento (línea azul) disminuye de manera constante, indicando aprendizaje progresivo; sin embargo, la pérdida de validación (línea naranja) permanece prácticamente plana y no muestra una mejora significativa a lo largo de las épocas, lo que podría indicar que el modelo tiene dificultades para capturar patrones relevantes en los datos de validación o que el conjunto de datos es insuficiente o poco representativo.

Figura 39

Prueba del modelo LSTM B con 12 inputs.



Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 39, se observa que el modelo LSTM (B) no está prediciendo adecuadamente los valores del conjunto de datos del test. Si bien, hay indicios de comportamiento estacional, sin embargo, el modelo no logra ajustarse de manera efectiva a los valores observados en el conjunto de test, lo cual es evidente visualmente. Es importante destacar que, debido a los 12 inputs, el modelo está prediciendo los últimos 33 valores de la serie.

Tabla 14

Resumen de la Arquitectura del Modelo LSTM (C) creado en Keras3.

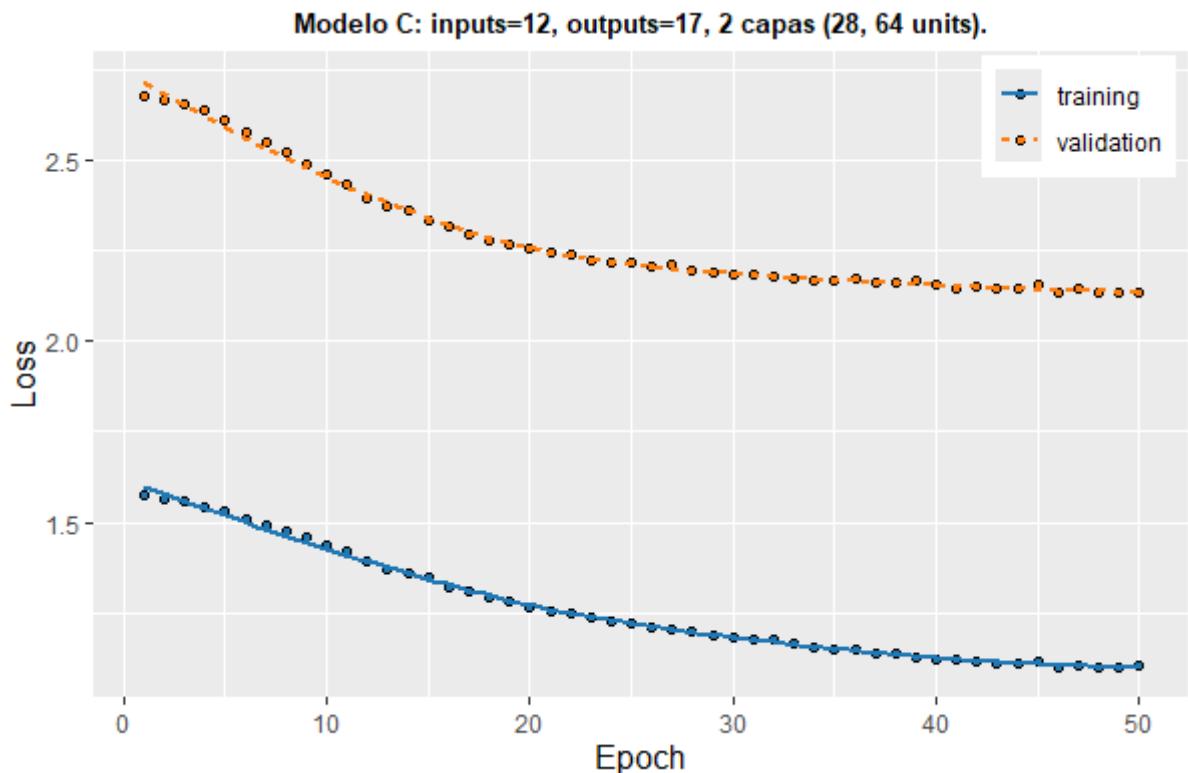
Layer (type)	Output Shape	Param Num
lstm_14 (LSTM)	(None, 12, 32)	4,352
dropout_7 (Dropout)	(None, 12, 32)	0
lstm_15 (LSTM)	(None, 64)	24,832
dense_7 (Dense)	(None, 17)	1,105

Nota: Total de parámetros: 90,869 (354.96 KB). Parámetros entrenables: 30,289 (118.32 KB). Parámetros no entrenables: 0 (0.00 B). Parámetros del optimizador: 60,580 (236.64 KB).

La Tabla 14 presenta la arquitectura del modelo LSTM (C) compuesta por dos capas LSTM (32 y 64 unidades), seguidas de una capa densa que permite generar las 17 predicciones. El modelo tiene un total de 90,869 parámetros, de los cuales 30,289 son entrenables y 60,580 corresponden al optimizador. Esta arquitectura con menos parámetros que los modelos anteriores, busca un equilibrio entre la simplicidad y la capacidad de modelado, manteniendo una estructura eficiente para la tarea de predicción.

Figura 40

Entrenamiento y validación de modelo LSTM (C)



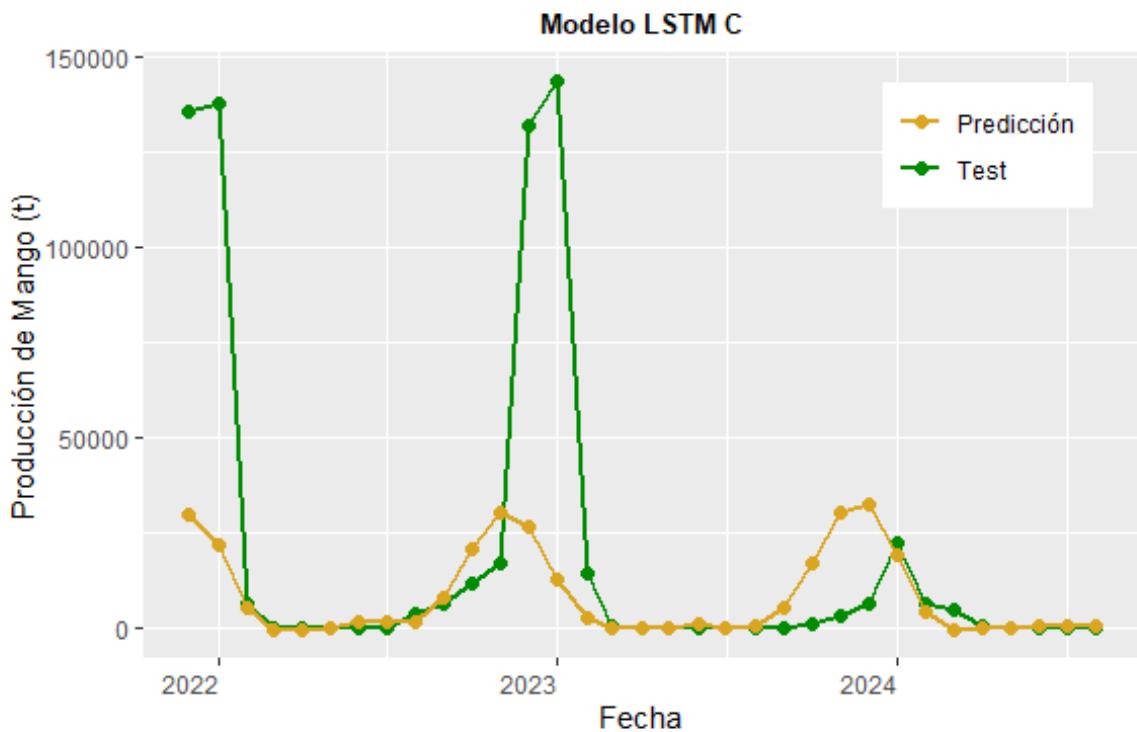
Nota: Creada con R a partir de los datos del MIDAGRI.

La Figura 40 presenta los resultados del entrenamiento y validación del modelo

LSTM (C). Se observa un entrenamiento casi adecuado y balanceado, pues la pérdida de entrenamiento (línea azul) disminuye de manera ligera y constante a lo largo de las épocas, mientras que la pérdida de validación (línea naranja) también disminuye progresivamente pero aun así no logra disminuir lo suficiente pues su ritmo es ligeramente más lento, y no logra acercarse a la curva de entrenamiento, lo que proporciona indicios de que la red no está capturando el comportamiento de la serie.

Figura 41

Prueba del modelo LSTM C con 12 inputs.



Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 41, se observa la prueba del modelo LSTM (C) con 12 inputs quien predice correctamente los valores nulos, correspondientes a la ausencia de producción de mango, y también intenta captar los picos estacionales que ocurren entre noviembre y enero. Sin embargo, a pesar de estos intentos, la red no logra predecir con precisión los valores más altos de producción en esos meses, lo que sugiere que el modelo tiene dificultades para ajustar correctamente los picos estacionales.

Tabla 15

Resumen de la Arquitectura del Modelo LSTM (D) creado en Keras3.

Layer (type)	Output Shape	Param Num
lstm_40 (LSTM)	(None, 12, 356)	509,792
dropout_26 (Dropout)	(None, 12, 356)	0
lstm_41 (LSTM)	(None, 228)	533,520
dense_26 (Dense)	(None, 17)	3,893

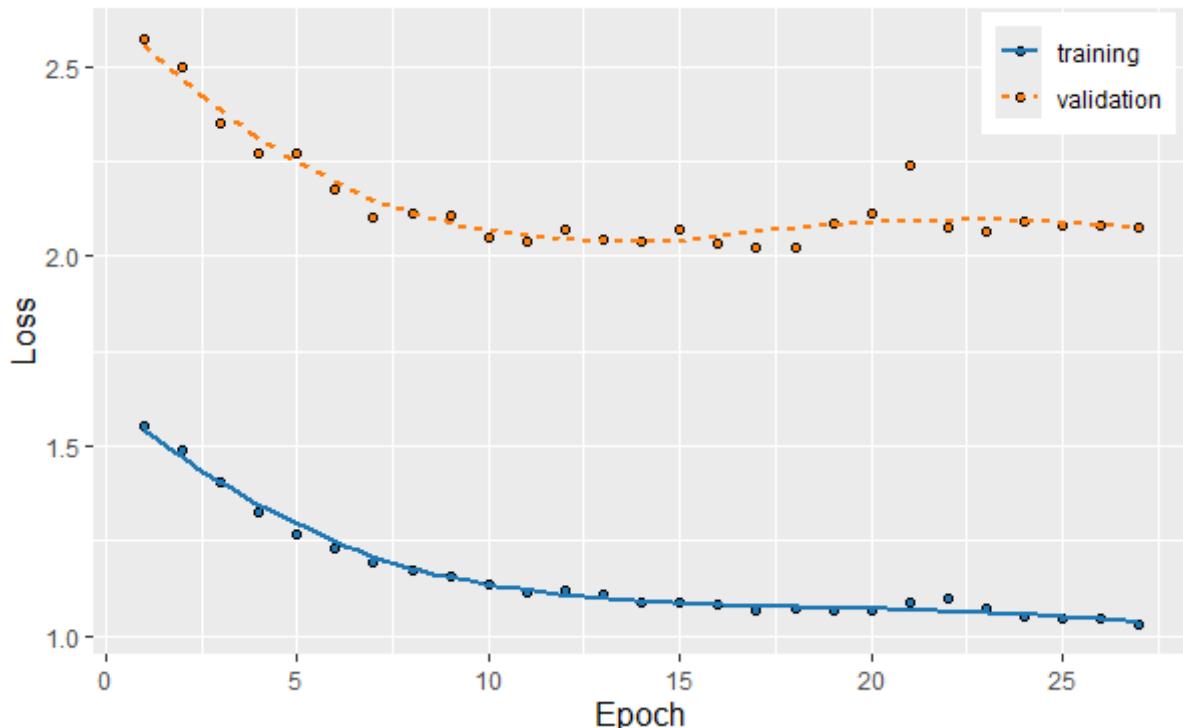
Nota: Total de parámetros: 3,141,617 (11.98 MB). Parámetros entrenables: 1,047,205 (3.99 MB). Parámetros no entrenables: 0 (0.00 B). Parámetros del optimizador: 2,094,412 (7.99 MB).

