

CHIKUNGUNYA EN BUCARAMANGA: UN ANÁLISIS EN SERIES DE TIEMPO

LAURA CECILIA ZÚÑIGA GARCÍA

UNIVERSIDAD INDUSTRIAL DE SANTANDER

FACULTAD DE CIENCIAS

ESCUELA DE MATEMÁTICAS

BUCARAMANGA

2025

CHIKUNGUNYA EN BUCARAMANGA: UN ANÁLISIS EN SERIES DE TIEMPO

LAURA CECILIA ZÚÑIGA GARCÍA

Trabajo de grado para optar al título de
Matemática

Director

Andrés Sebastián Ríos Gutiérrez

Candidato a Doctor

UNIVERSIDAD INDUSTRIAL DE SANTANDER

FACULTAD DE CIENCIAS

ESCUELA DE MATEMÁTICAS

BUCARAMANGA

2025

DEDICATORIA

Dedico este trabajo a mi madre, **Emilia**, y a mi hermano, **Juan David**, quienes me han apoyado en cada decisión y locura. Y a mi compañera fiel, que también se graduará conmigo: mi hermosa **Kiki**.

*«Vivo mi vida en círculos crecientes, que encima de las cosas se dibujan. El último quizás no lo complete, **pero quiero intentarlo...**»*

— R.M.R.

AGRADECIMIENTOS

Agradezco profundamente a mis profesores, en especial a mi tutor, el profesor **Andrés**, quien me ha brindado una paciencia ejemplar, una virtud que valoro enormemente en él. A **Jesús**, por estar siempre a mi lado, por su compañía incondicional y por todas las lecciones que me ha regalado con tanto amor.

No puedo olvidar a mis amigos que me hicieron volver a creer en lo que estudio y me mostraron que no siempre es bueno estar solo; ellos hicieron este camino mucho más ameno: **Catalina, Émerzon, Luis, Jimmy y Nicolás**.

CONTENIDO

	pág.
INTRODUCCIÓN	11
1 Planteamiento y justificación del problema	14
2 Objetivos	16
3 Antecedentes	17
4 Modelización ARIMA	19
4.1 Procesos estocásticos y series temporales	19
4.2 Procesos estocásticos estacionarios	21
4.3 Función de autocorrelación	25
4.4 Modelos lineales estacionarios	27
4.4.1 Modelo lineal general	30
4.4.2 Procesos autorregresivos (AR)	35
4.4.3 Procesos de medias móviles (MA)	37
4.4.4 Procesos ARMA(p,q)	39
4.5 Modelos ARIMA(p,d,q)	40
4.5.1 Selección de los órdenes p, q, d en un modelo ARIMA(p,d,q)	43
4.5.2 Estrategia de modelización ARIMA	43
4.6 Modelos ARIMA estacionales	44
4.6.1 Modelos ARIMA estacionales multiplicativos	46
4.7 Modelos de regresión dinámica	49
4.7.1 Modelo autorregresivo y de retardos distribuidos	49
4.7.2 Modelo dinámico con variables exógenas	50

5	Inferencia y predicción óptima con modelos ARIMA(p,d,q)	53
5.1	Estimación	53
5.2	Validación del modelo	59
5.2.1	Análisis de coeficientes estimados	60
5.2.2	Análisis de residuos	61
5.3	Predicción con modelos estacionarios	66
5.3.1	Predicción con modelos ARMA(p,q)	70
5.4	Predicción con modelos no estacionarios	72
6	Estudio de caso: Chikungunya en Bucaramanga	76
6.1	Metodología ARIMA	76
6.2	Descripción de las bases de datos	79
6.2.1	Chikungunya	79
6.2.2	Variables climáticas	82
6.2.2.1	Metodología de los datos	82
6.3	Análisis de la dinámica del Chikungunya en Bucaramanga	83
6.3.0.1	Temperatura	91
6.3.0.2	Precipitación	93
6.3.0.3	Humedad relativa	94
6.4	Validación del modelo	99
7	Conclusiones	103
	Conclusiones	103
	BIBLIOGRAFÍA	105

LISTA DE FIGURAS

	pág.
Figura 1 Serie simulada ARIMA(1,1,1)	59
Figura 2 Residuos del modelo AR(2)	64
Figura 3 FAC de los residuos AR(2)	65
Figura 4 Proceso de modelado ARIMA y validación de supuestos para la proyección de la incidencia del virus del chikungunya.	78
Figura 5 Casos diarios de chikungunya	84
Figura 6 Casos mensuales de chikungunya	85
Figura 7 Periodograma.	87
Figura 8 Serie temporal de casos de chikungunya con período de 10 meses	88
Figura 9 Serie temporal de casos de chikungunya con período de 14 meses	88
Figura 10 Temperatura promedio en la región de estudio	91
Figura 11 Descomposición de la serie temporal de temperatura	92
Figura 12 Precipitación promedio en la región de estudio	93
Figura 13 Descomposición de la serie temporal de precipitación	93
Figura 14 Humedad relativa promedio en la región de estudio	94
Figura 15 Descomposición de la serie temporal de humedad relativa	95
Figura 16 Gráfica del mejor modelo con mejor criterio AIC	101

LISTA DE TABLAS

	pág.
Tabla 1 Estadísticos principales para AR(2).	66
Tabla 2 Valores de frecuencia y espectro.	87
Tabla 3 Datos climáticos e incidencia del Chikungunya en Bucaramanga	91
Tabla 4 Mejores modelos ARIMA	98
Tabla 5 Coeficientes del modelo ARIMA con variables exógenas.	98
Tabla 6 Indicadores de ajuste del modelo.	99

RESUMEN

TÍTULO: CHIKUNGUNYA EN BUCARAMANGA: UN ANÁLISIS EN SERIES DE TIEMPO. *

AUTORES: LAURA CECILIA ZÚÑIGA GARCÍA **

PALABRAS CLAVE: CHIKUNGUNYA, SERIES DE TIEMPO, ARIMA, VARIABLES CLIMÁTICAS, INCIDENCIA, VIGILANCIA EPIDEMIOLÓGICA.

DESCRIPCIÓN:

El chikungunya (CHIKV) es un virus emergente que se propaga a través de los mosquitos *Aedes aegypti* y *Aedes albopictus*. Esta enfermedad se caracteriza por síntomas como fiebre alta y dolor en las articulaciones, los cuales pueden prolongarse durante meses, representando una gran carga para los sistemas de salud. Su identificación se complica debido a su similitud con el dengue y el Zika, ya que estas enfermedades comparten el mismo vector y presentan síntomas clínicos parecidos. A pesar de que en Colombia se han reportado pocos casos recientemente, las condiciones climáticas del país son propicias para el vector en gran parte de su área, lo que lo expone al riesgo de nuevos brotes. Varios estudios han señalado que las condiciones meteorológicas juegan un papel importante en la transmisión del virus. En este marco, el proyecto propone modelar los casos probables de CHIKV en Colombia durante 2014–2018, usando análisis de series de tiempo y variables climáticas de fuentes oficiales, con el objetivo de comprender mejor su comportamiento y contribuir a estrategias de vigilancia y alerta temprana basadas en evidencia.

* Trabajo de grado

** Facultad de Ciencias. Escuela de Matemáticas. Director: Andrés Sebastián Ríos Gutiérrez, Candidato a Doctor.

ABSTRACT

TITLE: CHIKUNGUNYA IN BUCARAMANGA: A TIME-SERIES ANALYSIS *

AUTORS: LAURA CECILIA ZÚÑIGA GARCÍA **

KEYWORDS: CHIKUNGUNYA, TIME-SERIES, ARIMA, CLIMATIC VARIABLES, INCIDENCE, EPIDEMIOLOGICAL SURVEILLANCE.

DESCRIPTION:

Chikungunya (CHIKV) is an emerging virus transmitted by *Aedes aegypti* and *Aedes albopictus* mosquitoes. This disease is characterized by symptoms such as high fever and joint pain, which can persist for months, representing a significant burden on healthcare systems. Its identification is complicated due to its similarity to dengue and Zika, as these diseases share the same vector and exhibit similar clinical symptoms. Although few cases have been reported recently in Colombia, the country's climatic conditions are favorable for the vector across much of its territory, exposing it to the risk of new outbreaks. Several studies have highlighted that meteorological conditions play an important role in the transmission of the virus. In this context, the project aims to model probable CHIKV cases in Colombia during 2014–2018 using time series analysis and climatic variables from official sources, with the goal of better understanding its behavior and contributing to evidence-based surveillance and early warning strategies.

* Bachelor Thesis

** Faculty of Science. School of Mathematics. Advisor: Andrés Sebastián Ríos Gutiérrez, Ph.D. Candidate.

INTRODUCCIÓN

El virus del chikungunya (CHIKV) es una enfermedad emergente transmitida de manera similar al dengue (DENV) y al Zika (ZIKV) a través de la picadura de hembras de los mosquitos *Aedes aegypti* y *Aedes albopictus*¹. Este virus ARN pertenece al género *alphavirus*, familia *Togorividae* y representa una significativa carga de morbilidad. Desde el punto de vista clínico, se caracteriza por fiebre superior a $38,5^{\circ}C$ y dolor articular incapacitante, el cual puede persistir hasta 32 meses después de la infección, lo que produce un impacto considerable en el sistema de salud y otros servicios de apoyo².

La Organización Mundial de la Salud ha reportado 113447 casos de chikungunya y 51 fallecimientos en las Américas entre el 1 de enero y el 4 de marzo de 2023, lo que representa un aumento significativo en comparación con el mismo período del año anterior, con una cifra cuatro veces mayor de casos y fallecimientos. En Colombia, en 2023 se registraron muy pocos casos de chikungunya, manteniéndose por debajo de los valores históricos desde la aparición del virus en 2014³. No obstante, la presencia de virus del dengue, chikungunya y Zika, todos transmitidos por el mismo vector y con manifestaciones clínicas similares, puede dificultar el diagnóstico preciso y afectar la

¹ M. A. Carrillo et al.: *Risk of dengue, Zika, and chikungunya transmission in the metropolitan area of Cucuta, Colombia: cross-sectional analysis, baseline for a cluster-randomised controlled trial of a novel vector tool for water containers*. En: *BMC Public Health* 23.1 (2023). DOI: <https://doi.org/10.1186/s12889-023-15893-4>.

² J. C. Rueda et al.: *Demographic and clinical characteristics of chikungunya patients from six Colombian cities, 2014–2015*. En: *Emerging Microbes and Infections* 8.1 (2019), págs. 1490-1500. DOI: <https://doi.org/10.1080/22221751.2019.1678366>.

³ Organización Mundial de la Salud: *Partes sobre brotes epidémicos; Expansión geográfica de los casos de dengue y chikungunya más allá de las áreas históricas de transmisión en la Región de las Américas*. 2023. <https://www.who.int/es/emergencies/disease-outbreak-news/item/2023-DON448>.

eficacia del sistema de vigilancia epidemiológica⁴.

Por otro lado, diversos estudios en Asia y Latinoamérica recalcan la importancia de los factores climáticos como determinantes del complejo ciclo epidemiológico de transmisión del chikungunya. Las condiciones climáticas, sujetas a variaciones regionales, tienen un impacto directo en el desarrollo de los mosquitos y la propagación del virus, siendo favorables alrededor de los $29^{\circ}C$ y en altitudes hasta 2300 metros sobre el nivel del mar⁵. En el país, aproximadamente el 81% de las 73 ciudades principales con más de 100000 habitantes experimentan estas condiciones, lo cual genera un entorno favorable para la aparición de patologías transmitidas por vectores. Por tanto, el objetivo de este proyecto es modelar el número de casos probables de CHIKV teniendo en cuenta variables meteorológicas usando análisis de series de tiempo, con el fin de comprender el comportamiento del virus en el país durante los años 2014-2018.

La metodología aplicada se fundamenta en datos recopilados por el Instituto Nacional de Salud de forma anónima a través de una base de datos abierta y disponible en la página del Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA), por lo que no se requiere la autorización del comité de ética para el desarrollo de esta investigación. Se utilizarán estos datos para ajustar un modelo ARIMA de regresión dinámica que represente apropiadamente el comportamiento, teniendo en cuenta factores climatológicos proporcionados por NASA Langley Research Center (LaRC) POWER

⁴ Alejandro Rico-Mendoza et al.: *Co-circulation of dengue, chikungunya, and Zika viruses in Colombia from 2008 to 2018*. En: *Revista Panamericana de Salud Pública* 43 (2019), e49. DOI: 10.26633/RPSP.2019.49. <https://iris.paho.org/handle/10665.2/51077>.

⁵ J. Morgan, C. Strode y J. E. Salcedo-Sora: *Climatic and socio-economic factors supporting the co-circulation of dengue, Zika and chikungunya in three different ecosystems in Colombia*. En: *PLOS Neglected Tropical Diseases* 15.3 (2021). DOI: <https://doi.org/10.1371/journal.pntd.0009259>.

Project⁶.

La combinación de métodos matemáticos y estadísticos con estrategias de vigilancia epidemiológica tradicionales permite una comprensión más profunda y cuantitativa de la dinámica de la enfermedad, lo que ayuda a reducir errores y proporciona una alerta temprana para futuros brotes. En consecuencia, la importancia de este trabajo se sustenta en la necesidad de integrar las matemáticas con las políticas de salud pública, reforzando la colaboración entre el sector salud y las ciencias exactas para ofrecer respuestas rápidas y eficaces fundamentadas en la evidencia.

⁶ NASA Prediction of Worldwide Energy Resources (POWER): *Nasa POWER Data Access Viewer*. <https://power.larc.nasa.gov/>.

1. Planteamiento y justificación del problema

Dentro del grupo de enfermedades tropicales, el virus del chikungunya (CHIKV), aunque no se caracteriza por su alta letalidad, representa una preocupación debido a su elevada capacidad de propagación y a la persistencia de síntomas incapacitantes que pueden prolongarse durante meses tras la infección. En la actualidad son pocos los tratamientos efectivos que pueden aliviar el cuadro clínico, y el escaso interés comercial por parte de la industria farmacéutica ha limitado el desarrollo de terapias específicas debido a altos costos de desarrollo y mercados acotados en países de bajos ingresos⁷.

En países tropicales como Colombia, las condiciones climáticas, sumadas a factores estructurales como el subdesarrollo económico, crean un entorno propicio para su propagación. En particular, los cambios en los patrones de precipitación durante la última década, asociados a fenómenos globales como El Niño-Oscilación del Sur (ENOS), han intensificado el riesgo de brotes epidémicos.

En Bucaramanga, ciudad ubicada en el nororiente colombiano, con una población aproximada de 623 000 habitantes (DANE 2025), con una temperatura promedio anual de 24°C y dos períodos de lluvias marcadas. Estas circunstancias favorecen la presencia del mosquito *Aedes aegypti*, que es el principal transmisor del CHIKV. De acuerdo con informes del Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA), Bucaramanga y su área metropolitana han experimentado varios brotes de arbovirus en la última década, incluyendo dengue, chikungunya y Zika, con tasas de incidencia diversas y

⁷ Bryan Walsh: *1 Million People Have a Disease You've Never Heard Of*. Accessed: 2025-07-02. 2014. <https://time.com/3631711/chikungunya-epidemic-2014/>; Manu Goyal et al.: *Recent development in the strategies projected for chikungunya vaccine in humans*. En: *Drug Design, Development and Therapy* 12 (2018). Accessed: 2025-07-02, págs. 4195-4206. DOI: 10.2147/DDDT.S181574. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6292406/>.

una carga significativa para los sistemas de salud locales. No obstante, a pesar de la evidencia de una transmisión continua, no existen estimaciones sistemáticas ni modelos de predicción que incluyan factores climáticos para prever el comportamiento de la enfermedad en la región.

Esta carencia de investigaciones a nivel local constituyen una barrera para elaborar estrategias de control que se ajusten a la dinámica epistemológica de la ciudad. Por ejemplo, no hay claridad de cómo las variaciones climáticas locales impactan la frecuencia del virus en la ciudad, ni de la posibilidad de realizar predicciones de casos que guíen la distribución de recursos en salud pública con el fin de comprender la relación entre el clima y la propagación de enfermedades como el chikungunya es fundamental para fortalecer la respuesta en salud pública. Se ha evidenciado que las fluctuaciones en la temperatura ambiental pueden modificar la capacidad vectorial⁸. En este marco, los modelos ARIMA se consolidan como herramientas estadísticas robustas para cuantificar asociaciones no lineales entre predictores climáticos e incidencia de casos, identificar patrones estacionales y generar pronósticos de corto a mediano plazo con intervalos de confianza comprobables. Su aplicación en Bucaramanga, zona endémica⁹, permitiría optimizar la asignación de recursos mediante alertas tempranas basadas en evidencia. Al incorporar la autocorrelación temporal de la incidencia y el efecto de variables exógenas, esta metodología mejora la capacidad predictiva y la planificación estratégica en salud pública.

⁸ Erin A. Mordecai et al.: *Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models*. En: *PLOS Neglected Tropical Diseases* 11.10 (2017), e0006451. DOI: 10.1371/journal.pntd.0006451. <https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0006451>.

⁹ Instituto Nacional de Salud de Colombia: *Portal SIVIGILA - Sistema de Vigilancia en Salud Pública*. <https://portalsivigila.ins.gov.co/>.

2. Objetivos

Objetivo general

Modelar la dinámica de los casos confirmados por el virus del chikungunya (CHIKV) en Bucaramanga, Santander, Colombia, así como su variabilidad, al analizar la posible influencia de factores climáticos a través de un enfoque basado en una serie de tiempo con regresión dinámica.

Objetivos específicos

- Describir la evolución de la enfermedad en Bucaramanga, así como la relación contemporánea y dinámica entre sus componentes, a través de un modelo ARIMA con regresión dinámica.
- Proponer un modelo predictivo para estimar futuros casos en función de condiciones climáticas.
- Evaluar la influencia climática sobre la incidencia del CHIKV, validando su impacto en la precisión del modelo.

3. Antecedentes

Uno de los primeros trabajos orientados a modelar la transmisión del virus chikungunya fue desarrollado por Yakob et al., quienes analizaron datos de Brasil mediante técnicas de suavizamiento exponencial y modelos autorregresivos ARIMA, incorporando factores exógenos como la temperatura y la precipitación¹⁰. De forma similar, Karansinghe et al. aplicaron modelos ARIMA para anticipar casos de dengue en Colombo, Sri Lanka, con hasta cuatro semanas de anticipación, lo que facilitó la planificación hospitalaria y el despliegue de recursos¹¹.

En la región latinoamericana, Lowe et al. demostraron en Barbados que la inclusión de variables climáticas mejora significativamente la precisión de las predicciones¹², mientras que en Colombia, Medina et al. implementaron modelos SARIMAX para predecir la incidencia de dengue en Córdoba, resaltando el valor de integrar covariables exógenas en contextos de vigilancia limitada¹³. Por otro lado, Banu et al. en Asia identificaron una relación directa entre condiciones meteorológicas y la aparición de brotes de chikungunya, enfatizando la necesidad de considerar los factores contextuales para

¹⁰ Laith Yakob: *Modeling the transmission of chikungunya in Brazil using time series methods*. En: *PLOS Neglected Tropical Diseases* 16.5 (2022), e0010472. DOI: 10.1371/journal.pntd.0010472.

¹¹ Sanjaya Karasinghe, Ruwan Silva y Tharaka Perera: *Short-term forecasting of dengue outbreaks using ARIMA models: A case study in Colombo, Sri Lanka*. En: *Infectious Disease Modelling* 9 (2024), págs. 15-27. DOI: 10.1016/j.idm.2024.01.002.

¹² Rachel Lowe et al.: *Nonlinear and delayed impacts of climate on dengue risk in Barbados: A modelling study*. En: *PLOS Medicine* 15.7 (2018), e1002613. DOI: 10.1371/journal.pmed.1002613.

¹³ Juliana Medina, Andrés Pérez y Laura Gómez: *Predicción de casos de dengue en Córdoba, Colombia usando modelos SARIMAX con variables climáticas*. En: *Revista Colombiana de Salud Pública* 26.1 (2024). Artículo en prensa, págs. 45-58.

mejorar la capacidad de respuesta sanitaria¹⁴.

En conjunto, estos estudios evidencian que la combinación de series temporales con variables adicionales (principalmente climáticas) mejora la eficacia de los modelos para prever brotes. Este enfoque no solo permite representar la dinámica temporal del virus, sino también integrar elementos contextuales que afectan su diseminación. Como resultado, los modelos de predicción que emplean datos multidimensionales se convierten en herramientas clave para la vigilancia y la formulación de políticas en salud pública¹⁵.

¹⁴ Shamima Banu et al.: *Space-time clusters of dengue fever in Bangladesh*. En: *Tropical Medicine and International Health* 20.8 (2015), págs. 1086-1091. DOI: 10.1111/tmi.12521.

¹⁵ Joshua A. Salomon et al.: *Forecasting epidemiological transitions: multilevel modeling of time-series and contextual data for health policy planning*. En: *The Lancet* 380.9859 (2012), págs. 2063-2071. DOI: 10.1016/S0140-6736(12)61719-7.

4. Modelización ARIMA

Este capítulo proporciona los fundamentos teóricos necesarios para el análisis de series de tiempo. Se describen conceptos clave como la estacionariedad, la autocorrelación, los modelos autorregresivos integrados de media móvil (ARIMA) y los modelos de regresión dinámica, los cuales son fundamentales para la modelación y predicción de fenómenos temporales.

Las definiciones incluidas a continuación han sido extraídas, en su mayoría, de¹⁶.

4.1. Procesos estocásticos y series temporales

Definición 4.1. Un proceso estocástico es una colección de variables aleatorias $(X_t)_{t \in T}$, indexadas por un conjunto T . Cada una de estas variables están definidas sobre el espacio de probabilidad $(\Omega, \mathfrak{F}, P)$.

Se puede interpretar el conjunto T como un parámetro temporal $\{0, \dots, n\}$. De este modo, un proceso estocástico puede entenderse como la evolución en el tiempo de un fenómeno cuya dinámica está determinada por el azar.

¹⁶ Benjamín Vega: *Notas de Clase: Series de Tiempo*. Último acceso: 18 de marzo de 2025. 2023. <https://github.com/benjov/Series-Tiempo/blob/main/docs/Notas-Series-Tiempo.pdf>; José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>; Universitat Oberta de Catalunya (UOC): *Econometría - Módulo 5: Modelos de regresión dinámicos y multiecuacionales*. Último acceso: 19 de marzo de 2025. 2018. https://openaccess.uoc.edu/bitstream/10609/71665/1/Econometr%C3%ADa_M%C3%B3dulo%205_Modelos%20de%20regresi%C3%B3n%20din%C3%A1micos%20y%20multiecuacionales.pdf.

Definición 4.2. Una serie de tiempo es un conjunto de variables aleatorias indexadas con respecto al tiempo, es decir, una serie de tiempo es el proceso estocástico $\{X_t\}_{t \geq 0}$ con $t \geq 0$.

A partir de esta sección, se utilizará la notación $\{Y_t\}_{t \geq 0}$ para referirse a una serie de tiempo. Por ejemplo, considérese $Y(t)$ como el número de casos confirmados de Chikungunya en Bucaramanga, Santander, Colombia en el tiempo t . El proceso $\{Y_t\}_{t \geq 0}$ denota el número de casos confirmados durante el tiempo $t \geq 0$.

Para simplificar la notación, no se distinguirá entre el proceso estocástico y una de sus realizaciones; en otras palabras, una serie de tiempo se representará siempre como la secuencia mostrada en la ecuación correspondiente. El proceso estocástico puede describirse completamente mediante su función de probabilidad conjunta.

Definición 4.3. Sea (Y_1, Y_2, \dots, Y_n) un vector de variables aleatorias reales definidas sobre un mismo espacio de probabilidad. La **función de distribución conjunta** $F_{Y_1, \dots, Y_n} : \mathbb{R}^n \rightarrow [0, 1]$ se define como:

$$F_{Y_1, \dots, Y_n}(y_1, y_2, \dots, y_n) = P(Y_1 \leq y_1, \dots, Y_n \leq y_n)$$

Esta función proporciona la probabilidad de que cada variable aleatoria Y_i tome un valor menor o igual que y_i de forma simultánea. Esta definición se utilizará más adelante. Por tanto, nos centraremos únicamente en los dos primeros momentos¹⁷.

¹⁷ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

El primer momento de un proceso estocástico viene dado por el conjunto de *medias* de todas las variables aleatorias del proceso:

$$\mathbb{E}[Y_t], \quad t \geq 0$$

El segundo momento centrado del proceso viene dado por el conjunto de *varianzas* de todas las variables aleatorias del proceso:

$$\text{Var}[Y_t] = \mathbb{E}[(Y_t - \mathbb{E}[Y_t])^2] \quad \text{para } t \geq 0$$

Y las $\frac{T(T-1)}{2}$ covarianzas asociadas entre los elementos de la serie.

$$\text{Cov}[Y_t, Y_s] = \mathbb{E}[(Y_t - \mathbb{E}[Y_t])(Y_s - \mathbb{E}[Y_s])] \quad \forall s \neq t, s, t \geq 0$$

Las covarianzas son distintas entre sí debido a que cumplen con la propiedad de simetría, es decir, $\text{Cov}[Y_t, Y_s] = \text{Cov}[Y_s, Y_t]$ cuando $t \neq s$.

4.2. Procesos estocásticos estacionarios

Se dice que un proceso estocástico es **estrictamente estacionario** si su comportamiento probabilístico no se ve afectado por un desplazamiento en el tiempo¹⁸. Es decir, la función de distribución conjunta de cualquier subconjunto de n observaciones permanece invariante ante un cambio en el origen temporal.

¹⁸ Instituto de Estadísticas de Puerto Rico: *Análisis de Series de Tiempo: Fundamentos y Aplicaciones*. Inf. téc. Consultado el [Fecha] en [Hora]. Gobierno de Puerto Rico, 2023, pág. 4.54. http://www.estadisticas.gobierno.pr/iepr/LinkClick.aspx?fileticket=4_BxecUaZmg%3D.

Definición 4.4. Se dice que un proceso estocástico $\{Y_t\}_{t \geq 0}$ es **estrictamente estacionario** si, para cualquier entero k , cualquier número natural n , y cualquier conjunto de tiempos t_1, t_2, \dots, t_n se cumple que:

$$F_{Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}}(y_1, y_2, \dots, y_n) = F_{Y_{t_1+k}, Y_{t_2+k}, \dots, Y_{t_n+k}}(y_1, y_2, \dots, y_n) \\ \forall t > 0 \text{ y todo } (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$$

En la práctica, **la identificación de un proceso estrictamente estacionario es una tarea compleja**, principalmente debido a las dificultades en la recolección y disponibilidad de datos. Verificar esta propiedad exige demostrar que **todas las funciones de distribución conjunta** del proceso permanecen invariantes ante cualquier desplazamiento temporal, lo cual es generalmente inalcanzable en aplicaciones empíricas, dado que se requiere información completa sobre la estructura probabilística del proceso en diferentes momentos del tiempo¹⁹.

En teoría, verificar si un proceso es estrictamente estacionario se deben seguir:

1. Obtener la función de distribución conjunta del proceso para cualquier número n de variables aleatorias y cualquier conjunto de tiempos t_1, \dots, t_n .
2. Verificar que para cualquier desplazamiento $k \in \mathbb{Z}$, se cumpla que:

¹⁹ Peter J. Brockwell y Richard A. Davis: *Introduction to Time Series and Forecasting*. 3rd. Springer Texts in Statistics. Springer, 2016. ISBN: 978-3-319-29852-8. DOI: 10.1007/978-3-319-29854-2. <https://doi.org/10.1007/978-3-319-29854-2>; José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>; Benjamín Vega: *Notas de Clase: Series de Tiempo*. Último acceso: 18 de marzo de 2025. 2023. <https://github.com/benjov/Series-Tiempo/blob/main/docs/Notas-Series-Tiempo.pdf>.

$$F_{Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}}(y_1, y_2, \dots, y_n) = F_{Y_{t_1+k}, Y_{t_2+k}, \dots, Y_{t_n+k}}(y_1, y_2, \dots, y_n) \\ \forall (y_1, \dots, y_n) \in \mathbb{R}^n$$

3. Esto debe cumplirse para **todo** $n \geq 1$, todos los conjuntos de tiempos t_1, \dots, t_n , y para **todo** k , lo cual implica un conocimiento completo del proceso estocástico, algo que no se puede garantizar en la práctica con una muestra finita.

Por esta razón, en la mayoría de los casos prácticos se trabaja bajo el supuesto de **estacionariedad débil** o **estacionariedad en los momentos**, que requiere condiciones menos estrictas y puede evaluarse empíricamente mediante técnicas estadísticas²⁰.

Definición 4.5. Un proceso estocástico $\{Y_t\}_{t \geq 0}$ es débilmente estacionario (o sólo estacionario) si cumple con las siguientes condiciones:

1. *Estacionariedad en media:* Un proceso estocástico es estacionario en media si $\mathbb{E}[Y_t] = \mu_t = \mu$ es constante $\forall t > 0$.
2. *Estacionariedad en varianza:* Un proceso estocástico es estacionario en varianza si $Var[Y_t] = \mathbb{E}[(Y_t - \mu_t)^2] = \sigma_Y^2$ es constante y finita para $\forall t > 0$.
3. *Estacionariedad en covarianza:* Un proceso estocástico es estacionario en covarianza si $Cov[Y_t, Y_s] = \mathbb{E}[(Y_t - \mu_t)(Y_s - \mu_s)]$ es una función del tiempo y la distancia entre dos variables aleatorias.

²⁰ Peter J. Brockwell y Richard A. Davis: *Introduction to Time Series and Forecasting*. 3rd. Springer Texts in Statistics. Springer, 2016. ISBN: 978-3-319-29852-8. DOI: 10.1007/978-3-319-29854-2. <https://doi.org/10.1007/978-3-319-29854-2>; José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>; Benjamín Vega: *Notas de Clase: Series de Tiempo*. Último acceso: 18 de marzo de 2025. 2023. <https://github.com/benjov/Series-Tiempo/blob/main/docs/Notas-Series-Tiempo.pdf>.

Dado que la varianza es un caso particular de la covarianza cuando $t = s$, al cumplirse la estacionariedad en covarianza también se garantiza que la varianza sea constante. Por lo tanto, la estacionariedad débil implica tanto estacionariedad media como en varianza²¹.

A partir de ahora, cada vez que se mencione que un proceso es estacionario, nos referiremos al caso débil y utilizaremos únicamente el término estacionario.

Ejemplo 4.6. Sea una serie de tiempo $\{U_t\}_{t \geq 0}$. Se dice que el proceso estocástico es un *proceso puramente aleatorio* o un *ruido blanco*, si cumple las siguientes características:

- $\mathbb{E}[U_t] = 0, \forall t;$
- $Var[U_t] = \mathbb{E}[(U_t - \mu_t)^2] = \mathbb{E}[(U_t - \mu)^2] = \mathbb{E}[(U_t)^2] = \sigma^2, \forall t, y$
- $Cov[U_t, U_s] = \mathbb{E}[(U_t - \mu_t)(U_s - \mu_s)] = \mathbb{E}[(U_t - \mu)(U_s - \mu)] = \mathbb{E}[U_t U_s] = 0, \forall t \neq s.$

Un proceso $\{U_t\}_{t \geq 0}$ que cumple estas condiciones se denomina ruido blanco. En otras palabras, un ruido blanco es un proceso estocástico con media cero, varianza constante y finita, y sin correlación entre observaciones en distintos momentos del tiempo²².

Esto implica que el valor U_t no depende del pasado, puesto que la covarianza entre U_t y U_s con $s \neq t$ es cero. Es decir, no existe conexión lineal entre las muestras del proceso

²¹ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

²² Benjamín Vega: *Notas de Clase: Series de Tiempo*. Último acceso: 18 de marzo de 2025. 2023. <https://github.com/benjov/Series-Tiempo/blob/main/docs/Notas-Series-Tiempo.pdf>.

en diferentes momentos, lo que asegura su independencia lineal a lo largo del tiempo.

A partir de ahora, se denotará como $\{U_t\}_{t \geq 0} \sim RB(0, \sigma^2)$, indicando que se trata de un ruido blanco con media cero y varianza σ^2 .