La Tabla 15 muestra la arquitectura del modelo LSTM (D), que consta de dos capas LSTM, la primera con 356 unidades y la segunda con 228 unidades, seguidas de una capa densa que genera las 17 predicciones; por otro lado, este modelo cuenta con un total de 3,141,617 parámetros, de los cuales 1,047,205 son entrenables y 2,094,412 corresponden a los parámetros del optimizador. Con una arquitectura más compleja y un número significativo de parámetros entrenables, este modelo está diseñado para capturar patrones más complejos en los datos a costa de un mayor requerimiento computacional.

Figura 42

Entrenamiento y validación de modelo LSTM (D)

Modelo D: inputs=12, outputs=17, 2 capas (356, 228 units).

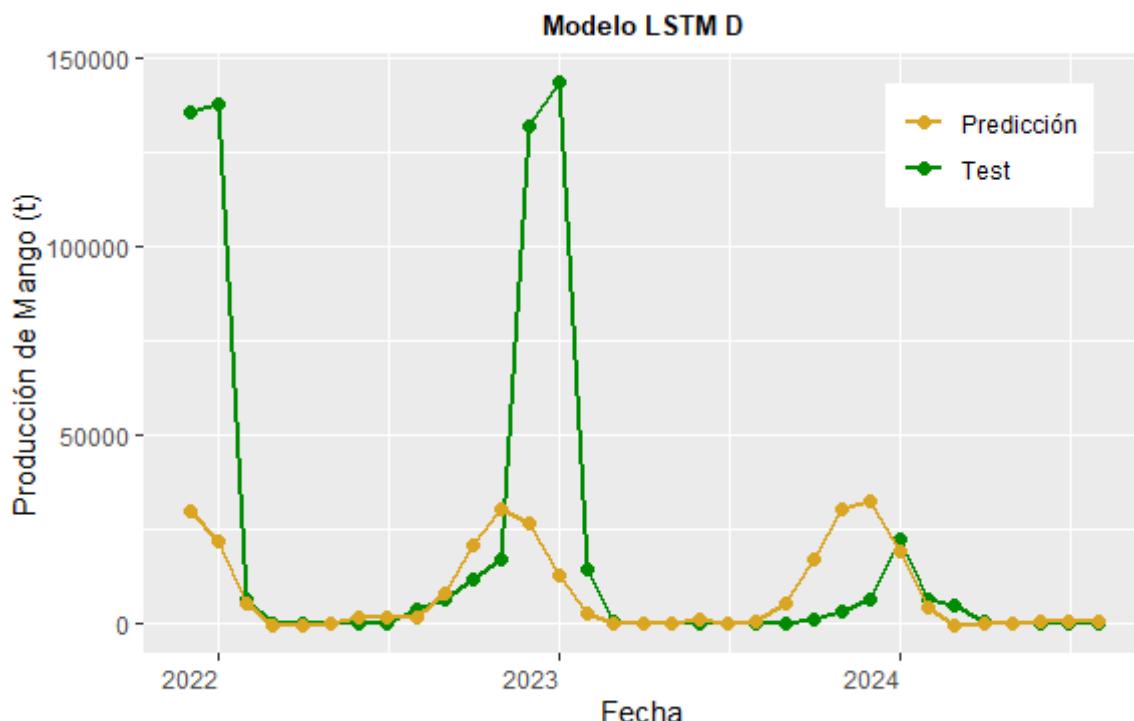


Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 42, se observa el entrenamiento y la validación del modelo LSTM (D), donde ambas curvas disminuyen gradualmente hasta el epoch 15, sin embargo, a partir de ese punto, las curvas se estabilizan y dejan de mostrar una disminución continua, lo que indica que el proceso de entrenamiento se detiene en el epoch 27, debido a la intervención del Early Stopping, que previene el sobreajuste al detectar que la mejora en la función de pérdida se ha estancado.

Figura 43

Prueba del modelo LSTM D con 12 inputs.



Nota: Creada con R a partir de los datos del MIDAGRI.

En la Figura 43, se muestra la prueba del modelo LSTM (D) con 12 inputs, el cual logra predecir de manera casi precisa los valores nulos, reflejando la ausencia de producción de mango; además, el modelo intenta identificar los picos estacionales entre noviembre y enero, pero no consigue ajustar adecuadamente los valores más altos de producción en esos meses.

Tabla 16

Monitoreo del Desempeño de los Modelos en Función de los Parámetros e Hiperparámetros.

Conjuntos	A	B	C	D
Entrenamiento	1.021	1.372	1.090	1.094
Validación	1.540	2.467	2.132	2.021
Prueba	1.126	1.928	2.064	2.002

Nota: Estos valores representan la evolución de la función de perdida (MAE)

La Tabla 16 presenta el monitoreo del desempeño de los modelos LSTM (A, B, C, D) en función de los parámetros e hiperparámetros utilizados y mostrados en las curvas de entrenamiento y validación. Por consiguiente, los valores de MAE corresponden a las métricas extraídas de las curvas de entrenamiento y validación, sin embargo, es importante señalar que estos valores no se encuentran en su escala original. A partir de estas métricas, se puede observar cómo el modelo (A) muestra el mejor desempeño en comparación con los otros modelos en los conjuntos de entrenamiento y prueba, mientras que los modelos (B), (C) y (D) presentan mayores errores, especialmente en los conjuntos de validación, por lo tanto, dado que no existe una variabilidad entre los 3 conjuntos del modelo LSTM (A).

Fase 3: Comparación y selección del mejor modelo LSTM

Tabla 17

Evaluación de los modelos LSTM sobre conjunto de prueba (Test vs Predicciones).

Métricas	A	B	C	D
MASE	1.119	1.218	1.203	1.218
MAE	14959.886	18094.563	17878.661	18091.381
MDAE	1034.715	1809.062	1487.025	2334.676
SMAPE	1.354	1.526	1.476	1.541

Nota: Modelos evaluados "test vs predicciones" usando R

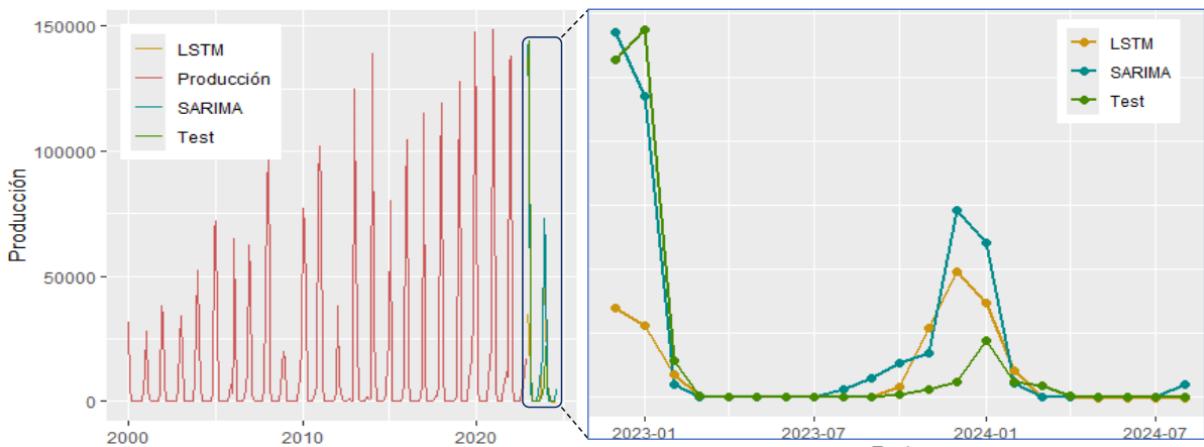
La Tabla 17 resume el desempeño de los modelos LSTM (A, B, C y D) sobre el conjunto de test. El modelo A destaca con los mejores resultados: menor MASE (1.119), MAE (14,959.886), MDAE (1034.715) y SMAPE (1.354), lo que indica mayor precisión y mejor ajuste. En conclusión, el modelo LSTM (A), con 24 inputs, presenta los mejores parámetros/hiperparámetros para predecir la producción mensual de mango.

Contrastación de los modelos utilizando medidas de precisión basadas en los errores del modelo y pronosticar la producción mensual de mango desde setiembre de 2024 hasta enero de 2026.

Para este caso, dado que se comparan 2 clases de modelos totalmente distintos, “el seleccionado previamente entre modelos SARIMA y entre modelos LSTM”, se optó por una división de datos de tal manera que ambos modelos fueron comparables, es así que ambos modelos se ejecutaron, pero dejando el conjunto test de 21 últimos meses para probar y elegir finalmente el mejor modelo de los 2.

Figura 44

Comparación de los mejores modelos LSTM (A) y SARIMA frente al test



Nota: Creada con R a partir de los datos del MIDAGRI.

Tabla 18

Contrastación de los modelos mediante métricas basadas en los errores del modelo.

Modelos	MASE	MAE	MDAE	SMAPE	RMSE
LSTM (A)	1.119	14959.886	1034.715	1.354	34876.800
SARIMA (1,0,0)(2,1,0)[12]	0.9461713	9478.744	398.321	1.332687	18614.440

Nota: Creada con R a partir de los datos del MIDAGRI.

De la Figura 44, se aprecia la comparación visual de los modelos de pronósticos, LSTM(A) y SARIMA(1,0,0)(2,1,0)[12], frente al test, mencionando que el test está compuesto por 21 meses (debido al input=24), con los cuales se probó la red LSTM previo al entrenamiento y validación. Dado que la mejor red LSTM(A) se probó con esos parámetros, estos datos test se usaron también para el SARIMA, garantizando una comparación equitativa. Los resultados evidencian que ambos modelos logran capturar la temporalidad de la serie, especialmente en las épocas de ausencia de producción; sin embargo, el modelo LSTM presenta dificultades para predecir valores extremos, como