4.3. Función de autocorrelación

Se puede medir la fuerza de la interdependencia en el tiempo entre las variables aleatorias de un proceso estocástico utilizando la función de autocovarianza. Se le llama autocovarianza ya que evalúa la covarianza del proceso en relación a sí mismo, lo que significa que se analiza la relación entre dos observaciones Y_t y Y_{t+k} del mismo proceso, que están distanciadas por un desfase $k \in \mathbb{Z}^{23}$.

La función de autocovarianza para un proceso con media constante μ se define como:

$$\gamma(k) = Cov[Y_t, Y_{t+k}] = \mathbb{E}[(Y_t - \mu)(Y_{t+k} - \mu)]$$

Al normalizar esta función al dividirla por la varianza del proceso, se logra obtener la función de autocorrelación:

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{Cov[Y_t, Y_{t+k}]}{Var[Y_t]} \quad (1)$$

Esta modificación genera un valor adimensional, lo que significa que no depende de las unidades empleadas para medir, dado que las escalas específicas de la variable aleatoria

²³ Benjamín Vega: *Notas de Clase: Series de Tiempo*. Último acceso: 18 de marzo de 2025. 2023. <https://github.com/benjov/Series-Tiempo/blob/main/docs/Notas-Series-Tiempo.pdf>.

se suprimen. Por este motivo, la autocorrelación permite comparar la dependencia entre diferentes procesos, independientemente de las magnitudes absolutas que tomen sus valores²⁴.

La autocorrelación se representa frecuentemente con un gráfico de barras conocido como correlograma. En esta visualización, el eje horizontal (x) representa los diversos desfases temporales, también conocidos como lags o rezagos, mientras que el eje vertical (y) ilustra los coeficientes de autocorrelación correspondientes a cada desfase²⁵.

El correlograma es una herramienta esencial para detectar patrones de dependencia temporal en los datos, tales como estacionalidad o aleatoriedad.

Este tipo de análisis se utiliza principalmente en procesos que son estacionarios, dado que en estos la autocorrelación se basa solo en el desfase y no en el momento específico. No obstante, también puede usarse de manera exploratoria en series que no son estacionarias, aunque en esos casos la interpretación es más restringida y puede llevar a conclusiones erróneas si los datos no se transforman adecuadamente²⁶.

Definición 4.7. El coeficiente de autocorrelación de orden k de un proceso estocástico

²⁴ Benjamín Vega: *Notas de Clase: Series de Tiempo*. Último acceso: 18 de marzo de 2025. 2023. <https://github.com/benjov/Series-Tiempo/blob/main/docs/Notas-Series-Tiempo.pdf>.

²⁵ George EP Box, Gwilym M Jenkins y Gregory C Reinsel: *Time Series Analysis: Forecasting and Control*. 4th. Wiley, 2008; José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

²⁶ George EP Box, Gwilym M Jenkins y Gregory C Reinsel: *Time Series Analysis: Forecasting and Control*. 4th. Wiley, 2008.

estacionario mide el grado de asociación lineal existente entre dos variables aleatorias del proceso separadas de k períodos.

$$\rho_k = \frac{Cov[Y_t, Y_{t+k}]}{\sqrt{Var[Y_t]Var[Y_{t+k}]}}$$

Por ser un coeficiente de correlación, no depende de unidades y $|\rho_k| \leq 1, \forall k$.

4.4. Modelos lineales estacionarios

Debido a la compleja dinámica que presentan las series de tiempo, los modelos lineales resultan insuficientes para capturar y explicar adecuadamente su comportamiento, ya que no permiten representar patrones no lineales o estacionales inherentes a este tipo de datos. Esta limitación reduce la capacidad predictiva de dichos modelos.

En esta sección se examinan los métodos propuestos por Box y Jenkins (1970)²⁷; en este contexto, se estudian los procesos autorregresivos (Autoregressive, AR) y los procesos de medias móviles (Moving Average, MA). La combinación de ambos da lugar a los modelos ARMA.

Para la construcción de modelos efectivos, es fundamental introducir el operador de rezago, el cual se utiliza para representar la relación entre observaciones en distintos momentos del tiempo.

²⁷ Desconocido: *Modelos ARIMA: Metodología Box-Jenkins*. Último acceso: 17 de marzo de 2025. s.f. <https://www.estadistica.net/ECONOMETRIA/SERIES-TEMPORALES/modelo-arima.pdf>.

Supóngase una serie de tiempo $\{Y_t\}_{t \geq 0}$ donde cada Y_t representa el valor de una variable observada en el instante t . A partir de esta secuencia, se puede definir el operador de rezago como se indica a continuación:

Definición 4.8. El operador de rezago se denota por \mathcal{L} (lag, en inglés) y es tal que $\mathcal{L}Y_t = Y_{t-1}$.

Esto significa que el operador de rezago \mathcal{L} desplaza la variable una unidad de tiempo hacia el pasado. A partir de la definición, al aplicarlo a Y_{t-1} se puede obtener Y_{t-2} , lo que nos lleva a establecer la siguiente proposición:

Proposición 4.9. Sea $\{Y_t\}_{t \geq 0}$ una serie de tiempo. El operador de rezago \mathcal{L} definido como $\mathcal{L}Y_t = Y_{t-1}$, cumple que:

$$\mathcal{L}(Y_{t-1}) = \mathcal{L}\mathcal{L}(Y_t) = \mathcal{L}^2Y_t = Y_{t-2} \quad (2)$$

Por lo que, en general, obtendremos:

$$\mathcal{L}^p Y_t = Y_{t-p} \quad p \in \mathbb{Z} \quad (3)$$

Demostración. Se demuestra la propiedad por inducción matemática sobre p , con $p \geq 2$.

Caso base: Para $p = 2$, se tiene:

$$\mathcal{L}^2 Y_t = \mathcal{L}(\mathcal{L}Y_t) = \mathcal{L}(Y_{t-1}) = Y_{t-2}.$$

Por tanto, la propiedad se cumple para $p = 2$.

Paso inductivo: Supóngase que para algún $k \geq 2$ se cumple la hipótesis de inducción:

$$\mathcal{L}^k Y_t = Y_{t-k}. \quad (\text{HI})$$

Se desea demostrar que:

$$\mathcal{L}^{k+1} Y_t = Y_{t-(k+1)}.$$

Aplicando la definición del operador de rezago y la hipótesis de inducción:

$$\mathcal{L}^{k+1} Y_t = \mathcal{L}(\mathcal{L}^k Y_t) = \mathcal{L}(Y_{t-k}) = Y_{t-k-1} = Y_{t-(k+1)}.$$

Con esto se concluye que la propiedad es válida para $p = k + 1$. Por lo tanto, por el principio de inducción matemática, la afirmación es cierta para todo $p \in \mathbb{Z}_{\geq 2}$. \square

Observación 4.10. De lo anterior, podemos inferir el siguiente resultado:

$$\Delta Y_t = Y_t - Y_{t-1} = (1 - \mathcal{L})Y_t \quad (4)$$

Lo que nos permite generalizar esta relación de diferencias a un orden k . En efecto, considerando $\Delta = (1 - \mathcal{L})$ y utilizando la Proposición 4.9 junto con la Definición 4.8, se tiene:

$$\begin{aligned} \Delta^{k-1} Y_t - \Delta^{k-1} Y_{t-1} &= (1 - \mathcal{L})^{k-1} Y_t - (1 - \mathcal{L})^{k-1} \mathcal{L} Y_t \\ &= (1 - \mathcal{L})^{k-1} (Y_t - \mathcal{L} Y_t) \\ &= (1 - \mathcal{L})^{k-1} (1 - \mathcal{L}) Y_t \\ &= (1 - \mathcal{L})^k Y_t = \Delta^k Y_t \end{aligned} \quad (5)$$

Esta notación resulta de gran utilidad cuando se quiere comparar períodos equivalentes como, por ejemplo, el mismo trimestre pero de un año anterior.

4.4.1. Modelo lineal general Dado que el objetivo es explicar el valor que toma, en el instante t , una variable con dependencia temporal, una forma adecuada de abordarlo radica en recopilar información sobre el pasado de dicha variable, observar su evolución temporal y aprovechar el patrón de regularidad reflejado en los datos. En un modelo univariante de series temporales, la serie $\{Y_t\}_{t \geq 0}$ se descompone en dos componentes: una que representa el patrón de regularidad, o parte sistemática, y otra puramente aleatoria, conocida como *innovación*²⁸.

La parte *sistemática* de un modelo de series de tiempo corresponde a la componente predecible con base en el conjunto de información utilizado para su construcción, es decir, la propia serie temporal. Por su parte, la *innovación* representa la parte aleatoria del modelo, cuya característica principal radica en que sus valores no representan relación ni dependencia entre sí, en relación con la información previamente disponible²⁹.

En el estudio de los procesos estocásticos estacionarios, se considera únicamente la clase de procesos lineales, los cuales se caracterizan por poder representarse como combinaciones lineales de variables aleatorias. En particular, cuando se trata de procesos

²⁸ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

²⁹ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

estacionarios con distribución normal y media cero, la descomposición de Wold garantiza que cualquier proceso estocástico débilmente estacionario puede expresarse como una combinación lineal infinita de innovaciones no anticipantes³⁰:

$$Y_t = \sum_{j=0}^{\infty} \pi_j U_{t-j}, \quad \text{con} \quad \sum_{j=0}^{\infty} \pi_j^2 < \infty,$$

Donde $\{U_t\}_{t \geq 0} \sim RB(0, \sigma^2)$, no correlacionado y con media cero.

Esta representación, requiere que el proceso sea no anticipante, es decir que el valor presente Y_t solo dependa de innovaciones presentes y pasadas, no futuras ($Y_t = f(U_s : s \leq t)$). Esta propiedad garantiza la causalidad del modelo, indispensable tanto para la interpretación como para la predicción práctica³¹.

Por otro lado, la invertibilidad implica que las innovaciones $\{U_t\}_{t \geq 0}$ pueden construirse a partir de las observaciones pasadas de $\{X_t\}_{t \geq 0}$. Formalmente, un proceso es invertible cuando la expansión:

$$U_t = \sum_{j=0}^{\infty} \psi_j Y_{t-j}, \quad \text{con} \quad \sum_{j=0}^{\infty} \psi_j^2 < \infty.$$

³⁰ Peter J. Brockwell y Richard A. Davis: *Introduction to Time Series and Forecasting*. 3rd. Springer Texts in Statistics. Springer, 2016. ISBN: 978-3-319-29852-8. DOI: 10.1007/978-3-319-29854-2. <https://doi.org/10.1007/978-3-319-29854-2>; José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

³¹ Peter J. Brockwell y Richard A. Davis: *Introduction to Time Series and Forecasting*. 3rd. Springer Texts in Statistics. Springer, 2016. ISBN: 978-3-319-29852-8. DOI: 10.1007/978-3-319-29854-2. <https://doi.org/10.1007/978-3-319-29854-2>.

La ecuación indica que estas innovaciones pueden recuperarse a partir de los valores pasados observados de la serie $\{Y_t\}_{t \geq 0}$, siempre que la secuencia de coeficientes $\{\psi_j\}_{j \geq 0}$ es cuadrado sumable, es decir, si la serie de los cuadrados de sus términos converge³².

Sea un proceso $\{Y_t\}_{t \geq 0}$ que tiene una representación del tipo:

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \pi_3 Y_{t-3} + \dots + U_t, \quad \forall t > 0 \quad (6)$$

donde $\{U_t\}_{t \geq 0} \sim RB(0, \sigma^2)$. El modelo (6) se puede escribir de forma más compacta en términos del operador de rezagos:

$$\begin{aligned} Y_t &= (\pi_1 \mathcal{L} + \pi_2 \mathcal{L}^2 + \dots) Y_t + U_t \\ (1 - \pi_1 \mathcal{L} + \pi_2 \mathcal{L}^2 + \dots) Y_t &= U_t \\ \implies \Pi_\infty(\mathcal{L}) Y_t &= U_t \end{aligned}$$

Otra representación alternativa de escribir el modelo (6) es:

$$Y_t = \frac{1}{\Pi_\infty(\mathcal{L})} U_t = \Psi_\infty(\mathcal{L}) U_t = (1 + \psi_1(\mathcal{L}) + \psi_2(\mathcal{L})^2 + \dots) U_t \quad (7)$$

$$Y_t = U_t - \psi_1 U_{t-1} + \psi_2 U_{t-2} + \dots \quad t = 1, 2, \dots$$

La representación dada en la ecuación (6) y (7) es válida para procesos estocásticos

³² Peter J. Brockwell y Richard A. Davis: *Introduction to Time Series and Forecasting*. 3rd. Springer Texts in Statistics. Springer, 2016. ISBN: 978-3-319-29852-8. DOI: 10.1007/978-3-319-29854-2. <https://doi.org/10.1007/978-3-319-29854-2>.

estacionarios siempre que se cumplan dos condiciones fundamentales: que el proceso sea no anticipante e invertible.

Dado que en la práctica se trabaja con series temporales de longitud finita, no es viable representar dependencias infinitas sin imponer restricciones adicionales. Por tanto, resulta necesario simplificar las formulaciones generales como la expresada en (6) y (7), utilizando modelos con un número finito de parámetros³³. Acudiendo a la teoría de polinomios, bajo condiciones muy generales, se puede aproximar un polinomio de orden infinito mediante un cociente de polinomios finitos :

$$\Pi_{\infty}(\mathcal{L}) \simeq \frac{\phi_p(\mathcal{L})}{\theta_q(\mathcal{L})}$$

donde $\phi_p(\mathcal{L})$ y $\theta_q(\mathcal{L})$ son polinomios en el operador de retardos finitos de orden p y q , respectivamente:

$$\begin{aligned}\phi_p(\mathcal{L}) &= 1 - \phi_1\mathcal{L} - \phi_2\mathcal{L}^2 - \dots - \phi_p\mathcal{L}^p \\ \theta_q(\mathcal{L}) &= 1 - \theta_1\mathcal{L} - \theta_2\mathcal{L}^2 - \dots - \theta_q\mathcal{L}^q\end{aligned}$$

Sustituyendo en el modelo (6), se obtiene:

$$\Pi_{\infty}(\mathcal{L})Y_t \simeq \frac{\phi_p(\mathcal{L})}{\theta_q(\mathcal{L})}Y_t = U_t \longrightarrow \phi_p(\mathcal{L})Y_t = \theta_q(\mathcal{L})U_t$$

³³ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

Por lo tanto, el modelo lineal general admite tres representaciones, todas igualmente válidas bajo los supuestos señalados.

- Representación *puramente autorregresiva* (6), $AR(\infty)$: el valor presente de la variable se representa en función de su propio pasado más una innovación contemporánea.
- Representación *puramente de medias móviles* (7), $MA(\infty)$: el valor presente de la variable se representa en función de todas las innovaciones presentes y pasadas.
- Representación *finita*:

$$\begin{aligned} \phi_p(\mathcal{L})Y_t &= \theta_q(\mathcal{L})U_t \\ (1 - \phi_1\mathcal{L} - \phi_2\mathcal{L}^2 - \dots - \phi_p\mathcal{L}^p)Y_t &= (1 - \theta_1\mathcal{L} - \theta_2\mathcal{L}^2 - \dots - \theta_q\mathcal{L}^q)U_t \\ Y_t &= \underbrace{\phi_1Y_{t-1} + \phi_2Y_{t-2} + \dots + \phi_pY_{t-p}}_{\text{parte autorregresiva}} + \underbrace{U_t + \theta_1U_{t-1} + \theta_2U_{t-2} + \dots + \theta_qU_{t-q}}_{\text{parte de medias móviles}} \end{aligned}$$

En este modelo finito, el valor de Y_t depende del proceso hasta valor en el tiempo $t - p$, es decir, Y_{t-p} (parte autorregresiva) y de la innovación hasta el momento $t - q$, U_{t-q} (parte medias móviles). Este modelo se denomina *Autorregresivo de Medias Móviles* de orden (p, q) , y se denota por $ARMA(p, q)$ ³⁴.

Dos casos particulares del modelo $ARMA(p, q)$ de gran interés son:

- $AR(p)$. Modelo que solo presenta parte autorregresiva, es decir, el polinomio de medias móviles es de orden 0.

³⁴ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

$$Y_t \sim AR(p) \longrightarrow Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + U_t, \quad \{U_t\}_{t \geq 0} \sim RB(0, \sigma^2)$$

- $MA(q)$. Modelo que solo presenta parte medias móviles, es decir, el polinomio autorregresivo es de orden 0.

$$Y_t \sim MA(q) \longrightarrow Y_t = U_t - \theta_1 U_{t-1} - \theta_2 U_{t-2} - \dots - \theta_q U_{t-q}, \quad \{U_t\}_{t \geq 0} \sim RB(0, \sigma^2)$$

Cuando el modelo es conocido, cualquiera de sus representaciones es válida según el propósito analítico. Sin embargo, si debe especificarse y estimarse a partir de datos, se recurre necesariamente a una formulación finita. En la construcción de modelos univariantes, el objetivo no es encontrar el “modelo verdadero”, sino obtener una representación parsimoniosa y adecuada de la dinámica reflejada en la función de autocorrelación³⁵.

4.4.2. Procesos autorregresivos (AR) Los procesos autorregresivos tienen su origen en el trabajo de Cochrane y Orcutt en 1949, mediante el cual analizaron los residuales de una regresión clásica como un proceso autorregresivo³⁶.

Estos representan un tipo de proceso aleatorio en el que la variable de interés se explica a través de una combinación lineal de sus observaciones pasadas y un término de error³⁷.

³⁵ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

³⁶ Benjamín Vega: *Notas de Clase: Series de Tiempo*. Último acceso: 18 de marzo de 2025. 2023. <https://github.com/benjov/Series-Tiempo/blob/main/docs/Notas-Series-Tiempo.pdf>.

³⁷ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

Como una generalización de los procesos autorregresivos (AR), podemos definir el modelo AR(p) de la siguiente manera:

Definición 4.11. Sea $\{Y_t\}_{t \geq 0}$ una serie temporal definida sobre un espacio de probabilidad (Ω, \mathcal{F}, P) . Se dice que $\{Y_t\}_{t \geq 0}$ sigue un modelo autorregresivo de orden p , denotado como AR(p), si satisface la siguiente ecuación:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + U_t, \quad \forall t > 0, \quad (8)$$

donde $\phi_1, \dots, \phi_p \in \mathbb{R}$ son los coeficientes del modelo y $\{U_t\}_{t \geq 0} \sim RB(0, \sigma^2)$.

Observación 4.12. En términos del operador de rezago (4.8). Obtenemos, la siguiente expresión de la ecuación (8):

$$(1 - \phi_1 \mathcal{L} - \phi_2 \mathcal{L}^2 - \phi_3 \mathcal{L}^3 - \cdots - \phi_p \mathcal{L}^p) Y_t = U_t \quad (9)$$

Definamos el polinomio $\phi_p(\mathcal{L})$ como el polinomio autorregresivo y $(\phi_1, \phi_2, \dots, \phi_p)$ es el vector de parámetros autorregresivos.

$$\phi(\mathcal{L}) = 1 - \phi_1 \mathcal{L} - \phi_2 \mathcal{L}^2 - \phi_3 \mathcal{L}^3 - \cdots - \phi_p \mathcal{L}^p \quad (10)$$

Así,

$$\phi_p(\mathcal{L}) Y_t = U_t \quad (11)$$

En primer lugar, es preciso comprobar si el proceso AR(p) cumple con las condiciones de estacionariedad para cualquier valor de los parámetros. Esta comprobación se complica

para modelos autorregresivos de orden mayor, cuyos parámetros han de satisfacer restricciones complejas para ser estacionarios³⁸.

El siguiente teorema proporciona condiciones necesarias y suficientes para que el modelo AR(p) sea estacionario.

Teorema 4.13. *Un proceso autorregresivo finito AR(p) es estacionario si, y solo si, el módulo de las raíces del polinomio autorregresivo está fuera del círculo unitario.*

Observación 4.14. La demostración del resultado anterior se basa en encontrar las soluciones de la ecuación característica (10). Si las raíces son reales, su valor absoluto debe ser mayor que 1. Si son complejas, su módulo debe ser mayor que 1³⁹.

4.4.3. Procesos de medias móviles (MA) Los procesos de medias móviles representan un tipo de proceso aleatorio en el que los valores actuales dependen linealmente del error presente y de varios errores pasados.

Definición 4.15. Sea la serie temporal $\{Y_t\}_{t \geq 0}$ definida sobre un espacio de probabilidad (Ω, \mathcal{F}, P) . El proceso de medias móviles de orden q , MA(q), puede ser escrito como:

$$Y_t = U_t - \theta_1 U_{t-1} - \theta_2 U_{t-2} - \dots - \theta_q U_{t-q}, \quad \{U_t\}_{t \geq 0} \sim RB(0, \sigma^2). \quad (12)$$

³⁸ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

³⁹ Peter J. Brockwell y Richard A. Davis: *Introduction to Time Series and Forecasting*. 3rd. Springer Texts in Statistics. Springer, 2016. ISBN: 978-3-319-29852-8. DOI: 10.1007/978-3-319-29854-2. <https://doi.org/10.1007/978-3-319-29854-2>; José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

Observación 4.16. Se puede reescribir la ecuación (12) utilizando el operador de rezago. Así, obtenemos el proceso de MA(q) como sigue a continuación:

$$Y_t = (1 - \theta_1 \mathcal{L} - \theta_2 \mathcal{L}^2 - \theta_3 \mathcal{L}^3 - \dots - \theta_q \mathcal{L}^q) U_t \quad (13)$$

Definamos el polinomio $\theta_q(\mathcal{L})$ como el polinomio de medias móviles y $(\theta_1, \theta_2, \theta_3, \dots, \theta_q)$ es el vector de parámetros de medias móviles:

$$Y_t = \theta_q(\mathcal{L})U_t \quad (14)$$

El modelo MA(q) es *no anticipante* porque el futuro no influye en el pasado y será *invertible* si su representación autorregresiva es tal que la influencia Y_{t-k} es menor conforme nos alejamos en el pasado⁴⁰. Esta condición se cumple si:

$$\sum_{i=1}^{\infty} \pi_i^2 < \infty$$

Por lo tanto, el modelo no es invertible para cualquier valor del vector de parámetros de medias móviles, sino que estas tendrán que cumplir algunas restricciones⁴¹.

⁴⁰ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

⁴¹ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

El siguiente teorema proporciona condiciones necesarias y suficientes para que un modelo de medias móviles sea invertible.

Teorema 4.17. *Un proceso de medias móviles finito $MA(q)$ es invertible si, y solo si, el módulo de las raíces del polinomio de medias móviles está fuera del círculo unitario.*

4.4.4. Procesos ARMA(p,q) Ahora analicemos cómo los modelos autorregresivos (AR) y medias móviles (MA) se combinan en un modelo general llamado ARMA(p, q).

Definición 4.18. Sea la serie temporal $\{Y_t\}_{t \geq 0}$ definida sobre un espacio de probabilidad (Ω, \mathcal{F}, P) . La especificación general de un modelo ARMA(p, q), donde $p, q \in \mathbb{N}$, consiste en expresar $\{Y_t\}_{t \geq 0}$ como una función lineal de sus valores pasados hasta el retardo p , de la innovación contemporánea y de los retardos hasta el orden q del proceso de innovaciones $\{U_t\}_{t \geq 0}$. Formalmente:

$$Y_t = \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + U_t + \theta_1 U_{t-1} + \dots + \theta_q U_{t-q}, \quad \{U_t\}_{t \geq 0} \sim RB(0, \sigma^2) \quad (15)$$

Observación 4.19. Mediante el uso del operador de rezago (4.8) se puede escribir la ecuación (15) como:

$$(1 - \phi_1 \mathcal{L} - \dots - \phi_p \mathcal{L}^p) Y_t = (1 - \theta_1 \mathcal{L} - \dots - \theta_q \mathcal{L}^q) U_t \quad (16)$$

En la ecuación (16), $\phi_p(\mathcal{L})$ es el polinomio autorregresivo y $\theta_q(\mathcal{L})$ es el polinomio de medias móviles. Así,

$$\phi_p(\mathcal{L}) Y_t = \theta_q(\mathcal{L}) U_t \quad (17)$$

Teorema 4.20. *Un proceso autorregresivo de medias móviles finito ARMA(p, q) es estacionario si, y solo si, el módulo de las raíces del polinomio autorregresivo está fuera del círculo unitario.*

Las condiciones de estacionariedad vienen impuestas por la parte autorregresiva, dado que la parte de medias móviles finita siempre es estacionaria. Para comprobar si el modelo no es anticipante e invertible, se estudia su representación autorregresiva general⁴².

Teorema 4.21. *Un proceso autorregresivo de medias móviles finito ARMA(p, q) es invertible si, y solo si, el módulo de las raíces del polinomio medias móviles está fuera del círculo unitario.*

Las condiciones de invertibilidad del modelo ARMA(p, q) vienen impuestas por la parte de medias móviles, dado que la parte autorregresiva finita siempre es invertible porque está directamente escrita en forma autorregresiva⁴³.

4.5. Modelos ARIMA(p, d, q)

Supongamos el siguiente modelo ARMA(p, q):

⁴² Peter J. Brockwell y Richard A. Davis: *Introduction to Time Series and Forecasting*. 3rd. Springer Texts in Statistics. Springer, 2016. ISBN: 978-3-319-29852-8. DOI: 10.1007/978-3-319-29854-2. <https://doi.org/10.1007/978-3-319-29854-2>; José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

⁴³ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

$$\Phi_p(\mathcal{L})Y_t = \Theta_q(\mathcal{L})U_t, \quad \{U_t\}_{t \geq 0} \sim \text{RB}(0, \sigma^2) \quad (18)$$

Donde el polinomio AR se puede factorizar en función de sus p raíces $\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_p$,

$$\Phi_p(\mathcal{L}) = (1 - \mathcal{L}_1^{-1}\mathcal{L})(1 - \mathcal{L}_2^{-1}\mathcal{L}) \cdots (1 - \mathcal{L}_p^{-1}\mathcal{L})$$

Considere que las $(p - 1)$ raíces son estacionarias (con módulo fuera del círculo unitario) y una de ellas es unitaria, $\mathcal{L}_i = 1$. Entonces, el polinomio AR, se puede reescribir como sigue:

$$\Phi_p(\mathcal{L}) = (1 - \mathcal{L}_1^{-1}\mathcal{L})(1 - \mathcal{L}_2^{-1}\mathcal{L}) \cdots (1 - \mathcal{L}_p^{-1}\mathcal{L}) = \varphi_{p-1}(\mathcal{L})(1 - \mathcal{L})$$

Donde el polinomio $\varphi_{p-1}(\mathcal{L})$ resulta del producto de los $(p - 1)$ polinomios de orden 1 asociados a las raíces L_i con módulo fuera del círculo unitario. Sustituyendo en el modelo (18) se tiene que:

$$\varphi_{p-1}(\mathcal{L})(1 - \mathcal{L})Y_t = \Theta_q(\mathcal{L})U_t \longrightarrow \varphi_{p-1}(\mathcal{L})\Delta Y_t = \Theta_q(\mathcal{L})U_t \quad (19)$$

El polinomio $\varphi_{p-1}(\mathcal{L})$ es estacionario porque todas sus raíces tienen módulo fuera del círculo unitario y el polinomio $\Delta = (1 - \mathcal{L})$ es el que recoge la raíz unitaria.

El modelo (19) representa el comportamiento del proceso Y_t que no es estacionario porque tiene una raíz unitaria. A un proceso Y_t con estas características se le denomina *proceso integrado de orden 1*. En general, el polinomio del modelo (18) puede contener más de una raíz unitaria; por ejemplo, d , entonces se puede descomponer como:

$$\Phi_p(\mathcal{L}) = \varphi_{p-d}(\mathcal{L})(1 - \mathcal{L})^d$$

y sustituyendo, de nuevo, en el modelo ARMA(p, q) (18), se tiene:

$$\varphi_{p-d}(\mathcal{L})\Delta^d Y_t = \Theta_q(\mathcal{L})U_t$$

El polinomio $\varphi_{p-d}(\mathcal{L})$ es estacionario porque sus $(p - d)$ raíces tienen módulo fuera del círculo unitario, y el polinomio $\Delta^d = (1 - \mathcal{L})^d$, de orden d , contiene las d raíces unitarias no estacionarias. A un proceso Y_t con estas características se le denomina *proceso integrado de orden d* y se denota por $Y_t \sim I(d)$.

Definición 4.22. Un proceso Y_t es integrado de orden d , $Y_t \sim I(d)$, si Y_t no es estacionario, pero su diferencia de orden d , $\Delta^d Y_t$, sigue un proceso ARMA($p - d, q$) estacionario e invertible.

El orden de integración del proceso es el número de diferencias que hay que tomar al proceso para conseguir la estacionariedad en media, o lo que es lo mismo, el número de raíces unitarias del proceso.

Definición 4.23. Se dice que una serie $\{Y_t\}_{t \geq 0}$ es integrada de orden d , si se puede representar por el siguiente modelo:

$$\Phi_p(\mathcal{L})\Delta^d Y_t = \delta + \Theta_q(\mathcal{L})U_t, \quad U_t \sim RB(0, \sigma^2) \quad (20)$$

Donde el polinomio autorregresivo estacionario $\Phi_p(\mathcal{L})$ y el invertible de medias móviles $\Theta_q(\mathcal{L})$ no tienen raíces comunes, δ es una constante para que su media no sea cero. El modelo (20) se denomina **Autoregresivo Integrado de Medias Móviles de orden (p, d, q)** o **ARIMA**(p, d, q), donde p es el orden del polinomio autorregresivo estacionario,

d es el orden de integración de la serie, es decir, el número de diferencias que hay que tomar a la serie para que sea estacionaria y q es el orden del polinomio de medias móviles invertible⁴⁴.

4.5.1. Selección de los órdenes p , q , d en un modelo ARIMA(p,d,q) La manera más convencional y formal para determinar el orden p y q del proceso ARMA(p, q) es utilizar los criterios de información, como AIC (Criterio de Información de Akaike) o el BIC (Criterio de Información Bayesiano). Así, el orden se elige de acuerdo con el valor mínimo del criterio de información. En el caso del parámetro d , se selecciona revisando la gráfica de autocorrelación parcial o la serie temporal diferenciada. El valor de d indica el número de diferenciaciones necesarias para que la serie de tiempo sea estacionaria, es decir, que tenga propiedades estadísticas constantes a lo largo del tiempo como la media y varianza⁴⁵.

4.5.2. Estrategia de modelización ARIMA La construcción de modelos ARIMA(p, d, q) se lleva a cabo de forma iterativa mediante un proceso en el que se pueden distinguir cuatro etapas⁴⁶.

⁴⁴ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

⁴⁵ Desconocido: *Modelos ARIMA: Metodología Box-Jenkins*. Último acceso: 17 de marzo de 2025. s.f. <https://www.estadistica.net/ECONOMETRIA/SERIES-TEMPORALES/modelo-arima.pdf>; José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>; Benjamín Vega: *Notas de Clase: Series de Tiempo*. Último acceso: 18 de marzo de 2025. 2023. <https://github.com/benjov/Series-Tiempo/blob/main/docs/Notas-Series-Tiempo.pdf>.

⁴⁶ Desconocido: *Modelos ARIMA: Metodología Box-Jenkins*. Último acceso: 17 de marzo de 2025. s.f. <https://www.estadistica.net/ECONOMETRIA/SERIES-TEMPORALES/modelo-arima.pdf>; José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025.

1. **Identificación.** Utilizando los datos y/o cualquier tipo de información disponible sobre cómo ha sido generada la serie, se intentará sugerir una subclase de modelos $ARIMA(p, d, q)$ que merezca la pena ser investigada. El objetivo es determinar los órdenes p, d, q que parecen apropiadas para reproducir las características de la serie bajo estudio y si se incluye o no la constante δ . En esta etapa es posible identificar más de un modelo candidato a haber podido generar la serie.
2. **Estimación.** Usando de forma eficiente los datos, se realiza inferencia sobre los parámetros condicionados a que el modelo investigado sea el apropiado.

Dado un determinado proceso propuesto, se trata de cuantificar los parámetros del mismo, $\theta_1, \dots, \theta_p, \phi_1, \dots, \phi_q$ y, en su caso δ .
3. **Validación.** Se realizan contrastes de diagnóstico para comprobar si el modelo se ajusta a los datos, o, si no es así, revelar las posibles discrepancias del modelo propuesto para mejorarlo.
4. **Predicción.** Obtener pronósticos en términos probabilísticos de los valores futuros de la variable. En esta etapa se tratará también de evaluar la capacidad predictiva del modelo.