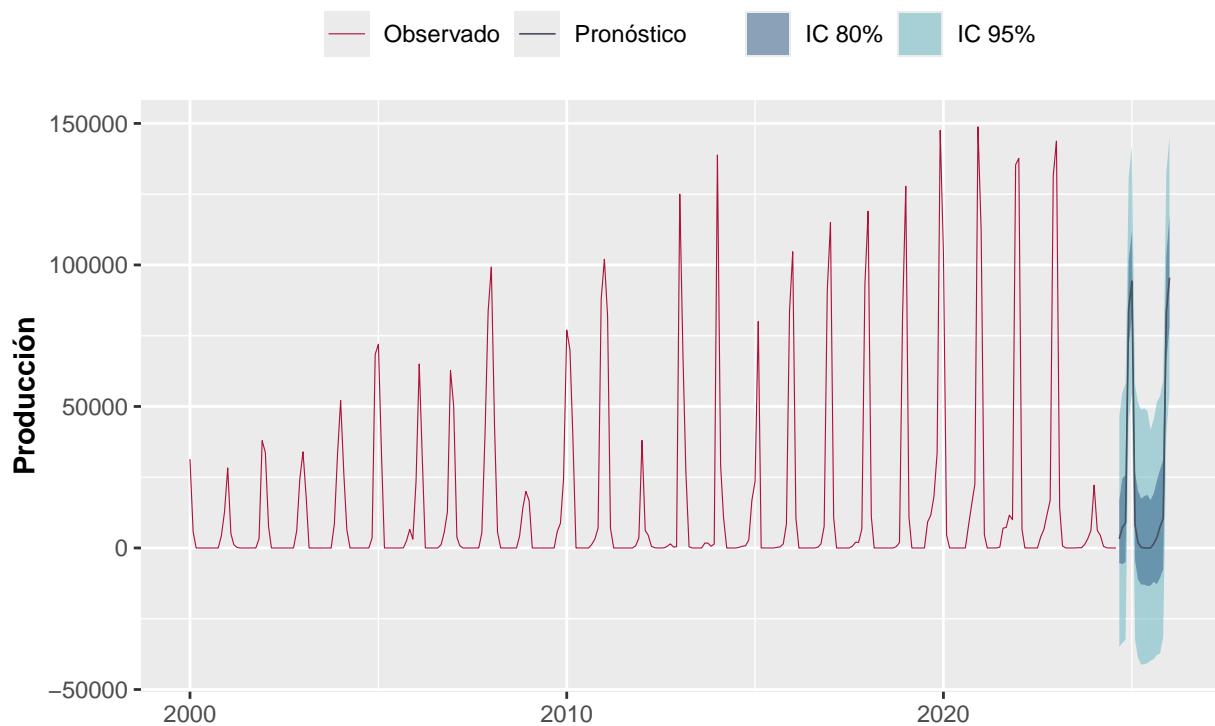
los observados en diciembre de 2022 y enero de 2023, meses en los cuales no alcanza los valores de producción observados. Este comportamiento puede deberse a la limitada cantidad de datos utilizados para su entrenamiento. Por otro lado, el modelo SARIMA muestra un mejor desempeño visual, ya que logra capturar con mayor precisión los meses de alta producción. A pesar de estas diferencias, ambos modelos reflejan adecuadamente el patrón de baja producción reportado entre diciembre de 2023 y febrero de 2024. Para reforzar esta comparación visual, en la Tabla 18 se observan las métricas de evaluación, las cuales confirman que el modelo SARIMA tiene un mejor desempeño en términos de error absoluto y error cuadrático medio, en particular, el SARIMA presenta un MAE (9478.74) y un RMSE (18614.44) significativamente menores en comparación con el LSTM, que tiene un MAE (14959.89) y un RMSE (34876.8). Además, el SARIMA alcanza un MASE (0.946) más bajo que el LSTM (1.119), lo que indica un menor error relativo respecto a un modelo de referencia simple, por lo tanto, el SARIMA muestra un desempeño superior tanto visual como cuantitativamente, destacándose particularmente en la predicción de valores extremos y en métricas clave como MAE y RMSE.

Dada esta situación, se procede a pronosticar la producción mensual de mango utilizando el modelo SARIMA (1,0,0) (2,1,0) [12]; sin embargo, debido a que este modelo no cumple con el supuesto de normalidad de los residuos, se opta por implementar intervalos de pronóstico basados en el método **Bootstrap**.

Este enfoque no paramétrico permite generar intervalos sin depender de distribuciones teóricas, lo cual resulta especialmente útil cuando los errores no siguen una distribución normal.

Figura 45

Pronóstico de la producción (t) de mango usando SARIMA $(1,0,0)(2,1,0)$ [12], desde agosto de 2024 a enero de 2026.



Nota: Creada con R a partir de los datos del MIDAGRI.

Tabla 19

Pronósticos mensuales del modelo SARIMA $(1,0,0)(2,1,0)$ [12] con intervalos de confianza.

Fecha	Pronóstico	IC 80% Inf	IC 80% Sup	IC 95% Inf	IC 95% Sup
Sep 2024	3,239.58	-6,268.01	20,229.19	-37,426.41	51,090.10
Oct 2024	7,267.75	-5,766.16	25,376.37	-34,064.96	56,354.03
Nov 2024	8,936.35	-5,309.12	27,744.31	-31,967.13	57,946.62
Dic 2024	84,279.94	70,411.62	103,315.31	42,887.57	133,434.81
Ene 2025	94,299.06	81,315.11	112,189.03	54,421.39	142,380.96
Feb 2025	8,242.60	-4,912.42	25,753.02	-32,838.71	55,274.30
Mar 2025	1,914.02	-11,427.80	18,742.27	-39,171.45	46,727.92
Abr 2025	232.75	-13,463.00	17,757.91	-40,355.87	49,094.95
May 2025	19.35	-14,281.04	18,205.31	-40,786.34	47,254.81
Jun 2025	0.00	-14,055.10	18,686.23	-41,849.88	49,529.47
Jul 2025	0.00	-14,094.73	18,651.75	-40,668.55	48,688.18
Ago 2025	1,472.69	-12,090.99	19,084.67	-40,424.53	49,426.98
Sep 2025	3,702.81	-12,915.63	24,353.39	-37,991.40	52,673.84

Fecha	Pronóstico	IC 80% Inf	IC 80% Sup	IC 95% Inf	IC 95% Sup
Oct 2025	7,466.87	-10,911.77	30,237.95	-34,992.96	56,591.31
Nov 2025	10,520.33	-8,735.64	30,464.96	-33,107.02	58,585.80
Dic 2025	83,214.21	65,176.39	104,501.43	40,450.07	132,372.03
Ene 2026	95,478.92	79,147.80	116,260.87	52,069.69	145,866.12

Nota: Creada con R a partir de los datos del MIDAGRI.

De la Figura 45 se aprecia los pronósticos de la producción de mango en el valle de san lorenzo usando SARIMA (1,0,0) (2,1,0) [12], desde agosto de 2024 a enero de 2026, evidenciándose que los pronósticos siguen el comportamiento estacional natural de la producción de mango; además, el modelo logra captar la ausencia de producción de algunos periodos. Adicionalmente y debido a que no se cumplió el supuesto de normalidad de los residuos, la figura muestra los intervalos de pronóstico mediante la técnica de Bootstrap, usando muestras de 5000. Por otro lado, en la Tabla 19 se aprecian los pronósticos obtenidos para la producción de mango, donde se demuestra que luego de la caída significativa de la producción de mango en los años 2023 y 2024, para la campaña 2024-2025 e inicios de 2026, se prevé una recuperación de los niveles de producción. Los resultados destacan la capacidad del modelo para capturar patrones estacionales, característicos de la producción de mango, incluyendo períodos de alta y baja producción.

4.2 Discusión

El estudio se enfoca en la comparación de dos modelos, entre ellos, está el modelo SARIMA, el cual es parte de la familia de los modelos de Box Jenkins, caracterizados por su capacidad para incluir componentes estacionales, y el modelo de redes neuronales LSTM, el que es una extensión más compleja de las RNN que resuelve el problema del desvanecimiento del gradiente introduciendo una estructura de memoria más compleja. La eficiencia de esta comparación se realiza mediante la comparación de los errores (MAE, RMSE, MASE, MDAE y SMAPE).

Respecto al primer objetivo específico que consistió en realizar un análisis descriptivo de la serie temporal de la producción mensual de mango en la parte media del Valle de San Lorenzo, entre enero del 2000 a agosto de 2024, se logró obtener estadísticos descriptivos que revelan una alta variabilidad, lo que se explica por el comportamiento estacional de la producción con valores que van desde 0 hasta 148,800 toneladas; por otro lado, se evidenció que la producción se concentra principalmente en enero y diciembre, mientras que entre

abril y julio no se registra actividad. Estos resultados son consistentes con el patrón cíclico reportado por Robles y Semillan (2023) en su estudio para el pronóstico de frutas en la región Piura.

La estimación del modelo SARIMA se realizó siguiendo la metodología de Box-Jenkins. Inicialmente, se confirmó la estacionariedad de la serie temporal original mediante las pruebas de raíz unitaria ADF y KPSS. Como resultado de este análisis, se aplicó una diferencia estacional a la serie, tras lo cual se verificó nuevamente el cumplimiento de la condición de estacionariedad. Una vez garantizada la estacionariedad, se identificaron siete modelos candidatos potenciales. Sin embargo, solo cuatro de ellos superaron la etapa de diagnóstico debido a la presencia de parámetros no significativos en los demás casos. Los modelos seleccionados fueron: SARIMA(1,0,0)(2,1,0)[12], SARIMA(0,0,1)(2,1,0)[12], SARIMA(1,0,0)(0,1,1)[12], SARIMA(0,0,1)(0,1,1)[12]. El diagnóstico de estos modelos reveló que los residuos cumplían con los criterios de independencia, homocedasticidad y presentaban una media próxima a cero, mostrando un comportamiento compatible con ruido blanco. No obstante, se observó que los residuos no seguían una distribución normal. En cuanto a la capacidad predictiva, todos los modelos mostraron un buen ajuste a los datos observados. Sin embargo, el modelo SARIMA(1,0,0)(2,1,0)[12] demostró ser el más eficiente, registrando los menores errores de pronóstico (ME = 1,578. MAE = 8,217, MASE = 0.833, MDAE = 912)

Estos hallazgos concuerdan de manera significativa con los resultados obtenidos por Robles y Semillan (2023), quienes, al aplicar la metodología Box-Jenkins para el pronóstico de frutas en la región Piura, identificaron el modelo SARIMA como el más adecuado para proyectar la producción de mango, específicamente el modelo SARIMA (1,1,0)(0,1,1)[12]. Ambos estudios coinciden en que los modelos SARIMA permiten capturar adecuadamente la dinámica estacional de la producción de mango, y que su uso es idóneo para planificar y anticipar comportamientos futuros en series con estructuras cíclicas marcadas. En ese sentido, los resultados alcanzados fortalecen la validez del enfoque metodológico empleado y reafirman la utilidad del modelo SARIMA en contextos agrícolas como el del Valle de San Lorenzo, donde la producción depende de ciclos definidos por factores climáticos y biológicos.

Con respecto al objetivo específico 3, orientado a estimar los parámetros óptimos en del modelo de redes neuronales LSTM, se implementó una metodología basada en división de datos y ajuste iterativo de hiperparámetros. El conjunto de datos se dividió en

tres subconjuntos: entrenamiento (70%), validación (15%) y prueba (15%). Esta partición permitió ajustar los parámetros del modelo durante la fase de entrenamiento, para luego validar su desempeño en tiempo real para evitar sobreajuste (*overfitting*), una ventaja frente al modelo SARIMA, que no requiere un proceso iterativo de optimización de hiperparámetros. Se evaluaron cuatro arquitecturas LSTM candidatas con distintas configuraciones de capas, neuronas e hiperparámetros. Entre ellas, el modelo LSTM (A), implementado en Keras, demostró un rendimiento superior. Su arquitectura consta de dos capas LSTM (256 y 128 unidades, respectivamente), 24 variables de entrada y 17 salidas (horizonte de pronóstico). Este modelo superó significativamente a las demás configuraciones en términos de precisión predictiva, validando su selección como la mejor opción para la serie temporal analizada.

Estos resultados complementan y amplía los aportes de Almeyda (2022), demostrando que el uso de redes neuronales LSTM puede ser efectivo, pero que su eficiencia relativa frente a modelos estadísticos tradicionales dependerá de las características particulares de la serie temporal analizada y del enfoque metodológico adoptado. Así mismo, los resultados se pueden contrastar, ya que comparó distintas arquitecturas de redes neuronales (RNN, LSTM, GRU y MLP) para pronosticar la demanda de banano orgánico peruano. A pesar de que Almeyda concluyó que las redes RNN ofrecían un mejor rendimiento predictivo, su estudio se circunscribió exclusivamente al uso de modelos de inteligencia artificial, sin incorporar enfoques estadísticos como los SARIMA. Además, la serie temporal utilizada por Almeyda (249 observaciones) fue más corta que la empleada en la presente investigación (296 observaciones), lo cual podría haber limitado la capacidad de generalización de modelos más complejos como las LSTM.