4.6. Modelos ARIMA estacionales

Muchas series económicas, si se observan varias veces a lo largo del año, trimestral o mensualmente, presentan un comportamiento estacional. Este tipo de comportamiento puede ser debido a factores meteorológicos, tales como temperatura, pluviosidad, etc.⁴⁷

Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

⁴⁷ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025.

Así a la hora de elaborar un modelo ARIMA adecuado para una serie temporal se ha de tener en cuenta el comportamiento estacional, si lo hubiere, porque implica que la observación de un mes y observación del mismo mes del año anterior tienen una pauta de comportamiento similar por lo que estarán correlacionas. Por tanto, para este tipo de series deberá recoger las dos clases de dependencia inter temporal que presentan, a saber⁴⁸.

- La relación lineal existente entre observaciones sucesivas (comportamiento tendencial).
- La relación lineal existente entre observaciones del mismo mes en años sucesivos (comportamiento estacional).

Supongamos que la serie $\{Y_t\}_{t \geq 0}$ presenta un componente estacional y se especifica un modelo ARIMA(p, d, q) general:

$$\phi_p(L)Y_t = \theta_q U_t$$

Para recoger las dos estructuras de correlación anteriormente mencionadas, la regular y la estacional, bastaría con añadir al modelo los retardos de Y_t y U_t necesarios. Así, si la serie estacional es mensual y se quiere representar, además de la dependencia

Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

⁴⁸ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

entre observaciones consecutivas, la autocorrelación entre observaciones del mismo mes separadas, un año, dos años, etc., será necesario incluir en el modelo retardos hasta de orden 12,24, etc., con lo que supone de un aumento en el número de los parámetros del modelo⁴⁹.

4.6.1. Modelos ARIMA estacionales multiplicativos Consideremos una serie estacional $\{Y_t\}_{t \geq 0}$ con período s conocido, por ejemplo, una serie mensual vista durante N años, de forma que en total contamos con $T = 12N$ observaciones. Si la serie es estacional, podemos dividirla en 12 subseries, una por mes, de N observaciones que denotaremos

$$y_\tau^{(1)}, y_\tau^{(2)}, y_\tau^{(3)}, \dots, y_\tau^{(12)}, \quad \tau = 1, 2, \dots, N$$

La relación entre estas subseries y la serie de partida es:

$$y_\tau^{(j)} = Y_{j+12(\tau-1)} = Y_t, \quad \tau = 1, 2, \dots, N \quad j = 1, 2, \dots, 12 \quad (21)$$

Cada una de estas doce subseries no representa comportamiento estacional, por lo que podemos representarlas mediante los modelos ARIMA(p, d, q). Supongamos que el modelo ARIMA adecuado para las doce subseries y_τ^j es el mismo:

$$(1 - \Phi_1 \mathcal{L} - \dots - \Phi_p \mathcal{L}^p)(1 - \mathcal{L})^D y_\tau^{(j)} = (1 - \Theta_1 \mathcal{L} - \dots - \theta_q \mathcal{L}^q) u_\tau^{(j)}, \quad \tau = 1, 2, \dots, N \quad (22)$$

⁴⁹ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

Si hay estacionalidad en la serie de partida, Y_t se cumple que $D \geq 1$. Nótese que si $D = 0$ y, por lo tanto, las series $y_\tau^{(j)}$ fueran estacionarias, todos los modelos (22), al ser el mismo, tendrían la misma media, lo que es incompatible con el supuesto de estacionalidad que implica que la media de cada mes, o sea, de cada subserie es diferente.

Los modelos para las doce subseries, al ser todos iguales, se pueden escribir conjuntamente en función de la serie de partida Y_t , teniendo en cuenta que, la relación (21):

$$\mathcal{L}y_\tau^{(j)} = y_{\tau-1}^{(j)} = Y_{j+12(\tau-2)} = Y_{j-12+12(\tau-1)} = \mathcal{L}^{12}Y_{j+12(\tau-1)}$$

Lo que implica que aplicar el operador de rezago \mathcal{L} a $y_\tau^{(j)}$ es equivalente a aplicar el operador \mathcal{L}^{12} a la serie original $Y_{j+12(\tau-1)}$. Además, habrá que definir una serie de ruido común para las doce subseries, α_t , asignando a cada mes t el ruido del modelo univariante correspondiente a dicho mes.

$$u_\tau^{(j)} = \alpha_{j+12(\tau-1)}, \quad \tau = 1, 2, \dots, N$$

Aplicando ambos resultados al modelo (22), se obtiene el modelo conjunto para toda la serie mensual, Y_t :

$$(1 - \Phi_1\mathcal{L}^{12} - \dots - \Phi_p\mathcal{L}^{12p})(1 - \mathcal{L}^{12})^D y_\tau^{(j)} = (1 - \Theta_1\mathcal{L}^{12} - \dots - \theta_q\mathcal{L}^{12q})\alpha_t, \quad t = 1, 2, \dots, 12N$$

Para cada uno de los modelos (22) las series $u_\tau^{(j)}$ son, por construcción, ruido blanco, pero la serie conjunta, $\alpha_t, t = 1, 2, \dots, T$, no tiene por qué serlo, en general, ya que la mayoría de las veces existirá dependencia entre las observaciones contiguas que, como no ha sido todavía tenida en cuenta, quedará recogida en esta serie con ruido. Es lo

que se denomina *estructura regular* o estructura asociada a los intervalos naturales de medida de la serie (meses en nuestro caso). Suponiendo que α_t sigue el modelo ARIMA no estacional siguiente:

$$\phi_p(\mathcal{L})(1 - \mathcal{L})^d \alpha_t = \theta_q(\mathcal{L})U_t, \quad U_t \sim RB(0, \sigma^2)$$

Sustituyendo este modelo en la ecuación general de Y_t , se obtiene el modelo completo para la serie observada:

$$\Phi_P(\mathcal{L}^s) \phi_p(\mathcal{L}) \Delta^d \Delta_s^D Y_t = \theta_q(\mathcal{L}) \Theta_Q(\mathcal{L}^s) U_t \quad (23)$$

Donde $\phi_p(\mathcal{L})$ y $\theta_q(\mathcal{L})$ son los polinomios autorregresivos y medias móviles de la parte regular y d es el orden de integración de la parte regular; mientras que Φ_P y $\Theta_Q(\mathcal{L})$ son los polinomios autorregresivos y medias móviles de la parte estacional y D es el orden de integración de la parte estacional.

Los *modelos ARIMA estacionales multiplicativos*, $ARIMA(p, d, q) \times (P, D, Q)_s$, son más flexivos en el sentido que especifican estacionalidades estocásticas, tendencias estocásticas y además recogen la posible interacción entre ambos componentes. Esta clase de modelos, se basa en la hipótesis central de que la relación de dependencia estacional es la misma para todos los períodos⁵⁰.

⁵⁰ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

4.7. Modelos de regresión dinámica

En un modelo de regresión, una variable Y (endógena) se explica por un conjunto k de variables X (exógena, independiente o explicativa). De esta forma, si modificamos el valor de alguna de las X , también estaremos cambiando Y .

Definición 4.24. Se denomina exógena a las variables que no son afectadas por otras en el sistema.

4.7.1. Modelo autorregresivo y de retardos distribuidos La componente autorregresiva de la variable dependiente se fusionan con las variables exógenas mediante retrasos⁵¹.

Definición 4.25. El modelo ADL combina componentes autorregresivos (retrasos de la variable dependiente) y retardos distribuidos de variables exógenas. Se define como:

$$A(\mathcal{L})Y_t = \mu + B(\mathcal{L})X_t + u_t$$
$$(1 - \alpha_1\mathcal{L} - \alpha_2\mathcal{L}^2 - \dots - \alpha_r\mathcal{L}^r)Y_t = \mu + (\beta_0 + \beta_1\mathcal{L} + \beta_2\mathcal{L}^2 + \dots + \beta_s\mathcal{L}^s)X_t + u_t$$

Donde las componentes de los polinomios $A(\mathcal{L})$ y $B(\mathcal{L})$ representan:

- Componente autorregresivo $A(\mathcal{L})$:

$$A(\mathcal{L}) = 1 - \alpha_1\mathcal{L} - \alpha_2\mathcal{L}^2 - \dots - \alpha_r\mathcal{L}^r$$

⁵¹ Universitat Oberta de Catalunya (UOC): *Econometría - Módulo 5: Modelos de regresión dinámicos y multiecuacionales*. Último acceso: 19 de marzo de 2025. 2018. https://openaccess.uoc.edu/bitstream/10609/71665/1/Econometr%C3%ADa_M%C3%B3dulo%205_Modelos%20de%20regresi%C3%B3n%20din%C3%A1micos%20y%20multiecuacionales.pdf.

Donde $\alpha_1, \dots, \alpha_r$ son los coeficientes de los retardos de Y_t (orden r).

- Componente de retardos distribuidos $B(\mathcal{L})$

$$B(\mathcal{L}) = \beta_0 + \beta_1\mathcal{L} + \beta_2\mathcal{L}^2 + \dots + \beta_s\mathcal{L}^s$$

Donde β_0, \dots, β_s son los coeficientes de los retardos X_t (orden s).

Esta estructura general sustenta la creación de modelos más sofisticados a partir del ARIMA anexando variables exógenas, que además incluyen diferenciación para analizar series no estacionarias.

4.7.2. Modelo dinámico con variables exógenas Para las series $\{Y_t\}_{t \geq 0}$ que presentan patrones cíclicos o estacionarios (como las tendencias de temperatura, precipitación durante meses o años, brotes de enfermedades en determinadas épocas del año), la estacionalidad se integra en el proceso para elaborar el modelo ARIMA estacional multiplicativo con variables exógenas. Este estudio empleará este modelo para examinar y proyectar las tendencias temporales de casos confirmados de chikungunya, considerando patrones climáticos.

La formulación general del modelo es la siguiente:

$$\underbrace{\Phi_P(\mathcal{L}^s)\phi_p(\mathcal{L})}_{\text{Componente AR}} \underbrace{(1 - \mathcal{L})^d(1 - \mathcal{L}^s)^D}_{\text{Diferenciación}} Y_t = \beta_0 + \underbrace{\sum_{i=1}^k \beta_i X_{i,t}}_{\text{Variables Exógenas}} + \underbrace{\Theta_Q(\mathcal{L}^s)\theta_q(\mathcal{L})}_{\text{Componente MA}} U_t, \quad (24)$$

Donde:

- $\Phi_P(\mathcal{L}^s)$: Polinomio AR estacional de orden P

- $\phi_p(\mathcal{L})$: Polinomio AR no estacional de orden p
- $(1 - \mathcal{L})^d$: Diferenciación regular de orden d
- $(1 - \mathcal{L}^s)^D$: Diferenciación estacional de orden D
- $\Theta_Q(\mathcal{L}^s)$: Polinomio MA estacional de orden Q
- $\theta_q(\mathcal{L})$: Polinomio MA no estacional de orden q
- $X_{i,t}$: Variables exógenas ($i = 1, \dots, k$)
- U_t : Término de error (ruido blanco)

El modelo incorpora variables exógenas que facilitan la captura de influencias externas que inciden en la dinámica de la variable endógena. No obstante, es crucial considerar algunas consideraciones fundamentales⁵²:

- Importancia tanto teórica como empírica: Las variables exógenas necesitan contar con una razón clara en relación con el fenómeno que se pretende modelar. Por ejemplo, en investigaciones epidemiológicas, estudios como el de Medina et al.⁵³ han demostrado como factores climáticos podrían tener una relación directa con la difusión de enfermedades transmitidas por vectores.
- Disponibilidad temporal: Es crucial que las variables exógenas estén accesibles durante el período de proyección.

⁵² Fahad R. Alharbi y Denes Csala: *A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) Forecasting Model-Based Time Series Approach*. En: *Inventions* 7.4 (2022), pág. 94. DOI: 10.3390/inventions7040094. <https://www.mdpi.com/2411-5134/7/4/94>.

⁵³ Juliana Medina, Andrés Pérez y Laura Gómez: *Predicción de casos de dengue en Córdoba, Colombia usando modelos SARIMAX con variables climáticas*. En: *Revista Colombiana de Salud Pública* 26.1 (2024). Artículo en prensa, págs. 45-58.

- Multicolinealidad: La incorporación de diversas variables exógenas exige comprobar la falta de colinealidad entre las mismas.
- Transformaciones previas: Similar a la serie endógena, es necesario examinar si las variables exógenas son estacionarias y, de ser necesario, modificar las variables para asegurar su estacionariedad.

5. Inferencia y predicción óptima con modelos ARIMA(p,d,q)

Una vez determinado el proceso estocástico que dio lugar a la serie de tiempo, siguiendo la metodología de Box-Jenkins⁵⁴, se analizan patrones (tendencias, estacionalidad) para entender relaciones históricas. Luego, se estiman parámetros del modelo⁵⁵ y se predicen valores futuros, con opción de ajustes iterativos.

5.1. Estimación

Dado un proceso, el objetivo es cuantificar los parámetros del mismo $\theta_1, \theta_2, \dots, \theta_q, \phi_1, \phi_2, \dots, \phi_p, \sigma^2$ y, si es necesario la constante δ . Para ello, es necesario disponer de un método para identificar estos parámetros. Para ello, se emplea el método de máxima verosimilitud⁵⁶.

Definición 5.1. Para una muestra aleatoria discreta X_1, X_2, \dots, X_n con función de masa de probabilidad $P(x | \theta)$, donde θ son los parámetros del modelo, la función de verosimilitud se define como:

$$L(\theta | \mathbf{x}) = \prod_{i=1}^n P(X_i = x_i | \theta). \quad (25)$$

Esta función permite comparar la plausibilidad de distintos valores de θ dados los datos

⁵⁴ George EP Box, Gwilym M Jenkins y Gregory C Reinsel: *Time Series Analysis: Forecasting and Control*. 4th. Wiley, 2008.

⁵⁵ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

⁵⁶ George Casella y Roger L. Berger: *Statistical Inference*. 2nd. Pacific Grove, CA: Duxbury, 2002.

observados.

- **Caso discreto:** Si X_i es una variable aleatoria discreta con distribución $P(x | \theta)$ conocida salvo por los parámetros θ , entonces la probabilidad conjunta de observar la muestra $\mathbf{x} = (x_1, \dots, x_n)$ es:

$$P(X_1 = x_1, \dots, X_n = x_n | \theta) = \prod_{i=1}^n P(X_i = x_i | \theta) = \prod_{i=1}^n P(x_i | \theta).$$

Esta expresión se define como la función de verosimilitud:

$$L(\theta | \mathbf{x}) = \prod_{i=1}^n P(x_i | \theta).$$

- **Caso continuo:** Si X_i es una variable aleatoria continua con función de densidad $f(x | \theta)$, conocida salvo por los parámetros θ , entonces la densidad conjunta de la muestra es:

$$f(x_1, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta),$$

lo cual define la función de verosimilitud como:

$$L(\theta | \mathbf{x}) = \prod_{i=1}^n f(x_i | \theta).$$

Se busca determinar el valor de $\hat{\theta}$ (estimador) que se encuentre lo más cercano posible al valor real θ_0 . En función de la distribución que produjo los datos, generalmente se emplea el logaritmo de la siguiente función:

$$\hat{\ell}(\theta | x_1, x_2, \dots, x_n) = \ln L = \sum_{i=1}^n \ln f(x_i | \theta)$$

Este método estima θ_0 en términos del valor θ que maximice InL . Este es el denominado cálculo de máxima verosimilitud (MLE) de θ_0 :

$$\hat{\theta}_{MLE} = \arg \max_{\theta} \hat{\ell}(\theta | x_1, \dots, x_n), \quad \theta \in \Theta$$

Definición 5.2. ⁵⁷ Sea $\theta_0 \in \Theta$ el vector de **parámetros verdaderos** que generan el modelo teórico, y sea $\theta \in \Theta$ un vector de **parámetros candidatos** para la aproximación del modelo. La **discrepancia** entre el modelo verdadero y el modelo aproximado se define mediante una función $D : \Theta \times \Theta \rightarrow \mathbb{R}^+$ que satisface:

$$d_n(\theta_n, \theta_0) = E_0\{-2InL(\theta|Y_n)\}, \quad (26)$$

Donde E_0 denota la expectativa bajo el modelo generador, y $L(\theta|Y_n)$ la verosimilitud.