Finalmente, se tuvo en cuenta los últimos 21 datos del test para la comparación visual entre los dos tipos de modelos de pronósticos, SARIMA (1,0,0)(2,1,0)[12] y LSTM(A). A partir de las métricas de evaluación, se confirmó que el modelo SARIMA tiene un mejor desempeño en términos de error absoluto y error cuadrático medio; en particular, el modelo SARIMA presenta un MAE (9478.74) y un RMSE (18614.44) significativamente menores en comparación con el LSTM, que tiene un MAE (14959.89) y un RMSE (34876.8). Además, el modelo SARIMA alcanza un MASE (0.946) más bajo que el LSTM (1.119), lo que indica un menor error relativo respecto a un modelo de referencia simple. Por lo tanto, el modelo SARIMA muestra un desempeño superior tanto visual como cuantitativamente, destacándose particularmente en la predicción de valores extremos y en métricas clave como MAE y RMSE. Estos resultados nos llevan a elegir al modelo SARIMA (1,0,0) (2,1,0) [12] para pronosticar

la producción mensual de mango; sin embargo, debido a que este modelo no cumple con el supuesto de normalidad de los residuos, se optó por implementar intervalos de pronóstico basados en el método Bootstrap.

Esta conclusión encuentra respaldo en investigaciones recientes sobre modelado de cultivos. El trabajo de Patrick et al. (2023) en plantaciones de plátano tanzanas demostró que la inclusión de variables climáticas exógenas mediante modelos SARIMAX y LSTM permitía capturar patrones complejos, aunque con un desempeño final inferior al enfoque State Space. Dicho hallazgo refuerza nuestra observación sobre la importancia de adaptar la selección de modelos a las particularidades de cada sistema agrícola, donde factores como la disponibilidad de datos auxiliares o la intensidad estacional pueden inclinar la balanza hacia soluciones técnicas distintas. En similar dirección apuntan los resultados de Ramos (2020) en el ámbito tributario peruano, donde un SARIMA(2,1,0)(1,1,1)[12] alcanzó un R^2 de 0.957, superando a alternativas neuronales - coincidencia particularmente relevante pues, al igual que en nuestro caso, el componente estacional resultó decisivo para la precisión predictiva.

La información analizada permite postular un principio metodológico trascendente: la elección entre modelos estadísticos y de inteligencia artificial no responde a una jerarquía universal, sino a cada contexto particular, donde interactúan la longitud de la serie, la complejidad de los patrones subyacentes y la disponibilidad de variables explicativas. Nuestros resultados, en consonancia con la literatura citada, plantean que las redes neuronales despliegan su máximo potencial cuando se dispone de volúmenes sustanciales de datos y patrones no lineales marcados, mientras que los enfoques SARIMA mantienen notable eficacia en contextos de muestras moderadas con estacionalidad pronunciada.

CONCLUSIONES

La serie temporal de la producción mensual de mango en la parte media del Valle de San Lorenzo muestra una alta dispersión alrededor de la media, el cual se evidencia en la desviación estándar de 32 368.201. Así mismo, se observa que la serie presenta una fuerte estacionalidad acompañada de meses sin producción lo cual se justifica por la ausencia de producción. Por otro lado, la serie refleja clara estacionariedad y una ligera tendencia, la cual, no se percibe debido a los períodos de ausencia de producción.

El modelo estimado para el pronóstico de la producción de mango en el valle de San Lorenzo es un modelo SARIMA(1,0,0)(2,1,0)[12], el cual muestra un desempeño competitivo y equilibrado, obteniendo un menor error absoluto y relativo en comparación con los modelos alternativos.

En la aplicación de las redes neuronales LSTM a la serie de tiempo de la producción de mango en el valle de San Lorenzo se estimó la arquitectura del modelo en Keras, donde se eligió la opción A compuesta por dos capas LSTM (256 y 128 unidades) seguidas de una capa densa para generar las 17 predicciones, así mismo, el modelo tiene un total de 1,390,517 parámetros, de los cuales 463,505 son entrenables, y 927,012 corresponden a parámetros del optimizador; además, presentó un MASE de 1.119, acompañado de un MAE de 14959.886; los cuales, son menores en comparación a las demás opciones de los modelos evaluados.

El modelo elegido para realizar el pronóstico de la producción de mango en el valle de San Lorenzo es el modelo SARIMA(1,0,0)(2,1,0)[12], el cual presenta un MAE (9478.74) y un RMSE (18614.44) significativamente menores en comparación con el LSTM.

El modelo SARIMA (1,0,0)(2,1,0)[12] se identificó como el más eficiente para el pronóstico de la producción mensual de mango en el Valle de San Lorenzo. Según las estimaciones, se prevé una producción de 1,472.69 toneladas para agosto de 2025, con un incremento progresivo en los meses siguientes: 3,702.81 toneladas en septiembre, 7,466.87 toneladas en octubre y 10,520.33 toneladas en noviembre. Sin embargo, en diciembre se proyecta una leve disminución, con una producción estimada de 8,3214.21 toneladas, seguida de un repunte en enero de 2026, alcanzando las 9,547.92 toneladas.

RECOMENDACIONES

Según los resultados del estudio, se procede a realizar las siguientes recomendaciones:

- Se recomienda realizar un seguimiento constante de los pronósticos, incorporando datos actualizados y empleando el modelo que demostró mayor eficacia en la presente investigación; para su implementación práctica, se sugiere utilizar la aplicación web desarrollada como parte complementaria del estudio, cuya descripción se encuentra en el Anexo 6. Esta herramienta permite actualizar periódicamente los pronósticos y ajustar los parámetros del modelo SARIMA de acuerdo con la evolución de los datos, lo que favorecerá una toma de decisiones más informada, oportuna y eficiente en la gestión de la producción.
- Para realizar un estudio más óptimo en investigaciones futuras, se recomienda integrar variables climáticas como precipitación, temperatura máxima y mínima, humedad relativa y humedad del suelo; ya que, pueden ser fundamentales para capturar patrones no explicados exclusivamente por la estacionalidad de la serie. Así mismo, ello, hace referencia al uso del modelo SARIMAX (una extensión de SARIMA que permite incluir variables exógenas), el cual, podría mejorar la precisión del pronóstico.
- Se recomienda implementar estaciones meteorológicas locales en el Valle de San Lorenzo, el cual, permitirá obtener datos precisos y en tiempo real sobre las condiciones climáticas. De esta forma, les ayudará para que puedan basarse en los pronósticos para planificar actividades agrícolas, como riego y fertilización, de acuerdo con las predicciones de producción.

REFERENCIAS BIBLIOGRÁFICAS

- Almeyda, E. M. (2022). *Pronóstico de la demanda internacional del banano orgánico de Perú usando algoritmos de Machine Learning*. <https://hdl.handle.net/11042/5718>
- Amat Rodrigo, J., y Escobar Ortiz, J. (2025). *skforecast* (Versión 0.16.0). <https://doi.org/10.5281/zenodo.8382788>
- Aragón, D. F. (2022). *Estudio de tendencias de mercado - Mango*. Instituto Nacional de Innovación Agraria. <https://repositorio.inia.gob.pe/handle/20.500.12955/2025>
- Arce, B. R. J. B., Granda Wong, C. A., Javier Alva, J., y San Martin Zapata, C. E. (2019). Manejo Integrado del Cultivo de Mango Kent. *Instituto Nacional de Innovación Agraria*. <https://repositorio.inia.gob.pe/handle/20.500.12955/966>
- Bowerman, B. L., y Koehler, A. B. (2007). *Pronósticos, Series de Tiempo Y Regresión: Un Enfoque Aplicado*. International Thomson Editores, S.A. de C.V.
- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., y Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control* (5.^a ed., p. 720). Wiley.
- Brockwell, P. J., y Davis, R. A. (2006). *Introduction to Time Series and Forecasting*. Springer Science & Business Media.
- Canova, F., y Hansen, B. E. (1995). Are seasonal patterns constant over time? A test for seasonal stability. *Journal of Business & Economic Statistics*, 13(3), 237-252. <https://doi.org/10.1080/07350015.1995.10524599>
- Castañeda, W. A., Polo Escobar, B. R., Vega Huincho, F., Castañeda, W. A., Polo Escobar, B. R., y Vega Huincho, F. (2007). Redes neuronales artificiales: una medición de aprendizajes de pronósticos como demanda potencial. *Universidad, Ciencia y Tecnología*, 27(118), 51-60. <https://doi.org/10.47460/uct.v27i118.686>
- Castro, J. J., Gómez Macho, L. K., y Camargo Casallas, E. (2023). La investigación aplicada y el desarrollo experimental en el fortalecimiento de las competencias de la sociedad del siglo XXI. *Tecnura*, 27(75), 140-157. <https://doi.org/10.14483/22487638.19171>
- Chatfield, C. (2016). *The Analysis of Time Series: An Introduction, Sixth Edition*. CRC Press. https://books.google.com.pe/books?id=qKzyAbdaDFAC&pg=PA11&source=gbs_toc_r&cad=2#v=onepage&q&f=false
- Chilón, A. L. (2023). *Redes neuronales recurrentes y modelos arima para el pronóstico de la inflación en el Perú*. <https://hdl.handle.net/20.500.14414/18318>
- Cho, K., Merrienboer, B. van, Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., y

- Bengio, Y. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. arXiv. <https://doi.org/10.48550/arXiv.1406.1078>
- Fernández Salguero, R. A. (2021). *Series Temporales Avanzadas: Aplicación de Redes Neuronales para el Pronóstico de Series de Tiempo*. Trabajo Fin de Máster, Máster en Estadística Aplicada, Universidad de Granada. https://masteres.ugr.es/estadistica-aplicada/sites/master/moea/public/inline-files/TFM_FernAndez%20SalgueroRicardo%20Alonzo.pdf
- García, J. (2010). *Fenología del cultivo del mango (*Mangifera indica L.*) en el alto y bajo Magdalena: bases conceptuales para su manipulación*. Corporación colombiana de investigación agropecuaria - AGROSAVIA. <https://repository.agrosavia.co/handle/20.50.0.12324/13003>
- Guerrero, V. M. (2009). *Análisis Estadístico y Pronostico de Series de Tiempo Económicas* (3a ed). JUST IN TIME PRESS.
- Hernández, R., Fernández Collado, C., y Baptista Lucio, M. del P. (2010). *Metodología de la investigación*. <http://148.202.167.116:8080/xmlui/handle/123456789/2707>
- Hesaraki, S. (2023). Long Short-Term Memory (LSTM). En *Medium*. <https://medium.com/@saba99/long-short-term-memory-lstm-fffc5eaebfdc>
- Hochreiter, S., y Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hylleberg, S., Engle, R. F., Granger, C. W., y Yoo, B. S. (1990). Seasonal integration and cointegration. *Journal of Econometrics*, 44(1-2), 215-238. [https://doi.org/10.1016/0304-4076\(90\)90080-7](https://doi.org/10.1016/0304-4076(90)90080-7)
- Hyndman, R., y Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed). OTexts: Melbourne. <https://otexts.com/fpp3/arima-reading.html>
- JUSHSAL. (2019). Breve Reseña Histórica del Valle San Lorenzo. <https://jusanlorenzo.org.pe/web/index.php/2019/07/24/breve-resena-historica-del-valle-san-lorenzo/>
- Kelleher, J. D. (2019). *Deep Learning*. MIT Press.
- Kumar, M. T., y Rao, M. C. (2023). Studies on predicting soil moisture levels at Andhra Loyola College, India, using SARIMA and LSTM models. *Environmental Monitoring and Assessment*, 195(12), 1426. <https://doi.org/10.1007/s10661-023-12080-1>
- Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., y Shin, Y. (1991). *Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root?* (Cowles Foundation Discussion Paper N.º 1222). Cowles Foundation for Research in Economics, Yale University. <https://elischolar.library.yale.edu>

u/cowles-discussion-paper-series/1222

- León, J. C. (2023). Promango: Perú produciría entre 100 mil y 120 mil toneladas de mango en la campaña 2023/2024. En *Agraria.pe Agencia Agraria de Noticias*. <https://agraria.pe/noticias/promango-peru-produciria-entre-100-mil-y-120-mil-toneladas-d-33708>
- Mahadeva, L., y Robinson, P. (2009). *Prueba de raíz unitaria para ayudar a la construcción de un modelo* (1a. ed). Centro de Estudios Monetarios Latinoamericanos (CEMLA).
- Martinez, J. A., Fajardo, A. G., Esquivel, J. S., González, D. M., Prieto, Á. G., y Rincón, D. (2020). Manejo integrado del cultivo de mango Mangifera indica L. *Revista Ciencias Agropecuarias (RCA)*, 6(1), 51-78. <https://dialnet.unirioja.es/servlet/articulo?codigo=8658103>
- Maya, M. C. (2022). *Pronóstico de series de tiempo empleando redes neuronales y meta-transferencia de aprendizaje* [{doctoralThesis}], Tesis (D.C.)–Centro de Investigación y de Estudios Avanzados del I.P.N. Departamento de Control Automático]. <https://repositorio.cinvestav.mx/handle/cinvestav/3835>
- Michel, R., Montaño, G., Mora, J., y Moncada, E. (2000). *Cultivo de mango*. <http://hdl.handle.net/11036/2470>
- MIDAGRI. (2010). Condiciones Agroclimáticas - Cultivo del Mango. *Ministerio de Desarrollo Agrario y Riego*. <http://repositorio.midagri.gob.pe:80/jspui/handle/20.500.13036/449>
- Ministerio de Desarrollo Agrario y Riego. (2023). *Perfil Productivo y Competitivo de los Principales Cultivos del Sector*. MIDAGRI. https://siae.midagri.gob.pe/portal/siae_bi/index.html
- Olah, C. (2015). *Understanding LSTM Networks*. <http://colah.github.io/posts/2015-08-Understanding-LSTMs>
- Osborn, D. R., Chui, A. P. L., Smith, J., y Birchenhall, C. R. (1988). Seasonality and the order of integration for consumption. *Oxford Bulletin of Economics and Statistics*, 50(4), 361-377. <https://doi.org/10.1111/j.1468-0084.1988.mp50004002.x>
- Pang, K. H. (2021, enero 23). *Multi-Step Multivariate Time-Series Forecasting using LSTM*. <https://pangkh98.medium.com/multi-step-multivariate-time-series-forecasting-using-lstm-92c6d22cd9c2>
- Patrick, S., Mirau, S., Mbalawata, I., y Leo, J. (2023). Time series and ensemble models to forecast banana crop yield in Tanzania, considering the effects of climate change. *Resources, Environment and Sustainability*, 14, 100138. <https://doi.org/10.1016/j.resenv.2023.100138>

- Peña, D. (2010). *Análisis de series temporales*. Alianza Editorial.
- Ramos, A. P. (2020). *Pronóstico de los ingresos tributarios mensuales del gobierno central peruano aplicando redes neuronales y modelos Sarima, en base a los años 2003 – 2018* [Tesis doctoral]. <https://repositorio.unp.edu.pe/handle/20.500.12676/2264>
- Robles, K. F., y Semillan, Y. P. (2023). *Creación de un modelo econométrico, para pronóstico de la producción de las principales frutas de la región Piura* [Tesis de licenciatura]. Universidad Nacional de Frontera. <http://repositorio.unf.edu.pe/handle/UNF/237>
- Rojas, K. (2022). *Capítulo 10 Redes Neuronales Artificiales y Aprendizaje Profundo | Ciencia de Datos para Ciencias Naturales*. https://bookdown.org/keilor_rojas/CienciaDatos/redes-neuronales-artificiales-y-aprendizaje-profundo.html
- SENAMHI. (2024). *Boletín Agroclimático del Desarrollo de la campaña agrícola 2024 Región Piura y Tumbes*. <https://cdn.www.gob.pe/uploads/document/file/5879117/5208252-riesgo-agroclimatico-enero-2024.pdf>
- Soria, E., Rodríguez, P., García Vidal, E., Vaquer, F., Vicent, J., y Vila, J. (2022). *Inteligencia artificial: Casos prácticos con aprendizaje profundo*. Ediciones de la U.
- Sotaquirá, M. (2023). Forecasting con Redes LSTM - Parte 1: tipos de predicción. En *Codificando Bits*. <https://www.codificandobits.com/blog/series-de-tiempo-redes-lstm-tipos-de-prediccion/>
- Tian, H., Wang, P., Tansey, K., Zhang, J., Zhang, S., y Li, H. (2021). An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong Plain, PR China. *Agricultural and Forest Meteorology*, 310, 108629. <https://doi.org/10.1016/j.agrformet.2021.108629>
- Vásquez, P. (2023). *ECONOMETRÍA DE LAS SERIES DE TIEMPO*.
- Villalobos, V. (2020). 4 condiciones ambientales para el cultivo del mango. En *AvoGo Consulting*. <https://avogoconsulting.com/subtropicales/4-condiciones-ambientales-claves-para-mango/>
- Villavicencio, J. (2007). *Introducción a series de tiempo*. http://www.estadisticas.gobierno.pr/iepr/LinkClick.aspx?fileticket=4_BxecUaZmg%3D
- Wang, X., Smith, K. A., y Hyndman, R. J. (2006). Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery*, 13(3), 335-364. <https://doi.org/10.1007/s10618-006-0027-7>
- Whitney, F. (1990). *Elementos de investigación* (Omega).
- Xu, L., y Chen, W. (2021). Construction and Simulation of Economic Statistics Measurement

Model Based on Time Series Analysis and Forecast [Retracted]. *Complexity*. <https://doi.org/10.1155/2021/5963516>

ANEXOS

ANEXO 01: Matriz de Consistencia

“MODELO SARIMA Y RED NEURONAL RECURRENTE PARA EL PRONÓSTICO DE LA PRODUCCIÓN DE MANGO EN EL VALLE DE SAN LORENZO, 2024 – 2026”					
Problema	Objetivo General	Objetivos	Hipótesis	Variables	Metodología
Problema General ¿Qué modelo, entre SARIMA y LSTM, es más eficiente para el pronóstico mensual de la producción de mango en el Valle de San Lorenzo, Piura, 2024 – 2026?	Determinar la eficiencia de los modelos SARIMA y redes neuronales (LSTM) para el pronóstico mensual de la producción de mango en el Valle de San Lorenzo, Piura, 2024 – 2026.		Hipótesis General Existen diferencias significativas en la precisión de los pronósticos de la producción mensual de mango de la parte media del Valle de San Lorenzo-Piura, obtenidos a partir del modelo SARIMA y las Redes Neuronales (LSTM).	Univariante: Producción de Mango	Enfoque: Cuantitativo Diseño: No experimental Nivel: Predictivo Tipo: Longitudinal Técnicas e instrumentos
Problemas Específicos ¿Cómo se comporta la serie temporal de la producción mensual de mango en la parte media del Valle de San Lorenzo desde enero del 2000 hasta agosto de 2024?	Efectuar un análisis descriptivo de la serie temporal de la producción mensual de mango en la parte media del Valle de San Lorenzo, entre enero del 2000 a agosto de 2024.	Objetivos Específicos Estimar el modelo SARIMA para el pronóstico de la producción mensual de mango en la parte media del Valle de San Lorenzo, con la metodología de Box Jenkins, basados en los datos históricos comprendidos entre enero del 2000 a agosto de 2024.			Población: Mes desde donde se empezo la Producción de mango mensuales en el valle de San Lorenzo.
	¿Cuáles son los estimadores del modelo SARIMA aplicado a la serie temporal de la producción mensual de mango en la parte media del Valle de San Lorenzo, utilizando la metodología de Box Jenkins desde enero del 2000 hasta agosto de 2024?	Estimar los parámetros óptimos de redes neuronales (LSTM) para el pronóstico de la producción mensual de mango en la parte media del Valle de San Lorenzo, basados en los datos históricos comprendidos entre enero del 2000 a agosto de 2024.			Muestra: Producción dada entre los meses de enero de 2000 a agosto de 2024, con un total de 296 valores.
	¿Cuáles son los estimadores óptimos de los parámetros de las redes neuronales LSTM que permiten pronostica la producción mensual de mango en la parte media del Valle de San Lorenzo, utilizando los datos históricos desde enero del 2000 hasta agosto de 2024?	Contrastar ambos modelos utilizando medidas de precisión basadas en los errores del modelo y pronosticar la producción mensual de mango desde setiembre de 2024 hasta enero de 2026.			

ANEXO 02: Carta de solicitud de datos

Datos solicitados preliminarmente



PERÚ
Ministerio
de Desarrollo Agrario
y Riego

Secretaría General

Oficina de Atención al
Ciudadano y Gestión Documental

"Decenio de la Igualdad de oportunidades para mujeres y hombres"
"Año del Bicentenario, de la consolidación de nuestra Independencia y de la conmemoración de las heroicas batallas de Junín y Ayacucho"

Lima, 02 de abril de 2024

CARTA Nro. 0410-2024-MIDAGRI-SG/OACID-TRANSP

Señora
Jahayra Sheridan Rodriguez Rodriguez
AA.HH. Los Medanos Mz. "H" lote 18 – Castilla
Piura. –
Jahayrarr0@gmail.com

Asunto : Solicitud de Acceso a la Información Pública

Referencia : CUT N.º 14945-2024-MIDAGRI

Tengo el agrado de dirigirme a usted; en atención a su solicitud de Acceso a la Información Pública, mediante la cual solicita: *"Información del mango en el departamento de Piura, desglosado por provincias y todos sus distritos, desde el año 1990 a febrero de 2024 de manera mensual: > Serie mensual de la producción (T) de mango"*

Al respecto, la Dirección General de Estadística, Seguimiento y Evaluación de Políticas, mediante Correo Electrónico N.º 0036-2024-MIDAGRI-DVPSDA-DGESEP/DEIA-JKCL de fecha 01.04.2024, remite en archivo digital la información disponible en la citada Dirección General, indicando que la información del año 2023 es preliminar.

Estando a lo expuesto y de conformidad con el artículo 13¹ del TUO de la Ley N° 27806, Ley de Transparencia y Acceso a la Información Pública, se remite parte de la información solicitada. (Se adjuntan copias de los documentos antes citados).

Aprovecho la ocasión para informarle que, el Ministerio de Desarrollo Agrario y Riego, cuenta con su Mesa de Partes Digital, cuyo acceso es el siguiente enlace: <https://mesadepartesdigital.midagri.gob.pe/> a través del cual, los ciudadanos pueden remitir sus solicitudes para todo tipo de trámites, sin acudir de forma presencial a un local del MIDAGRI, ahorrando tiempo, dinero y sobre todo cuidando su salud.

Finalmente, ponemos a su disposición el correo electrónico para consultas y seguimiento en el siguiente horario: lunes a viernes de 8:30 a.m. a 4:30 p.m. al info@midagri.gob.pe.