Para un conjunto dado de estimaciones de máxima verosimilitud $\hat{\theta}_n$,

$$d_n(\hat{\theta}_n, \theta_0) = E_0\{-2InL(\theta|Y_n)\}_{\theta=\hat{\theta}_n} \quad (27)$$

Es una medida útil de separación entre el modelo generador y el modelo candidato. No obstante, (27) no es posible de evaluar debido a que requiere el conocimiento de θ . Sin embargo, Akaike mostró que $-2InL(\theta|\hat{Y}_n)$ sirve como un estimador sesgado de (27), por tanto, el ajuste de sesgo

$$E_0\{E_0\{-2InL(\theta|Y_n)_{\theta=\hat{\theta}_n}\} - E_0\{-2InL(\hat{\theta}_n|Y_n)\}\} \quad (28)$$

⁵⁷ Joseph E. Cavanaugh: *Unifying Derivations for the Akaike and Corrected Akaike Information Criteria*. En: *Statistics and Probability Letters* 33.2 (1997), págs. 201-208. DOI: 10.1016/S0167-7152(96)00128-9.

Donde a menudo se puede estimar asintóticamente por el doble de la dimensión de $\hat{\theta}_n$. Por lo tanto, si dejamos representar la dimensión de $\hat{\theta}_n$, entonces, el valor esperado de

$$AIC = -2(\hat{\theta}_n|Y_n) + 2k \quad (29)$$

La función de máxima verosimilitud para ARIMA(p, q) se define de la siguiente manera:

Definición 5.3. Sean nuestras innovaciones U_t , con $t \in \{1, \dots, n\}$, provenientes de un proceso ARIMA(p, q) que siguen una distribución normal, es decir:

$$U_t \sim \mathcal{N}(0, \sigma^2).$$

Dado esto, el modelo ARIMA puede escribirse como:

$$U_t = Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \dots - \phi_p Y_{t-p} + \theta_1 U_{t-1} + \dots + \theta_q U_{t-q}.$$

donde ϕ_i son los coeficientes autorregresivos y θ_j los coeficientes de la parte de medias móviles.

Bajo el supuesto de normalidad de las innovaciones, la función de verosimilitud para una muestra de tamaño n es:

$$f(U_1, \dots, U_n) = \frac{K}{\sigma^n} \exp \left(-\frac{1}{2\sigma^2} \sum_{t=1}^n \left(Y_t - \sum_{i=1}^p \phi_i Y_{t-i} + \sum_{j=1}^q \theta_j U_{t-j} \right)^2 \right), \quad (30)$$

donde K es una constante que no depende de los parámetros.

Debemos encontrar los coeficientes $\hat{\theta}_1, \dots, \hat{\theta}_q$ y σ^2 que maximicen la función de verosimilitud $f(U_1, \dots, U_t)$ (30) y con dichos valores calculamos el criterio de información

$$AIC = -2 \ln f(U_1, \dots, U_t) + 2k \quad (31)$$

Definición 5.4. ⁵⁸ La prueba de McLeod-Li se utiliza para detectar heteroscedasticidad condicional en los residuos de un modelo de series temporales, como un modelo ARIMA. Consiste en aplicar la prueba de Ljung-Box a los residuos al cuadrado del modelo, con el fin de identificar autocorrelación en su varianza.

H_0 : No existe heteroscedasticidad condicional.

H_a : Existe heteroscedasticidad condicional.

Ejemplo 5.5. El modelo ARIMA(1,1,1) se representa como:

$$(1 - \phi_1 \mathcal{L})(1 - \mathcal{L})y_t = (1 + \theta_1 \mathcal{L})U_t,$$

lo que implica:

$$\Delta y_t - \phi_1 \Delta y_{t-1} = U_t + \theta_1 U_{t-1},$$

y por tanto:

$$U_t = \Delta y_t - \phi_1 \Delta y_{t-1} - \theta_1 U_{t-1}.$$

⁵⁸ David S. Stoffer: *TSA: Time Series Analysis*. R package version 1.3. 2023. <https://cran.r-project.org/web/packages/TSA/index.html>.

La función de verosimilitud bajo $U_t \sim \mathcal{N}(0, \sigma^2)$ es:

$$L(\phi_1, \theta_1, \sigma^2 | \mathbf{y}) = \prod_{t=2}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{U_t^2}{2\sigma^2}\right),$$

y la log-verosimilitud:

$$\ell(\phi_1, \theta_1, \sigma^2 | \mathbf{y}) = -\frac{n-1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=2}^n U_t^2.$$

Este valor está vinculado directamente con los U_t , los cuáles dependen de los parámetros ϕ_1 y θ_1 . Por otro lado, para obtener los valores óptimos $\hat{\phi}_1$, $\hat{\theta}_1$ y $\hat{\sigma}^2$, se maximiza la función de log-verosimilitud con respecto a los parámetros.

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{t=2}^n \hat{U}_t^2$$

Finalmente, se puede calcular el criterio de información de Akaike para comparar modelos:

$$AIC = -2\ell(\hat{\phi}_1, \hat{\theta}_1, \hat{\sigma}^2 | \mathbf{y}) + 2k$$

Donde $k = 3$ es el número total de parámetros estimados.

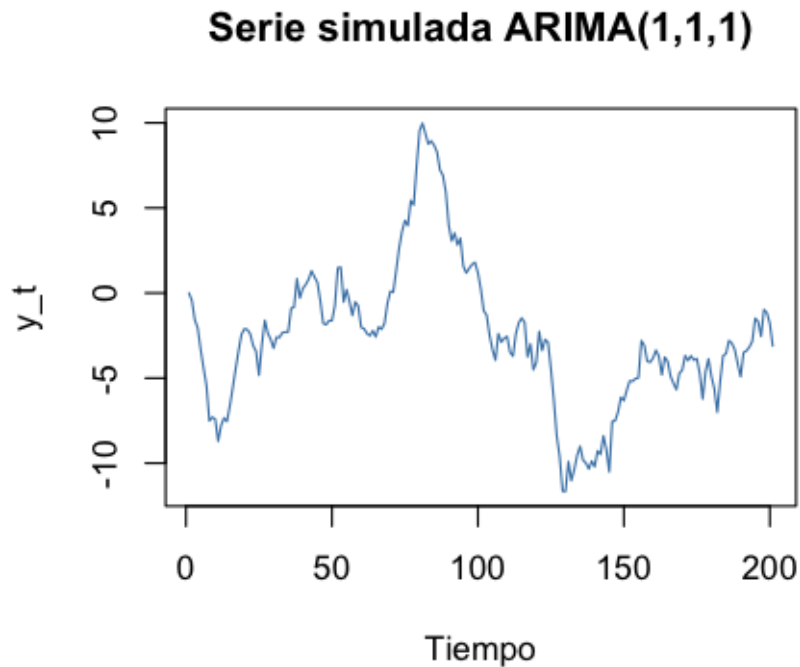


Figura 1. Serie simulada ARIMA(1,1,1)

5.2. Validación del modelo

Se debe asegurar que las innovaciones siguen una distribución normal. Dado que en este contexto la estimación por máxima verosimilitud supone normalidad. Por tanto, se introduce la prueba de Shapiro-Wilk como el método para verificar la suposición⁵⁹:

Definición 5.6. La prueba de Shapiro-Wilk es una prueba estadística utilizada para evaluar si una muestra proviene de una distribución normal.

$$H_0 : \text{Los datos siguen una distribución normal.}$$

⁵⁹ S. Shapiro y M. B. Wilk: *An analysis of variance test for normality*. En: *Biometrika* 52.3 (1965), págs. 591-611.

H_a : Los datos no siguen una distribución normal.

Si el valor p es menor que el nivel de significancia, se rechaza la hipótesis nula, indicando que los datos no siguen una distribución normal.

5.2.1. Análisis de coeficientes estimados En el caso más general de un modelo ARMA(p, q) con constante, se plantean los siguientes contrastes de hipótesis para evaluar la significancia individual de los parámetros del modelo:

$$\begin{aligned} H_0 : \delta = 0 \quad \text{frente a} \quad H_a : \delta \neq 0 \\ H_0 : \phi_i = 0 \quad \text{frente a} \quad H_a : \phi_i \neq 0 \\ H_0 : \theta_i = 0 \quad \text{frente a} \quad H_a : \theta_i \neq 0 \end{aligned} \tag{32}$$

Dado que los estimadores se obtienen por máxima verosimilitud, su distribución asintótica es normal:

$$\hat{\beta}_i \overset{a}{\sim} \mathcal{N}(\beta_i, \text{Var}(\hat{\beta}_i)), \quad \forall i,$$

donde la varianza asintótica se estima mediante la matriz de información de Fisher observada (evaluada en los valores estimados). Específicamente,

$$\text{Var}(\hat{\beta}_i) = \left[\mathcal{I}^{-1}(\hat{\beta}) \right]_{ii},$$

donde $\mathcal{I}(\hat{\beta})$ denota la matriz de información observada:

$$\mathcal{I}(\hat{\beta}) = - \left. \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta'} \right|_{\beta=\hat{\beta}}.$$

Para contrastar la hipótesis nula de que un parámetro individual es cero, se emplea el estadístico tipo t (que asintóticamente sigue una normal estándar):

$$t = \frac{\widehat{\beta}_i}{\sqrt{\widehat{\text{Var}}(\widehat{\beta}_i)}} \stackrel{a}{\sim} \mathcal{N}(0, 1)$$

Se rechazará la hipótesis nula al nivel de significancia $\alpha = 5\%$ si:

$$\left| \frac{\widehat{\beta}_i}{\sqrt{\widehat{\text{Var}}(\widehat{\beta}_i)}} \right| > z_{1-\alpha/2} \approx 1,96$$

Además, es crucial verificar que el modelo propuesto cumple con las condiciones de estacionariedad e invertibilidad. Para ello:

- Se calculan las raíces del polinomio autorregresivo estimado $\widehat{\phi}(L) = 0$.
- Se calculan las raíces del polinomio de medias móviles $\widehat{\theta}(L) = 0$.

El modelo es estacionario si todas las raíces del polinomio AR están fuera del círculo unitario (es decir, tienen módulo mayor que 1), e invertible si las raíces del polinomio MA también están fuera del círculo unitario. La presencia de raíces cercanas a la unidad puede indicar violaciones a estos supuestos.

5.2.2. Análisis de residuos Si el modelo ARMA(p,q) elegido para una serie estacionaria Z_t .

$$\Phi_p(L)Z_t = \Theta_q(L)a_t \tag{33}$$

Entonces, $a_t = \frac{\Phi_p(L)}{\Theta_q(L)}Z_t$ es un proceso de ruido blanco. Los residuos del modelo estimado vienen dados por:

$$\hat{a}_t = \frac{\widehat{\Phi}_p(L)}{\widehat{\Theta}_q(L)}Z_t \quad (34)$$

Son estimaciones de a_t . El análisis de residuos consiste en una serie de contrastes de diagnóstico con el objetivo de determinar si los residuos replican el comportamiento de un ruido blanco, es decir, si su media es cero, su varianza constante y las autocorrelaciones nulas.

1. Para comprobar si la *media es cero*, se efectúa un análisis gráfico, representando los residuos a lo largo del tiempo y observando si los valores oscilan alrededor de cero.

Además, se puede llevar a cabo el siguiente contraste de hipótesis:

$$H_0 : E(a_t) = 0$$

$$H_a : E(a_t) \neq 0$$

El estadístico de contraste se distribuye bajo la hipótesis nula como sigue:

$$t = \sqrt{T} \frac{\bar{\hat{a}}}{\sqrt{C_0(\hat{a})}} \sim N(0,1) \quad (35)$$

Donde \hat{a} y $C_0(\hat{a})$ son, respectivamente, la media y varianza muestrales de los residuos. No se rechaza la hipótesis nula al nivel de significancia del 5% sí $|t| < 1,96$.

2. *Varianza constante*. Si el gráfico de los residuos de dispersión de los mismos es constante, concluiremos que la varianza a_t permanece constante.

3. Ausencia de correlación serial.

Si los residuos se comportan como un ruido blanco, los coeficientes de la FAC Y FACP muestrales deben ser prácticamente nulos para todos los retardos. Para comprobarlo, se puede llevar a cabo:

- Contraste de significatividad *individual* sobre los coeficientes de autocorrelación:

$$H_0 : \rho_k(a) = 0$$

$$H_a : \rho_k(a) \neq 0$$

Bajo la hipótesis nula, se tiene que:

$$\hat{\rho}_k(a) \sim N\left(0, \frac{1}{T^*}\right)$$

Se dirá que a_t es un ruido blanco si los coeficientes de autocorrelación estimados están dentro del intervalo de no significación, es decir, si

$$|\hat{\rho}_k(a)| \leq \frac{2}{\sqrt{T^*}}$$

Para todo k o al menos para el 95 % de los coeficientes estimados.

4. Contraste de normalidad. El estadístico más utilizado es el de Jarque-Bera, que se define:

$$JB = \frac{T^* - p - q}{6} \left(\text{Asimetría}^2 + \frac{1}{4}(\text{Kurtosis} - 3)^2 \right)$$

Y que bajo la hipótesis nula de normalidad se distribuye asintóticamente como $\chi^2(2)$. Se rechaza la hipótesis nula de normalidad al nivel de significación del 5 % sí $JB > 5,99$.

Ejemplo 5.7. Se simuló una serie estacionaria AR(2) con la forma:

$$Y_t = 0,5Y_{t-1} - 0,3Y_{t-2} + U_t, \quad U_t \sim N(0, 1)$$

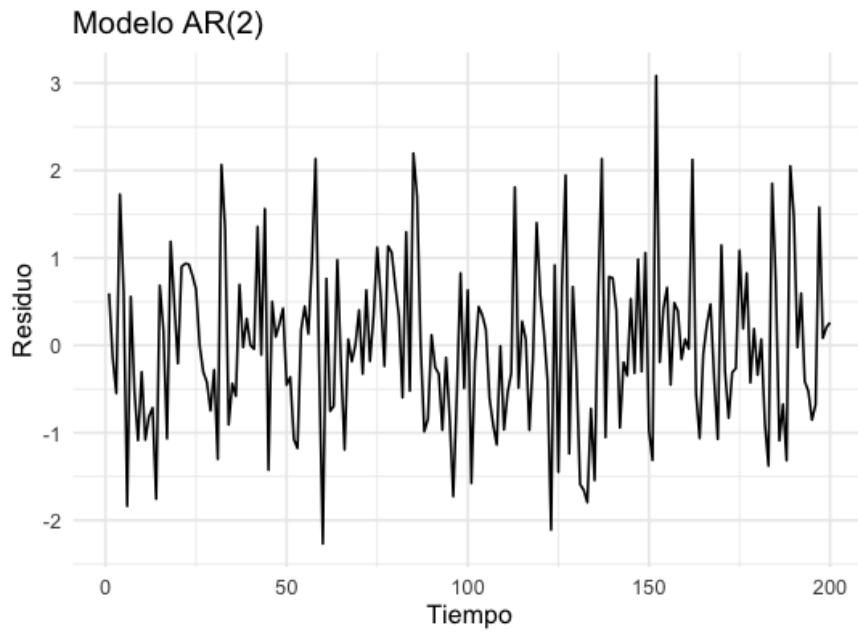


Figura 2. Residuos del modelo AR(2)

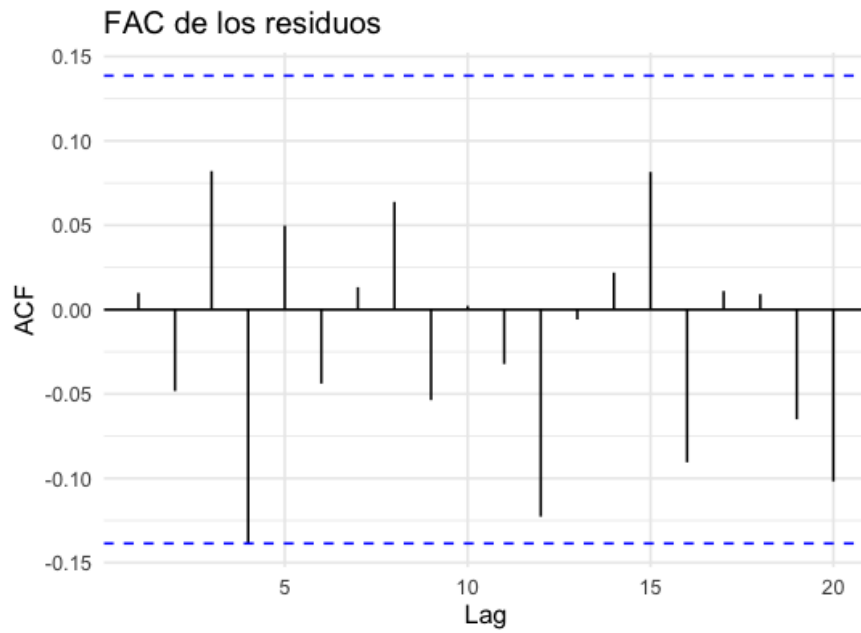


Figura 3. FAC de los residuos AR(2)

Se puede observar que el gráfico de los residuos oscila en torno a cero con varianza bastante homogénea. En lo que se refiere al correlograma de los residuos, los coeficientes de autocorrelación simple estimados se encuentran dentro de las bandas de no significación.