Atentamente,



Firmada digitalmente por CHARA
ARBIRO Marisol - Oficina FAU
20131372931 Hand
Motivo: Soy el autor del documento
Fecha: 02.04.2024 16:50:15 -05:00

DOCUMENTO FIRMADO DIGITALMENTE MARISOL F. CHARA ARBIRO

Directora (e)

Oficina de Atención al Ciudadano y Gestión Documental

CUT N.º: 14945-2024-MIDAGRI

MFCA/njco/mbuu

¹ Artículo 13¹... La solicitud de información no implica la obligación de las entidades de la Administración Pública de crear o producir información con la que no cuente o no tenga obligación de contar al momento de efectuarse el pedido.
En este caso, la entidad de la Administración Pública deberá comunicar por escrito que la denegatoria de la solicitud se debe a la inexistencia de datos en su poder respecto de la información solicitada. Esta Ley no faculta que los solicitantes exijan a las entidades que efectúen evaluaciones o análisis de la información que posean (...).



Esta es una copia auténtica imprimible de un documento electrónico archivado en el Ministerio de Desarrollo Agrario y Riego, aplicando lo dispuesto por el Art. 25 de D.S. 070-2013-PCM y la Tercera Disposición Complementaria Final el D.S. 026-2016-PCM. Su autenticidad e integridad pueden ser contrastadas a través de la siguiente dirección web:



Actualización de datos y data final



"Decenio de la Igualdad de oportunidades para mujeres y hombres"
"Año del Bicentenario, de la consolidación de nuestra Independencia y de la conmemoración de las heroicas batallas de Junín y Ayacucho"

Lima, 11 de setiembre de 2024

CARTA Nro. 1181-2024-MIDAGRI-SG/OACID-TRANSP

Señor
Jairon Kevin Ojeda Silupu
Pedregal S/N – Tambo Grande – Piura
Piura. –
Jairokevinojedas1999@gmail.com

Asunto: Solicitud de Acceso a la Información Pública

Referencia: CUT. N° 48518-2024-MIDAGRI

Tengo el agrado de dirigirme a usted; en atención a su solicitud de Acceso a la Información Pública, mediante la cual solicita: "Serie mensual de la producción de mango (miles de toneladas) desde enero de 2000 hasta agosto de 2024, en el departamento de Piura, distribuidos por provincias y distritos".

Al respecto, la Dirección General de Estadística, Seguimiento y Evaluación de Políticas, remite Memorando N.º 0539-2024-MIDAGRI-DVPSDA/DGESEP de fecha 11.09.2024, que contiene el enlace con parte de la información.

Estando a lo expuesto y de conformidad con el artículo 13¹ del TUO de la Ley N° 27806, Ley de Transparencia y Acceso a la Información Pública, se remite parte de la información solicitada. (Se adjunta copia del documento antes citado).

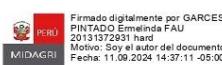
Aprovecho la ocasión para informarle que, el Ministerio de Desarrollo Agrario y Riego, cuenta con su Mesa de Parte Digital, cuyo acceso es el siguiente enlace: <https://mesadepartesdigital.midagri.gob.pe/>, a través del cual los ciudadanos pueden remitir sus solicitudes para todo tipo de trámites, sin acudir de forma presencial a un local del MIDAGRI, ahorrando tiempo, dinero y sobre todo cuidando su salud.

Finalmente, para cualquier consulta sobre esta respuesta ponemos a su disposición el correo electrónico: info@midagri.gob.pe. (Horario de Atención: lunes a viernes de 8:30 a.m. a 4:30 p.m.)

Atentamente,



Firmado digitalmente por
PINTADO Ermelinda FAU
20131372931 hard
Motivo: Doy Vº Bº
Fecha: 11.09.2024 14:17:33 -05:00



Firmado digitalmente por GARCES
PINTADO Ermelinda FAU
20131372931 hard
Motivo: Soy el autor del documento
Fecha: 11.09.2024 14:37:11 -05:00

DOCUMENTO FIRMADO DIGITALMENTE

ERMELINDA GARCES PINTADO

Directora

Oficina de Atención al Ciudadano y Gestión Documental

EGP/Njco/Mbuuh

CUT N°: 48518-2024-MIDAGRI

¹ Artículo 13º (...) La solicitud de información no implica la obligación de las entidades de la Administración Pública de crear o producir información con la que no cuente o no tenga obligación de contar al momento de efectuarse el pedido.
En este caso, la entidad de la Administración Pública deberá comunicar por escrito que la denegatoria de la solicitud se debe a la inexistencia de datos en su poder respecto de la información solicitada. Esta Ley no faculta que los solicitantes exijan a las entidades que efectúen evaluaciones o análisis de la información que posean. (...)



Esta es una copia auténtica imprimible de un documento electrónico archivado en el Ministerio de Desarrollo Agrario y Riego, aplicando lo dispuesto por el Art. 25 de D.S. 070-2013-PCM y la Tercera Disposición Complementaria Final el D.S. 026-2016-PCM. Su autenticidad e integridad pueden ser contrastadas a través de la siguiente dirección web: <https://sisgedconsultaxexterna.midagri.gob.pe/> ingresando el código KLMN8ElFB1 y el número de documento.



Jirón Cahuide 805
Jesús María – Lima, Perú
T: (511) 209-8600
[https://www.gob.pe/midagri](http://www.gob.pe/midagri)

ANEXO 03: Datos de la serie de tiempo empleada

Años	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2000	31400	5700	0	0	0	0	0	0	0	0	4400	12684
2001	28276	5058	1198	196	0	0	0	0	0	0	3296	37935
2002	33720	7587	0	0	0	0	0	0	0	0	6000	24300
2003	33962	18000	0	0	0	0	0	0	0	0	8500	32590
2004	52144	26072	6482	0	0	0	0	0	0	0	3593	68407
2005	72000	32000	0	0	0	0	0	0	40	2600	6580	3100
2006	23600	65000	33300	0	0	0	0	0	1200	5480	12630	62693
2007	50112	3800	790	0	0	0	0	0	0	5380	40635	83620
2008	99246	42800	5600	0	0	0	0	0	0	4200	14048	20000
2009	16800	0	0	0	0	0	0	0	0	5800	8750	24309
2010	77000	70000	40000	0	0	0	0	0	1300	3180	7000	88000
2011	102000	82000	7000	0	0	0	0	0	0	0	800	3500
2012	38000	6200	4500	600	100	0	0	55	560	1450	350	420
2013	125000	67500	27000	452	0	0	0	0	1750	1706	620	1400
2014	138843	29068	10900	0	0	0	0	270	600	800	2873	16800
2015	23560	80000	0	0	0	0	0	200	400	1430	8376	83716
2016	104664	10477	0	0	0	0	0	0	280	1450	7700	90970
2017	115000	10810	0	0	0	0	0	600	2000	2000	6700	93250
2018	119000	11000	0	0	0	0	0	0	0	480	1900	81900
2019	127800	10920	0	0	0	0	0	9225	11620	18200	33589	147500
2020	104438	4600	0	0	0	0	0	0	8100	15400	22500	148800
2021	111911	4725	0	0	0	0	308	7000	7280	11600	10000	135400
2022	137700	6552	0	0	0	0	0	3900	6500	11700	16800	131600
2023	143780	14200	700	0	0	0	0	84	100	1200	3315	6159
2024	22200	6200	4500	600	50	0	0	180				

ANEXO 05: Código R utilizado

Se presenta el código principal empleado.

```
##### Librerías necesarias #####
# Importación y manejo de datos
library(readxl)
library(dplyr)
library(lubridate)
library(tsibble)
library(tidyr)
library(scales)
library(tibble)

# Análisis exploratorio y visualización
library(ggplot2)
library(forecast)
library(e1071)
library(magrittr)
library(ggfortify)

# Modelado y ajuste de modelos SARIMA
library(fable)
library(forecast)
library(fabletools)
library(kableExtra)
library(Metrics)

# Modelado LSTM
library(tensorflow)
library(keras3)

# Carga las funciones personalizadas para apariencia de gráficos
# tablas, y configuraciones de secuencias LSTM.
source("https://raw.githubusercontent.com/jaironkevin/SARIMA_LSTM/main/0E1_0E2.R")
source("https://raw.githubusercontent.com/jaironkevin/SARIMA_LSTM/main/0E3.R")

# IMPORTACIÓN DE DATOS
AGRO <- readxl::read_excel("D:/A_TESIS_ESTADISTICA/CAP_RESULTADOS/DATASET/AGRO.xlsx")
##### O.E.1: DESCRIPCIÓN DE LA SERIE #####
##### Defino las medidas a calcular
`%>%` <- magrittr::`%>%
Medidas <- c("obs.", "Mínimo", "1st qu.", "Mediana", "Media", "3rd Qu.",
           "Máximo", "SD.", "Asimetria", "Curtosis", "Moda")
```

```

# Calculo de las estadísticas en AGRO$Producción
Z <- vector(length = length(Medidas))
Z[1] <- round(sum(!is.na(AGRO$Producción)), 0)
Z[2] <- round(min(AGRO$Producción, na.rm = TRUE), 0)
Z[3] <- round(quantile(AGRO$Producción, 0.25, na.rm = TRUE), 0)
Z[4] <- round(median(AGRO$Producción, na.rm = TRUE), 0)
Z[5] <- round(mean(AGRO$Producción, na.rm = TRUE), 0)
Z[6] <- round(quantile(AGRO$Producción, 0.75, na.rm = TRUE), 0)
Z[7] <- round(max(AGRO$Producción, na.rm = TRUE), 0)
Z[8] <- round(sd(AGRO$Producción, na.rm = TRUE), 3)
Z[9] <- round(e1071::skewness(AGRO$Producción, na.rm = TRUE), 3)
Z[10] <- round(e1071::kurtosis(AGRO$Producción, na.rm = TRUE), 3)
Z[11] <- round(Moda(AGRO$Producción), 3)

# Crear el data frame
Z <- data.frame("Estadística" = Medidas, Valor = Z)

# Grafica de la serie
Mango <- ts(AGRO$Producción,
            start = c(2000, 1),
            end = c(2024, 8),
            frequency = 12)

autoplot(Mango, colour = "#AE123A")+
  xlab("Tiempo") +
  ylab("Producción (toneladas)")

# Veo la estacionalidad
ggseasonplot(Mango) +
  labs(
    title = NULL,
    x = NULL,
    y = "Producción",
    caption = "Meses"
  ) +
  theme(
    plot.title = element_text(size = 18, hjust = 0.5, face = "bold"),
    axis.title = element_text(size = 14, face = "bold"),
    axis.text = element_text(size = 12),
    legend.title = element_text(size = 12),
    legend.text = element_text(size = 10),
    plot.caption = element_text(size = 12, hjust = 0.5)

```

```

) +
scale_fill_brewer(palette = "Dark2") +
geom_point(data = NULL)

# veo la distribución mensual
fechas <- seq(as.Date("2000-01-01"), as.Date("2024-08-01"), by = "month")

Box_mensual <- data.frame(
  Fecha = fechas,
  Produccion = AGRO$Producción
)
df_mango<-Box_mensual
Box_mensual$Mes <- factor(month.name[month(Box_mensual$Fecha)],
                            levels = month.name)

ggplot(Box_mensual, aes(x = Mes, y = Produccion, fill = Mes)) +
  geom_boxplot(alpha = 0.7) +
  geom_jitter(position = position_jitter(0.2), color = "black", alpha = 0.5) +
  labs(x = "Mes",
       y = "Producción (toneladas)") +
  scale_fill_brewer(palette = "Set3") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# reviso los componentes que tiene la serie
componetes <- stl(Mango, s.window = "periodic")
autoplot(componetes, facets = TRUE, ts.colour = "#CD5555")

#####
##### Estacionariedad #####
#####

adf_table(Mango)
kpss_table(Mango)

# diferenciación D=1
Diff_Mango <- diff(Mango, lag = 12)

p1 <- autoplot(Diff_Mango, colour = "#26456E") +
  theme(plot.title = element_text(hjust = 0.5, size = 10))

p1
#####
##### O.E.2: SARIMA-BOX-JENKINS #####
#####