Estadístico	Valor
Media	-0.00024
Desviación estándar	0.94266
Mínimo	-2.26942
1er Cuartil (Q1)	-0.59789
Mediana	-0.06873
3er Cuartil (Q3)	0.60313
Máximo	3.08413

Tabla 1. Estadísticos principales para AR(2).

El test de Ljung-Box aplicado a los residuos del modelo AR(2) arrojó un valor de $\chi^2 = 8.267$ con $df = 10$ y un valor-p de 0.6028, por lo que no se rechaza la hipótesis nula de independencia serial de los residuos.

El test de Jarque-Bera aplicado a los residuos del modelo AR(2) resultó en un estadístico de $\chi^2 = 3.553$ con $df = 2$ y un valor-p de 0.1692, lo cual indica que no se rechaza la hipótesis nula de normalidad.

5.3. Predicción con modelos estacionarios

La estrategia de predicción se basa en escribir el valor que se desea predecir, X_{T+l} , tal y como se genera en función del modelo, para luego obtener la predicción óptima calculando la esperanza condicionada al conjunto de información⁶⁰. Para la

⁶⁰ José Ramón González-Pereira y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

representación medias móviles general, $X_{t+\ell}$ viene dado por:

$$X_{T+\ell} = a_{T+\ell} + \psi_1 a_{T+\ell-1} + \psi_2 a_{T+\ell-2} + \cdots + \psi_{\ell-1} a_{T+1} + \psi_{\ell} a_T + \psi_{\ell+1} a_{T-1} + \psi_{\ell+2} a_{T-2} + \cdots$$

Tomando la esperanza condicionada al conjunto de información, se obtiene:

$$X_T(\ell) = E_T[X_{T+\ell}] = \psi_{\ell} a_T + \psi_{\ell+1} a_{T-1} + \psi_{\ell+2} a_{T-2} + \psi_{\ell+3} a_{T-3} + \cdots$$

Dado que:

$$E_T(a_{T+j}) = \begin{cases} a_{T+j} & j \leq 0 \\ E(a_{T+j}) = 0 & j > 0 \end{cases}$$

La perturbación a_t es la innovación en el momento t . Sí, dado el conjunto de información I_T , se conoce el verdadero valor de X_t , como la parte sistemática se puede predecir mediante el modelo, la perturbación $a_t = X_t - \widehat{PS}_t$ está determinada, es fija. Si dado I_T , no se conoce el verdadero valor de X_t , entonces la innovación a_t no está determinada por el conjunto de información, con lo que su media condicionada será la misma que su media no condicionada, es decir, cero.

Los errores de predicción son:

$$\begin{aligned}
e_T(1) &= X_{T+1} - X_T(1) = a_{T+1} + \psi_1 a_T + \psi_2 a_{T-1} + \psi_3 a_{T-2} + \cdots \\
&\quad - (\psi_1 a_T + \psi_2 a_{T-1} + \psi_3 a_{T-2} + \cdots) = a_{T+1} \\
e_T(2) &= X_{T+2} - X_T(2) = a_{T+2} + \psi_1 a_{T+1} + \psi_2 a_T + \psi_3 a_{T-1} + \cdots \\
&\quad - (\psi_2 a_T + \psi_3 a_{T-1} + \psi_4 a_{T-2} + \cdots) = a_{T+2} + \psi_1 a_{T+1} \\
&\quad \vdots \\
e_T(\ell) &= X_{T+\ell} - X_T(\ell) = a_{T+\ell} + \psi_1 a_{T+\ell-1} + \psi_2 a_{T+\ell-2} + \cdots + \psi_{\ell-1} a_{T+1} \\
&\quad + \psi_\ell a_T + \psi_{\ell+1} a_{T-1} + \psi_{\ell+2} a_{T-2} + \cdots \\
&\quad - (\psi_\ell a_T + \psi_{\ell+1} a_{T-1} + \psi_{\ell+2} a_{T-2} + \psi_{\ell+3} a_{T-3} + \cdots) \\
&= a_{T+\ell} + \psi_1 a_{T+\ell-1} + \psi_2 a_{T+\ell-2} + \cdots + \psi_{\ell-1} a_{T+1}
\end{aligned}$$

Los errores de predicción son una combinación lineal de las perturbaciones futuras $a_{T+\ell}$, $\ell = 1, 2, \dots$, con valor medio cero:

$$E_T(e_T(\ell)) = E_T = E_T[a_{T+\ell} + \psi_1 a_{T+\ell-1} + \psi_2 a_{T+\ell-2} + \cdots + \psi_{\ell-1} a_{T+1}] = 0$$

La varianza del error de predicción o error cuadrático medio de predicción viene dado por:

$$\begin{aligned}
V_T(e_T(1)) &= E_T[e_T(1)]^2 = E_T[a_{T+1}]^2 = \sigma^2 \\
V_T(e_T(2)) &= E_T[e_T(2)]^2 = E_T[a_{T+2} + \psi_1 a_{T+1}]^2 = (1 + \psi_1)^2 \sigma^2 \\
&\quad \vdots \\
V_T(e_T(\ell)) &= E_T[e_T(\ell)]^2 = E_T[a_{T+\ell} + \psi_1 a_{T+\ell-1} + \psi_2 a_{T+\ell-2} + \cdots + \psi_{\ell-1} a_{T+1}]^2 \\
&= (1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{\ell-1}^2) \sigma^2 = \sigma^2 \sum_{i=0}^{\ell-1} \psi_i^2
\end{aligned}$$

Como se puede observar, la varianza del error de predicción va creciendo conforme nos alejamos en el futuro. Ahora bien, si el proceso es estacionario, se cumple lo siguiente:

$$\sum_{i=1}^{\infty} \psi_i^2 < \infty$$

Por lo que en varianza no crece indefinitivamente, sino que tiene una cota máxima finita.

Se puede observar que la perturbación o innovación a_t y su varianza σ^2 tienen una nueva interpretación.

- $a_t = Y_t - Y_{t-1}(1)$, es el error de predicción un período hacia adelante.
- $V(a_t) = \sigma^2$, es la varianza del error de predicción un período hacia adelante.

Si el proceso de ruido blanco sigue una distribución normal, se sigue que:

$$e_T(\ell) = X_{T+\ell} - X_T(\ell) \sim N[0, V(e_T(\ell))]$$

Por lo que el intervalo de predicción de probabilidad $(1 - \alpha) \%$ es:

$$\begin{aligned} X_{T+1} &: [X_T(1) - N\frac{\alpha}{2}\sqrt{\sigma^2} \ ; \ X_T(1) + N\frac{\alpha}{2}\sqrt{\sigma^2}] \\ X_{T+2} &: [X_T(2) - N\frac{\alpha}{2}\sqrt{\sigma^2(1 + \psi_1^2)} \ ; \ X_T(2) + N\frac{\alpha}{2}\sqrt{\sigma^2(1 + \psi_1^2)}] \\ &\vdots \\ X_{T+\ell} &: \left[X_T(\ell) - N\frac{\alpha}{2}\sqrt{\sigma^2 \sum_{i=0}^{\ell-1} \psi_i^2} \ ; \ X_T(2) + N\frac{\alpha}{2}\sqrt{\sigma^2 \sum_{i=0}^{\ell-1} \psi_i^2} \right] \end{aligned}$$

5.3.1. Predicción con modelos ARMA(p,q) Consideremos un modelo ARMA(p,q) sencillo, ARMA(1,2):

$$X_t = \delta + \phi X_{t-1} + a_t - \phi_1 a_{t-1} - \phi_2 a_{t-2}, \quad a_t \sim RB(0, \sigma^2) \quad \forall t > 0 \quad (36)$$

La media de este proceso no es cero si $\delta \neq 0$:

$$E(X_t) = \frac{\delta}{1 - \phi}$$

Las predicciones por punto son:

$$X_{T+1} = \delta + \phi X_T + a_{T+1} - \theta_1 a_T - \theta_2 a_{T-1}$$

$$\begin{aligned} X_T(1) &= E_T[X_{T+1}] = E_T[\delta + \phi X_T + a_{T+1} - \theta_1 a_T - \theta_2 a_{T-1}] = \\ &= \delta + \phi X_T - \theta_1 a_T - \theta_2 a_{T-1} \end{aligned}$$

$$X_{T+2} = \delta + \phi X_{T+1} + a_{T+2} - \theta_1 a_{T+1} - \theta_2 a_T$$

$$\begin{aligned} X_T(2) &= E_T[X_{T+2}] = E_T[\delta + \phi X_{T+1} + a_{T+2} - \theta_1 a_{T+1} - \theta_2 a_T] = \\ &= \delta + \phi X_T(1) - \theta_2 a_T \end{aligned}$$

$$X_{T+3} = \delta + \phi X_{T+2} + a_{T+3} - \theta_1 a_{T+2} - \theta_2 a_{T+1}$$

$$\begin{aligned} X_T(3) &= E_T[X_{T+3}] = E_T[\delta + \phi X_{T+2} + a_{T+3} - \theta_1 a_{T+2} - \theta_2 a_{T+1}] = \\ &= \delta + \phi X_T(2) \end{aligned}$$

$$\longrightarrow X_T(\ell) = E_T[X_{T+\ell}] = \delta + \phi X_T(\ell - 1) \quad \forall \ell > 2$$

Las dos primeras predicciones se basan en la última observación X_T (parte autorregresiva) y en los últimos errores de predicción, un período hacia delante a_T y a_{T-1}

(parte de medias móviles). Para $\ell > 2$, la parte de medias móviles no aparece de forma explícita en la función de predicción, y cada predicción se va obteniendo de las anteriores siguiendo una *regla en cadena* marcada por la parte autorregresiva. Esta función se va acercando a la media del proceso conforme nos alejamos en el futuro:

$$\begin{aligned}
 X_T(3) &= \delta + \phi X_T(2) \\
 X_T(4) &= \delta + \phi X_T(3) = \delta + \phi(\delta + \phi X_T(2)) = \delta(1 + \phi) + \phi^2 X_T(2) \\
 X_T(5) &= \delta + \phi X_T(4) = \delta + \phi(\delta(1 + \phi) + \phi^2 X_T(2)) = \delta(1 + \delta + \delta^2) + \phi^3 X_T(2) \\
 &\vdots \\
 X_T(\ell) &= \delta + \phi X_T(\ell - 1) = \delta(1 + \phi + \phi^2 + \dots + \phi^{\ell-3}) + \phi^3 X_T(2)
 \end{aligned}$$

De forma que, como el modelo ARMA(2,1) es estacionario, $|\phi| < 1$ y:

$$\lim_{\ell \rightarrow \infty} X_T(\ell) = \delta \sum_{i=0}^{\ell-3} \phi^i = \frac{\delta}{1 - \phi} (= E(X_T))$$

Para obtener la varianza del error de predicción y, de esta forma, las predicciones por intervalo, se deriva de la representación de medias móviles infinitas.

$$\begin{aligned}
 (1 - \phi L)X_t &= (1 - \theta_1 L - \theta_2 L^2)a_t \\
 \longrightarrow X_t &= \frac{1 - \theta_1 L - \theta_2 L^2}{1 - \phi L} a_t = (1 + \psi_1 L + \psi_2 L^2 + \psi_3 L^3 + \dots) a_t
 \end{aligned}$$

De dónde:

$$\begin{aligned}
 \frac{1 - \theta_1 L - \theta_2 L^2}{1 - \phi L} &= 1 + \psi_1 L + \psi_2 L^2 + \psi_3 L^3 + \dots \\
 \longrightarrow 1 - \theta_1 L - \theta_2 L^2 &= (1 - \phi L)(1 + \psi_1 L + \psi_2 L^2 + \psi_3 L^3 + \dots)
 \end{aligned}$$

Es igualando coeficientes:

$$L - \theta_1 L = (\psi_1 - \phi)L \implies -\theta_1 = \psi_1 - \phi \implies \psi_1 = \phi - \theta_1$$

$$L^2 - \theta_2 L^2 = (\psi_2 - \phi\psi_1)L^2 \implies -\theta_2 = \psi_2 - \phi\psi_1 \implies \psi_2 = \phi\psi_1 - \theta_2 = \phi(\phi - \theta_1) - \theta_2$$

$$L^3 - \theta_3 L^3 = (\psi_3 - \phi\psi_2)L^3 \implies -\theta_3 = \psi_3 - \phi\psi_2 \implies \psi_3 = \phi\psi_2$$

...

Los pesos de la forma medias móviles infinita son

$$\psi_i = \begin{cases} k=0 & \psi_0 = 1 \\ k=1 & \psi_1 = \phi - \theta_1 \\ k=2 & \psi_2 = \phi\psi_1 - \theta_2 = \phi(\phi - \theta_1) - \theta_2 \\ k>2 & \psi_k = \phi\psi_{k-1} \end{cases}$$

Con estos pesos se pueden construir intervalos de predicción. Como el proceso ARMA(1,2) es estacionario, la amplitud de los intervalos irá creciendo conforme nos alejamos en el futuro, pero con una cota máxima dada por $\left[\pm N^{\frac{\alpha}{2}} \times \sqrt{V(X_T)} \right]$.

5.4. Predicción con modelos no estacionarios

La predicción con modelos no estacionarios ARIMA(p,d,q) se lleva a cabo de la misma manera que con modelos estacionarios ARMA(p,q). El predictor por punto óptimo de $X_{T+\ell}$ viene dado por la esperanza condicionada al conjunto de información $X_T(\ell) = E_T[X_{X_{T+\ell}}]$. Para obtener las esperanzas condicionadas, sabiendo que:

$$E_T[X_{T+j}] = \begin{cases} X_{T+j} & j \leq 0 \\ X_T(j) & j > 0 \end{cases}$$

$$E_T[a_{T+j}] = \begin{cases} a_{T+j} & j \leq 0 \\ 0 & j > 0 \end{cases}$$

Para construir los intervalos de predicción,

$$\left[X_{T+\ell} \pm N \frac{\alpha}{2} \sqrt{V e_t(\ell)} \right] \quad \text{donde} \quad V(e_t(\ell)) = \sigma^2 \sum_{j=0}^{\ell-1} \psi_j^2$$

El modelo ha de estar escrito en forma $MA(\infty)$, ya que ψ_j son los pesos del modelo ARIMA escrito en forma medias móviles.

Ejemplo 5.8. Modelo ARIMA(0,1,1).

Consideremos que la serie X_t ha sido generada por el siguiente modelo:

$$(1 - L)X_t = (1 + \theta L)a_t$$

$$X_t = X_{t-1} + a_t + \theta a_{t-1}$$

La función de predicción es:

$$\begin{aligned}
X_T(1) &= E_T[X_{T+1}] = x_T + \theta a_T \\
X_T(2) &= E_T[X_{T+2}] = X_T(1) = X_T + \theta a_T \\
X_T(3) &= E_T[X_{T+3}] = X_T(2) = X_T(1) = X_T + \theta a_T \\
&\quad \vdots \quad \quad \quad \vdots \\
X_T(\ell) &= E_T[X_{T+\ell}] = X_T(\ell - 1) = X_T + \theta a_t \quad \ell > 0
\end{aligned}$$

En consecuencia, la función de predicción se centra en la predicción de un período hacia delante, $X_T(1)$, que depende del conjunto de información a través de X_T y a_T y del parámetro del modelo θ , que se encuentra conforme con ℓ creciendo.