```

```

#####
# Identificación SARIMA
D1<-afc.id(Diff_Mango, lag.max = 25, period = 12, threshold = 0.05)
D2<-pacf.id(Diff_Mango, lag.max = 25, period = 12, threshold = 0.05)

# Estimación SARIMA
models_sarima <- AGRO %>%
  mutate(Date = yearmonth(Date)) %>%
  as_tsibble(index = Date)

models_sarima[,-1] <- apply(models_sarima[,-1], 2, as.numeric)

MODELS_SARIMA_G1 <- models_sarima %>%
  model(
    G1_MOD1 = ARIMA(Producción ~ pdq(1,0,1) + PDQ(2,1,0)),
    G1_MOD2 = ARIMA(Producción ~ pdq(1,0,1) + PDQ(2,1,1)),
    G1_MOD3 = ARIMA(Producción ~ pdq(1,0,0) + PDQ(2,1,0)),
    G1_MOD4 = ARIMA(Producción ~ pdq(0,0,1) + PDQ(2,1,0)),
    G1_MOD5 = ARIMA(Producción ~ pdq(1,0,0) + PDQ(0,1,1)),
    G1_MOD6 = ARIMA(Producción ~ pdq(0,0,1) + PDQ(2,1,1)),
    G1_MOD7 = ARIMA(Producción ~ pdq(0,0,1) + PDQ(0,1,1))
  ) %>%
  tidy() %>%
  select(.model, term, estimate, std.error, statistic, p.value) %>%
  mutate(
    estimate = sprintf("%.3f", estimate),
    std.error = sprintf("%.3f", std.error),
    statistic = sprintf("%.3f", statistic),
    p.value = sprintf("%.3f", p.value),
    .model = recode(.model,
      G1_MOD1 = "SARIMA(1,0,1)(2,1,0)[12]",
      G1_MOD2 = "SARIMA(1,0,1)(2,1,1)[12]",
      G1_MOD3 = "SARIMA(1,0,0)(2,1,0)[12]",
      G1_MOD4 = "SARIMA(0,0,1)(2,1,0)[12]",
      G1_MOD5 = "SARIMA(1,0,0)(0,1,1)[12]",
      G1_MOD6 = "SARIMA(0,0,1)(2,1,1)[12]",
      G1_MOD7 = "SARIMA(0,0,1)(0,1,1)[12]")
  )
  )

# Diagnóstico SARIMA

MOD3<-Arima(Mango, order=c(1,0,0), seasonal=c(2,1,0),method = "ML")

```

```

check_residuals(MOD3, lag.max = 25, title = "SARIMA (1,0,0)(2,1,0)[12]")

MOD3_test <- test.residuals(MOD3,
                             test_ac = "Ljung-Box",
                             test_normal = "Jarque-Bera",
                             test_hetero = "Levene",
                             test_stationarity = "KPSS",
                             test_media = TRUE)

MOD4<-Arima(Mango, order=c(0,0,1), seasonal=c(2,1,0), method = "ML")
check_residuals(MOD4, lag.max = 25, title = "SARIMA(0,0,1)(2,1,0)[12]")

MOD4_test <- test.residuals(MOD4,
                             test_ac = "Ljung-Box",
                             test_normal = "Jarque-Bera",
                             test_hetero = "Levene",
                             test_stationarity = "KPSS",
                             test_media = TRUE)

MOD5<-Arima(Mango, order=c(1,0,0), seasonal=c(0,1,1), method = "ML")
check_residuals(MOD5, lag.max = 25, title = "SARIMA(1,0,0)(0,1,1)[12]")



MOD5_test <- test.residuals(MOD5,
                             test_ac = "Ljung-Box",
                             test_normal = "Jarque-Bera",
                             test_hetero = "Levene",
                             test_stationarity = "KPSS",
                             test_media = TRUE)

MOD7 <- Arima(Mango, order = c(0,0,1), seasonal = list(order = c(0,1,1),
                                                       period = 12),
               method = "ML")
check_residuals(MOD7, lag.max = 25, title = "SARIMA(0,0,1)(0,1,1)[12]")

MOD7_test <- test.residuals(MOD7,
                             test_ac = "Ljung-Box",
                             test_normal = "Jarque-Bera",
                             test_hetero = "Levene",
                             test_stationarity = "KPSS",
                             test_media = TRUE)

# Comparación entre modelos los 4 SARIMA

```

```

# Preparo el conjunto de datos
models_sarima <- AGRO %>%
  mutate(Date = yearmonth(Date)) %>%
  as_tsibble(index = Date)

models_sarima[,-1] <- apply(models_sarima[,-1], 2, as.numeric)

# Definir los modelos SARIMA usando los órdenes ARIMA como nombres de modelo
models_sarima_fable <- models_sarima %>%
  model(
    `ARIMA(1,0,0)(2,1,0)` = ARIMA(Producción ~ pdq(1,0,0) + PDQ(2,1,0)),
    `ARIMA(0,0,1)(2,1,0)` = ARIMA(Producción ~ pdq(0,0,1) + PDQ(2,1,0)),
    `ARIMA(1,0,0)(0,1,1)` = ARIMA(Producción ~ pdq(1,0,0) + PDQ(0,1,1)),
    `ARIMA(0,0,1)(0,1,1)` = ARIMA(Producción ~ pdq(0,0,1) + PDQ(0,1,1))
  )

# Obtengo los ajustes
ajuste <- fitted(models_sarima_fable)
ajuste_limpio <- ajuste %>%
  filter(!is.na(.fitted))

# Gráfico: Todos los modelos en un solo gráfico
plot_all <- ggplot(models_sarima, aes(x = Date, y = Producción)) +
  geom_line(aes(color = "Observado"), linewidth = 0.3) +
  geom_line(data = ajuste_limpio, aes(y = .fitted, color = .model),
            linewidth = 0.5) +
  labs(color = "Modelos") +
  theme_minimal() +
  scale_color_manual(values = c("Observado" = "black",
                                "ARIMA(1,0,0)(2,1,0)" = "#1f77b4",
                                "ARIMA(0,0,1)(2,1,0)" = "#ff7f0e",
                                "ARIMA(1,0,0)(0,1,1)" = "#2ca02c",
                                "ARIMA(0,0,1)(0,1,1)" = "#d62728"))

plot_all
##### Métricas entre modelos SARIMA #####
# Extraer los valores observados y predichos
augment_data <- models_sarima_fable %>%
  augment() %>%
  select(.model, Producción, .fitted)

```

```

custom_metrics <- augment_data %>%
  as.data.frame() %>%
  group_by(.model) %>%
  summarise(
    MDAE = mdae(Producción, .fitted),
    SMAPE = smape(Producción, .fitted)
  )

custom_metrics

# Calcular las métricas estándar de fable
accuracy_df <- models_sarima_fable %>%
  forecast::accuracy()

# Combinar las métricas estándar con las de metrics
full_metrics_df <- accuracy_df %>%
  left_join(custom_metrics, by = ".model")

full_metrics_df
#####
##### O.E.3: Modelado LSTM #####
#####

# Nota: Para no extender mas el código, se presenta el código del cual
# se fueron modificando los hiperparametros y parámetros de los modelos
# A, B, C, D.

# Vuelvo a Carga los datos disponibles (Fuente: MIDAGRI)
file_path <- "D:/A_TESIS_ESTADISTICA/CAP_RESULTADOS/DATASET/AGRO.xlsx"
df <- read_excel(file_path)
df$date <- as.Date(paste0(df$date, "-01"), format = "%Y-%m-%d")
df <- df[order(df$date), ]
str(df)

#####
##### 1.Pre-procesamiento #####
#####

# 1.1. Partición del set en entrenamiento, validación y prueba
splits <- train_val_test_split(df$Producción, tr_size = 0.7,
                                vl_size = 0.15, ts_size = 0.15)
tr <- splits$train
vl <- splits$val
ts <- splits$test

```

```

cat(sprintf('Tamaño set de entrenamiento: %d\n', length(tr)))
cat(sprintf('Tamaño set de validación: %d\n', length(vl)))
cat(sprintf('Tamaño set de prueba: %d\n', length(ts)))

# 1.2. Generación del dataset supervisado (entrada y salida del modelo)

# Definir hiperparámetros
INPUT_LENGTH <- 24 # Número de pasos temporales usados como entrada
OUTPUT_LENGTH <- 17 # Número de pasos que queremos predecir o de salida

datasets <- crear_todos_los_datasets(tr, vl, ts, INPUT_LENGTH, OUTPUT_LENGTH)

x_tr <- datasets$x_tr
y_tr <- datasets$y_tr
x_vl <- datasets$x_vl
y_vl <- datasets$y_vl
x_ts <- datasets$x_ts
y_ts <- datasets$y_ts

# Preparar la lista de datos de entrada
data_in <- list(
  'x_tr' = x_tr, 'y_tr' = y_tr,
  'x_vl' = x_vl, 'y_vl' = y_vl,
  'x_ts' = x_ts, 'y_ts' = y_ts
)

# 1.3. Escalamiento
resultado <- escalar_dataset(data_in)
data_s <- resultado$data_scaled
scalers <- resultado$scalers

# Extraer los datasets escalados
x_tr_s <- data_s$x_tr_s
y_tr_s <- data_s$y_tr_s
x_vl_s <- data_s$x_vl_s
y_vl_s <- data_s$y_vl_s
x_ts_s <- data_s$x_ts_s
y_ts_s <- data_s$y_ts_s

# Ajustar las formas de las salidas (Y) para que coincidan con
# la salida del modelo
y_tr_s <- array(y_tr_s, dim = c(dim(y_tr_s)[1], dim(y_tr_s)[2]))

```

```

y_vl_s <- array(y_vl_s, dim = c(dim(y_vl_s)[1], dim(y_vl_s)[2]))
y_ts_s <- array(y_ts_s, dim = c(dim(y_ts_s)[1], dim(y_ts_s)[2]))

##### 2. Creación y entrenamiento del modelo #####
# Establecer una semilla para reproducibilidad
set.seed(123)
tensorflow::set_random_seed(123)

# 2.1. Definir hiperparámetros
INPUT_LENGTH <- 24 # Número de pasos temporales usados como entrada 12
OUTPUT_LENGTH <- 17 # Número de pasos que queremos predecir
N_UNITS_FIRST_LSTM <- 256 #128
N_UNITS_SECOND_LSTM <- 128 # 64
DROPOUT_RATE <- 0.2
EPOCHS <- 50
BATCH_SIZE <- 32

# 2.2. Modelado
# Definir la forma de la entrada
INPUT_SHAPE <- c(INPUT_LENGTH, dim(x_tr_s)[3])

modelo <- keras_model_sequential() %>%
  layer_lstm(units = N_UNITS_FIRST_LSTM,
             return_sequences = TRUE,
             input_shape = INPUT_SHAPE) %>%
  layer_dropout(rate = DROPOUT_RATE) %>%
  layer_lstm(units = N_UNITS_SECOND_LSTM) %>%
  layer_dense(units = OUTPUT_LENGTH, activation = 'linear')

# Compilar el modelo
modelo %>% compile(
  optimizer = optimizer_adam(learning_rate = 1e-3),
  loss = "mae"
)

# revisamos el esquema del modelo
summary(modelo)
plot(modelo)
plot(
  modelo,
  show_shapes = TRUE,

```

```

show_dtype = TRUE,
show_layer_names = TRUE,
rankdir = "TB",
expand_nested = TRUE,
dpi = 100,
show_layer_activations = TRUE,
)

# Definir EarlyStopping
early_stop <- callback_early_stopping(
  monitor = "val_loss",
  patience = 10,
  restore_best_weights = TRUE
)

# Entrenar el modelo
historiaA <- modelo %>% fit(
  x = x_tr_s,
  y = y_tr_s,
  batch_size = BATCH_SIZE,
  epochs = EPOCHS,
  validation_data = list(x_vl_s, y_vl_s),
  verbose = 2,
  callbacks = list(early_stop)
)

# ploteamos la historia del entrenamiento y validación
plotC<-plot_history(historiaA) +
  abs(title = "Modelo A: inputs=24, outputs=17, 2 capas (256, 128 units).") +
  theme(plot.title = element_text(hjust = 0.5, size = 10, face = "bold"))

# guardo y cargo el historial de entrenamiento
save(historia, file ="history1.RData")
load("history1.RData")
# Guardar el modelo completo en formato .keras
keras3::save_model(modelo,"LSTM_A.keras")

#####
##### 3. Desempeño del Modelo #####
#####

# Evaluar el modelo en los 3 conjuntos

mae_tr <- modelo %>% evaluate(x_tr_s, y_tr_s, verbose = 0)

```

```

mae_vl <- modelo %>% evaluate(x_vl_s, y_vl_s, verbose = 0)
mae_ts <- modelo %>% evaluate(x_ts_s, y_ts_s, verbose = 0)

cat('Comparativo desempeños:\n')
cat(sprintf(' MAE train:\t %.3f\n', mae_tr$loss))
cat(sprintf(' MAE val:\t %.3f\n', mae_vl$loss))
cat(sprintf(' MAE test:\t %.3f\n', mae_ts$loss))

#####
# 4. Evaluación del modelo sobre Test #####
# 4.1. Hacer predicciones en el conjunto de prueba
y_ts_pred_s <- modelo %>% predict(x_ts_s, verbose = 0)

# 4.2. Desescalar las predicciones y los valores observados
y_ts_pred <- desescalar(y_ts_pred_s, scalers$scaler_y)
y_ts_true <- desescalar(y_ts_s, scalers$scaler_y)

# 4.2. Graficos las secuencias que se generaran y pronósticos finales
# Debido al solapamiento de los outputs (many to many), se generan
# mas de un pronostico para cada punto de test observado.

data <- resultLSTM_testC(y_ts_pred, y_ts_true)
plot_resultLSTM_testC(data$test_completo, data$resultados)

data_completa <- resultLSTM_complete(y_ts_pred, y_ts_true)
plot_resultLSTM_complete(data_completa$full_data,
                         data_completa$test_completo,
                         data_completa$resultados)

# 4.2. se aplican medidas de precisión basadas en sus errores

# se consideran las mas robustas debido a la naturaleza de la serie
Test_obs<-tibble_pre_obs(y_ts_true, y_ts_pred, df)
metrics(Test_obs$Observado, Test_obs$Predicción)

#####
# 5. Pronostico SARIMA #####
#####

modelo_sarima <- forecast::Arima(Mango, order = c(1, 0, 0),
                                    seasonal = c(2, 1, 0))
predicciones <- forecast::forecast(modelo_sarima, h = 17,
                                    bootstrap = TRUE)

```

```

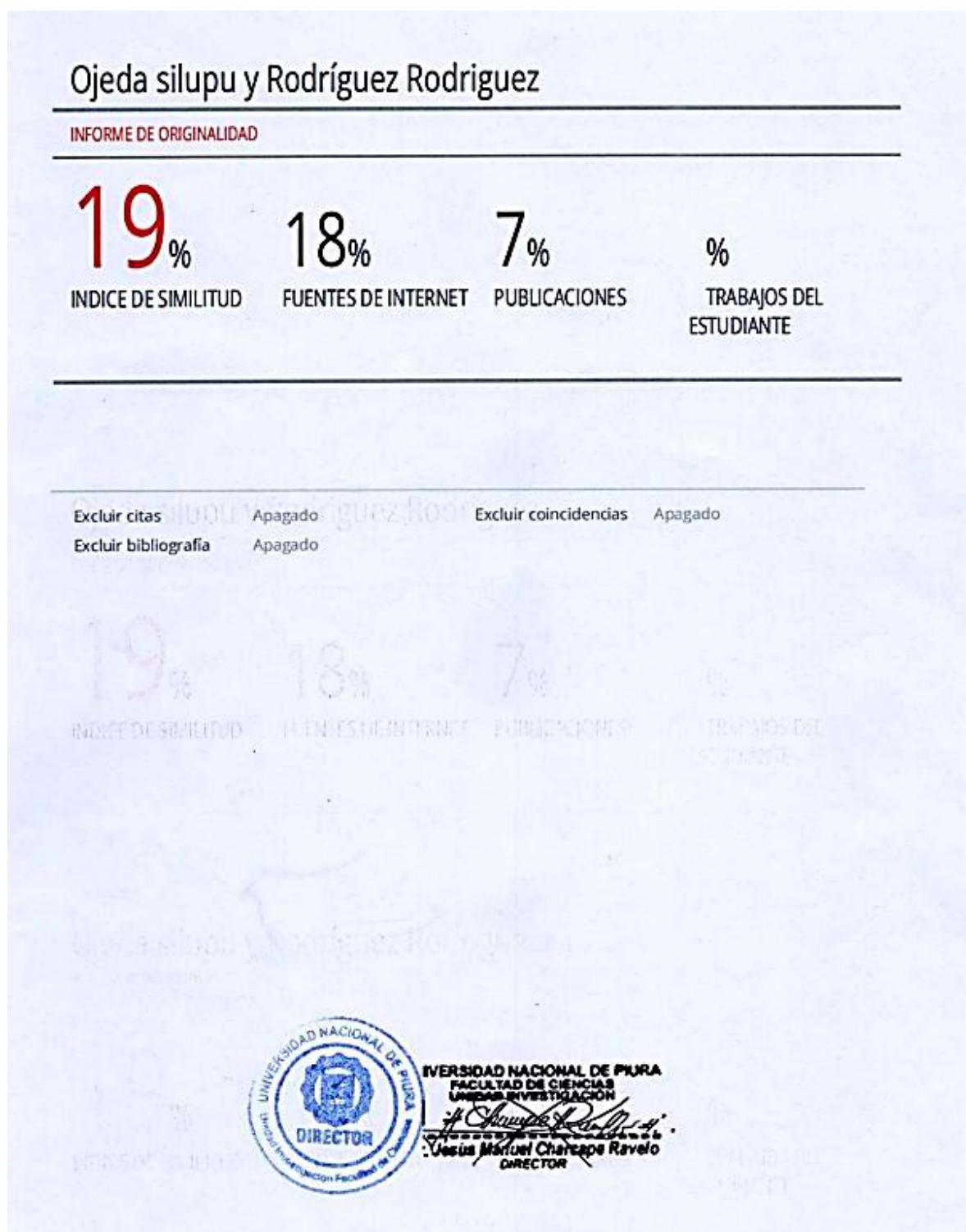
fechas_obs <- seq(as.Date("2000-01-01"), by = "month",
                   length.out = length(Mango))
df_obs <- tibble(Fecha = fechas_obs,
                  Producción = as.numeric(Mango))

df_pred <- as.data.frame(predicciones)
df_pred$Fecha <- seq(as.Date("2024-09-01"),
                      by = "month", length.out = 17)

ggplot() +
  geom_ribbon(data = df_pred, aes(x = Fecha, ymin = `Lo 95`,
                                   ymax = `Hi 95`, fill = "IC 95%"),
               alpha = 0.6) +
  geom_ribbon(data = df_pred, aes(x = Fecha,
                                   ymin = `Lo 80`, ymax = `Hi 80`,
                                   fill = "IC 80%"), alpha = 0.5) +
  geom_line(data = df_obs, aes(x = Fecha,
                                y = Producción, color = "Observado"),
            size = 0.2) +
  geom_line(data = df_pred, aes(x = Fecha, y = `Point Forecast`,
                                color = "Pronóstico"), size = 0.3) +
  scale_color_manual(
    values = c("Observado" = "#AE123A", "Pronóstico" = "#49525E"))
  ) +
  scale_fill_manual(
    values = c("IC 80%" = "#2C5985", "IC 95%" = "#7BBFC9"))
  ) +
  labs(
    x = NULL,
    y = "Producción"
  ) +
  theme(
    axis.title = element_text(face = "bold"),
    plot.title = element_text(face = "bold"),
    legend.title = element_blank(),
    legend.position = "top",
    legend.direction = "horizontal",
    legend.box = "horizontal",
    legend.spacing.x = unit(0.4, "cm"))

```

ANEXO 06: Informe de Turnitin

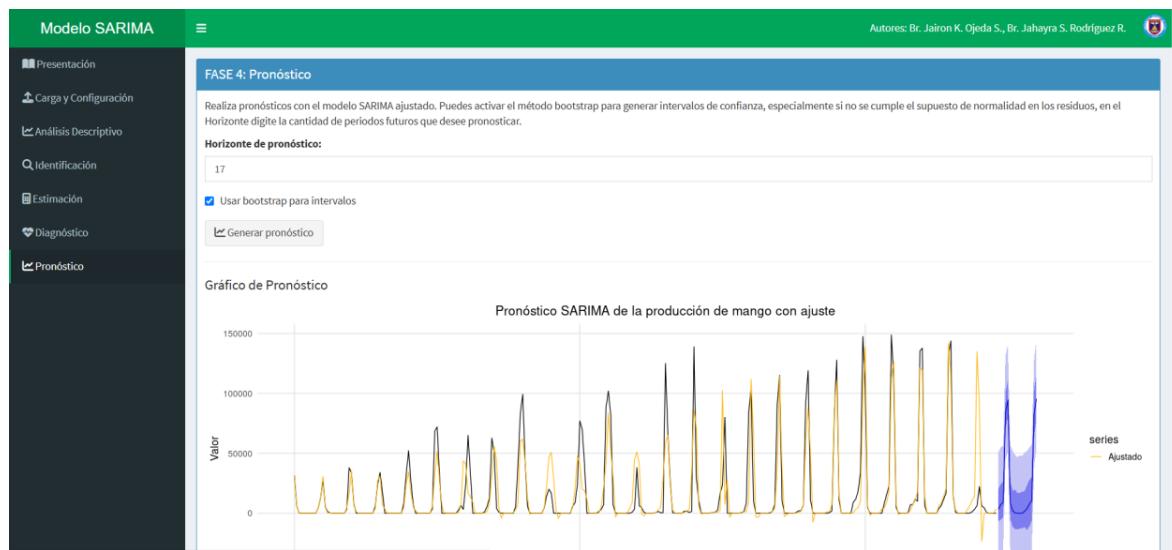


ANEXO 06: Seguimiento del modelo SARIMA

Debido a que se determinó que el modelo más eficiente para pronosticar la producción mensual de mango fue el SARIMA(1,0,0)(2,1,0)[12], se desarrolló una aplicación web utilizando R y Shiny con el propósito de automatizar y poner en producción dicho modelo. Esta herramienta permite al usuario interesado realizar un seguimiento continuo del modelo, así como modificar sus parámetros en caso de que la dinámica de la serie de tiempo así lo requiera.

Es importante destacar que la aplicación fue diseñada bajo el enfoque metodológico de Box-Jenkins, con el objetivo de garantizar un monitoreo más robusto y riguroso de los pronósticos. La interfaz está organizada en siete secciones:

- Una descripción general del objetivo del modelo,
- La carga y configuración de los datos,
- El análisis descriptivo de la serie,
- La identificación del modelo,
- La estimación de parámetros,
- El diagnóstico de residuos,
- Y, finalmente, la generación del pronóstico.



Puede hacer seguimiento al modelo SARIMA en la siguiente aplicación, echa con la library(shiny) en R en el siguiente enlace:

https://9t0pdl-jaironkevin.shinyapps.io/APP_SARIMA/