En general, se puede demostrar que si la serie ha sido generada por un proceso integrado de orden 1, de forma que se puede representar mediante un modelo ARIMA(p,1,q) sin término independiente ($\delta = 0$), la función de predicción cuando $\ell \rightarrow \infty$, tiende a ser una constante:

$$X_T(\ell) \rightarrow K^T$$

K^T no es la media de todo el proceso (como no es estacionario, el proceso puede seguir una tendencia), sino que K^T es una constante que depende del conjunto de información y de los parámetros AR y MA.

Cuando el modelo termina independiente, $\theta \neq 0$ entonces la función de predicción está dada por:

$$E_T(X_{T+1}) = E(\delta + X_T + a_{T+1} + \theta a_T) = \delta + X_T + \theta a_T$$

$$E_T(X_{T+2}) = E(\theta + X_{T+1} + a_{T+2} + \theta a_{T+1}) = \theta + E_T(X_{T+1}) = 2\delta + X_T + \theta a_T$$

⋮

$$E(X_{T+\ell}) = E(\delta + X_{T+\ell-1} + a_{T+\ell} + \theta a_{T+\ell-1}) = \delta + E_T(X_{T+\ell-1}) = \delta\ell + X_T + \theta a_T$$

La función de predicción tiende a una línea recta cuando $\ell \rightarrow \infty$,

$$X_T(\ell) \rightarrow K^T + \delta\ell$$

La función viene dada por la constante δ y el intercepto, K^T , que depende del conjunto de información y de los parámetros del modelo.

Si el proceso no es estacionario, el límite $\lim_{\ell \rightarrow \infty} V[e_T(\ell)]$ no necesariamente existe. Para calcular $V[e_T(\ell)]$ de un modelo no estacionario ARIMA(p,d,q) es necesario escribir el modelo.

6. Estudio de caso: Chikungunya en Bucaramanga

Este capítulo desarrolla un modelo predictivo basado en series temporales para analizar la dinámica de transmisión del virus chikungunya (CHIKV) en Bucaramanga, Colombia, durante el período epidémico 2014-2018. Mediante modelos de regresión dinámica, se cuantifica el impacto de variables climáticas clave (temperatura, precipitación y humedad relativa) sobre la incidencia mensual de casos sospechosos, permitiendo identificar patrones estacionales y establecer relaciones predictivas entre las condiciones meteorológicas y la ocurrencia de brotes.

6.1. Metodología ARIMA

La metodología aplicada en este estudio se basa en el uso de la función `auto.arima()` del paquete `forecast` en R, con el objetivo de identificar el modelo más adecuado para la serie temporal univariada de casos mensuales reportados de chikungunya en Bucaramanga.

El proceso se inició con la organización y transformación de la serie diaria original (2014–2018) en una serie mensual continua, mediante interpolación *spline*. A continuación, se llevó a cabo un análisis de estacionalidad utilizando el periodograma, lo que permitió identificar frecuencias dominantes relevantes para el ajuste del modelo.

Posteriormente, se incorporaron variables exógenas estandarizadas (temperatura, precipitación y humedad relativa), debido a su posible influencia en la dinámica de transmisión del virus. Estas variables fueron incluidas dentro del proceso de ajuste automático del modelo mediante `auto.arima()`, el cual selecciona la estructura ARIMA

óptima minimizando el criterio de información AIC, evaluando tanto componentes estacionales como no estacionales.

El modelo final obtenido corresponde a un modelo de regresión dinámica que integra dichas variables exógenas, como se describe en la ecuación (24). Para validar su idoneidad, se aplicaron pruebas diagnósticas a los residuos del modelo: la prueba de **Ljung-Box** para verificar independencia temporal, la prueba de **McLeod-Li** para evaluar homocedasticidad, y la prueba de **Shapiro-Wilk** para comprobar la normalidad.

A continuación, se presenta un esquema que resume el procedimiento metodológico seguido en este trabajo.

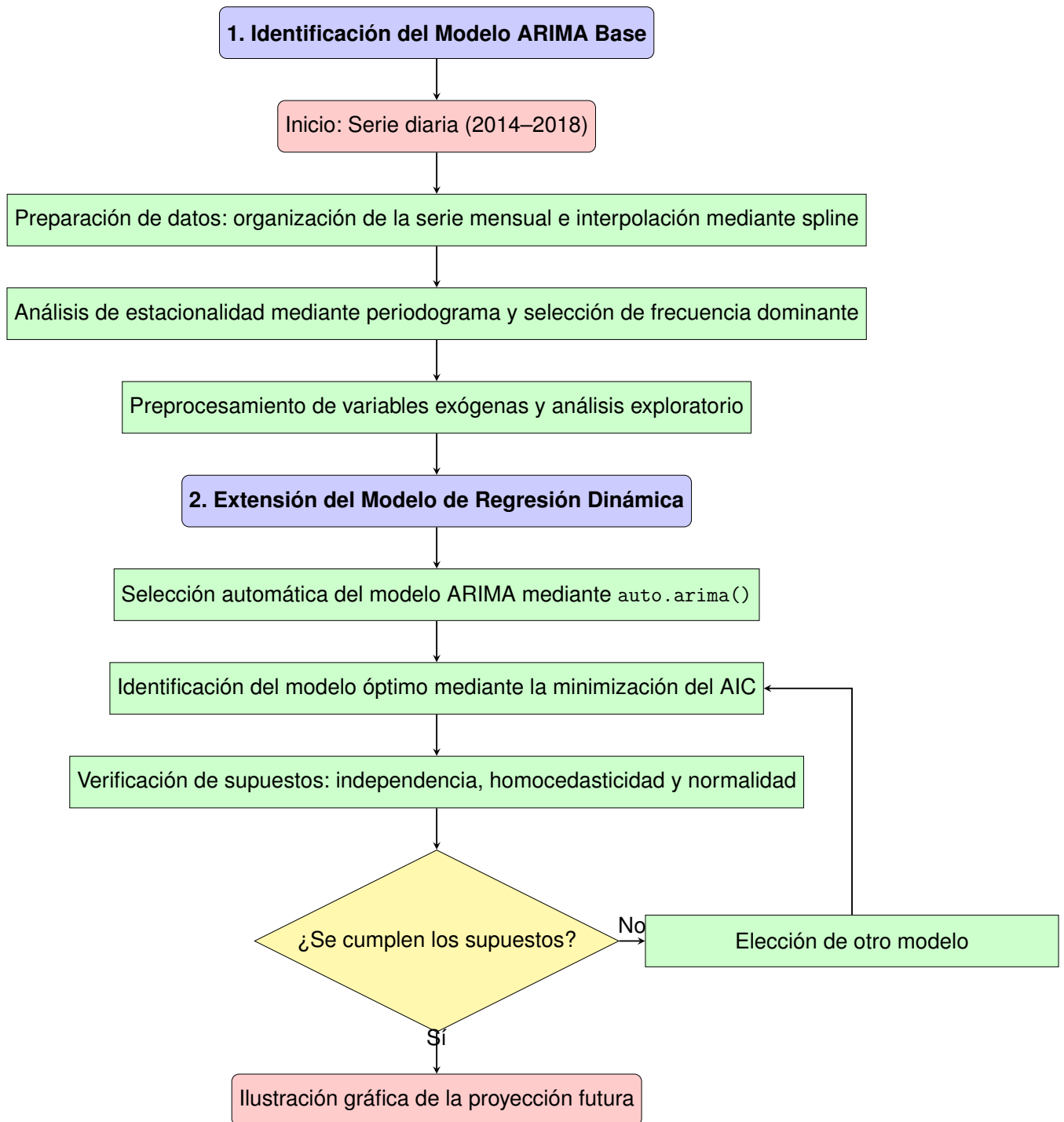


Figura 4. Proceso de modelado ARIMA y validación de supuestos para la proyección de la incidencia del virus del chikungunya.

6.2. Descripción de las bases de datos

6.2.1. Chikungunya La serie de tiempo a considerar corresponde a los casos reportados de chikungunya mensualmente en Santander, Bucaramanga, Colombia. Se incluyen como variables exógenas factores climáticos tales como la temperatura, la humedad y la precipitación, con el fin de comprender el impacto del ambiente en la dinámica de propagación del virus. En la literatura, se ha documentado ampliamente la relación entre estos factores y la proliferación del vector, el mosquito *Aedes aegypti*, autores como Tjaden et al.⁶¹ modelaron el impacto climático en la expansión de las zonas propicias para la transmisión del virus chikungunya, mostrando además cómo influyen significativamente en la eficiencia de transmisión a nivel local. Su modelo identificó cinco variables bioclimáticas determinantes: temperatura media anual, temperatura mínima del mes más frío, temperatura media del trimestre más lluvioso, temperatura media del trimestre más cálido, precipitación anual.

De igual forma, Mordecai et al.⁶² desarrollaron un modelo dinámico de transmisión de enfermedades para el virus del dengue y los mosquitos aedes que integra respuestas térmicas. El modelo implementado permitió identificar que las temperaturas intermedias (24–25°C) al inicio del brote favorecen epidemias más extensas. Además, exhibieron que tanto las ciudades tropicales como ciertas urbes con gran variabilidad térmica pueden tener una gran capacidad para la propagación del virus.

⁶¹ Nils B. Tjaden et al.: *Modelling the effects of global climate change on Chikungunya transmission in the 21st century*. En: *Scientific Reports* 7.1 (2017), pág. 3813. DOI: 10.1038/s41598-017-03566-3. <https://www.nature.com/articles/s41598-017-03566-3>.

⁶² Erin A. Mordecai et al.: *Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models*. En: *PLOS Neglected Tropical Diseases* 11.10 (2017), e0006451. DOI: 10.1371/journal.pntd.0006451. <https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0006451>.

Para el diagnóstico del virus del chikungunya se utilizan dos tipos de pruebas: RT-PCR y serología⁶³. La obtención de muestras en la etapa aguda de la enfermedad, es decir, entre los primeros cinco días desde la aparición de síntomas, deben ser examinadas a través de técnicas moleculares como RT-PCR. Por el contrario, aquellas recolectadas durante la fase de convalecencia (Período de recuperación de una enfermedad, de una intervención quirúrgica o de una lesión), a partir del sexto día de inicio de síntomas, deben ser analizadas mediante métodos serológicos. Ante la sospecha de casos fatales, se puede intentar la identificación del virus por RT-PCR en las muestras de tejidos disponibles⁶⁴.

La base de datos de chikungunya en Colombia se encuentra disponible en la página web del Instituto Nacional de Salud⁶⁵. En el portal **SIVIGILA 4.0**⁶⁶, se pueden consultar tableros que muestran el comportamiento histórico del virus, junto con la base de datos anonimizada (microdatos), la cual ha sido depurada anualmente desde 2014, dicho proceso se realiza de manera sistemática, el cual incluye primero la identificación y

⁶³ Organización Panamericana de la Salud (OPS): *Preparación y respuesta ante la introducción del virus chikungunya en las Américas*. Inf. téc. Disponible en: <https://www.paho.org>. Organización Panamericana de la Salud, Diciembre de 2011. <https://www.paho.org/sites/default/files/2024-02/preparacion-respuesta-introduccion-virus-chikungunya-americas-2011.pdf> visitado feb. de 2024.

⁶⁴ Instituto Nacional de Salud de Colombia: *Protocolo de Vigilancia en Salud Pública: Chikungunya 2024*. 2024. https://www.ins.gov.co/buscador-eventos/Lineamientos/Pro_Chikungunya%202024.pdf.

⁶⁵ Instituto Nacional de Salud de Colombia: *Portal SIVIGILA - Sistema de Vigilancia en Salud Pública*. <https://portalsivigila.ins.gov.co/>.

⁶⁶ <https://portalsivigila.ins.gov.co/>

eliminación de registros duplicados, además de la verificación y coherencia de los datos⁶⁷.

Se especifican y clarifican formalmente las siguientes variables:

- *FEC_NOT*, fecha de notificación del caso epidemiológico.
- *SEMANA*, semana epidemiológica en la que ocurrió el evento.
- *EDAD*, edad de la persona en el momento de la toma de la prueba.
- *SEXO*, masculino (M) o femenino (F).
- *FEC_CON*, fecha de consulta médica.
- *INI_SIN*, es la fecha estimada en la que la persona comenzó a presentar síntomas como dolor articular, fiebre, dolor muscular, dolor de cabeza, entre otros.
- *confirmados*, confirmado por clínica o laboratorio (1) y sospechoso (0).
- *nom_est_f_caso*, clasificación final del caso: sospechoso, confirmado por criterio clínico o por laboratorio.
- *Nom_updg*, nombre de la Unidad Primaria Generadora de Datos, es decir, nombre de la institución que reportó el caso.

Los datos a modelar corresponden al conteo mensual de la **Fecha de inicio de síntomas**, que se trata del número de casos mensuales notificados por mes.

⁶⁷ Instituto Nacional de Salud (INS): *Manual del Usuario SIVIGILA 4.0*. Disponible en: <https://www.ins.gov.co/BibliotecaDigital/manual-sivigila.pdf>. Instituto Nacional de Salud de Colombia. 2023.

6.2.2. Variables climáticas Los datos mensuales de temperatura, precipitación y humedad relativa fueron obtenidos de **NASA Langley Research Center (LaRC) POWER Project**⁶⁸⁶⁹. Estos datos meteorológicos, con una resolución espacial de $0,5 \times 0,625$, corresponden a las coordenadas $7,1389^\circ$ de latitud y $-73,1195^\circ$ de longitud, ubicadas en Bucaramanga, Santander, Colombia. Los datos se obtuvieron en el mismo período de tiempo que los datos de incidencia de la enfermedad (2014-2018).

6.2.2.1. Metodología de los datos Los valores de **humedad relativa (HR)** En el archivo POWER se calculan a partir de la presión (Pa en kPa), la temperatura de bulbo seco (T_a en $^\circ C$) y la razón de mezcla. son obtenidos de una rejilla reconfigurada y representan estimaciones a dos metros sobre la superficie local, promediadas según la resolución espacial de la fuente original.

Los valores de **precipitación** se extraen de un modelo que combina observaciones satelitales y simulaciones. Para confirmar su exactitud, se contrastan con registros de estaciones meteorológicas terrestres del NCEI⁷⁰.

Asimismo, la **temperatura** integra datos de observaciones terrestres, marítimos, aéreos y satelitales; brinda estimaciones a dos metros sobre la superficie local.

⁶⁸ <https://power.larc.nasa.gov/data-access-viewer/>

⁶⁹ NASA Prediction of Worldwide Energy Resources (POWER): *Nasa POWER Data Access Viewer*. <https://power.larc.nasa.gov/>.

⁷⁰ **National Centers for Environmental Information (NCEI)**..

6.3. Análisis de la dinámica del Chikungunya en Bucaramanga

Como se mencionó anteriormente, los casos mensuales de chikungunya se obtuvieron del Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA) del Instituto Nacional de Salud de Colombia para el período de septiembre de 2014 a diciembre de 2018⁷¹. Durante este período de estudio (2014-2018) se presentó un registro constante de casos, junto con las epidemias de Zika y dengue⁷².

Originalmente, la tabla presenta el número de casos diarios notificados del evento epidemiológico en el período comprendido entre el 5 de septiembre de 2014 y el 7 de diciembre de 2018. La ausencia de reporte en determinadas fechas no necesariamente indica “cero casos”, ya que pueden existir retrasos en la notificación o fallos en el sistema de vigilancia (registro). Por lo tanto, asumir que no hubo casos en esas fechas sería un error interpretativo.

Para abordar ese problema, se aplicó una interpolación spline con el fin de obtener una serie de tiempo regular. Este método utiliza polinomios de grado bajo para aproximar los datos, preservando la tendencia general de la serie y minimizando el error asociado a la estimación.

⁷¹ Instituto Nacional de Salud de Colombia: *Portal SIVIGILA - Sistema de Vigilancia en Salud Pública*. <https://portalsivigila.ins.gov.co/>.

⁷² Alejandro Rico-Mendoza et al.: *Co-circulation of dengue, chikungunya, and Zika viruses in Colombia from 2008 to 2018*. En: *Revista Panamericana de Salud Pública* 43 (2019), e49. DOI: 10.26633/RPSP.2019.49. <https://iris.paho.org/handle/10665.2/51077>.

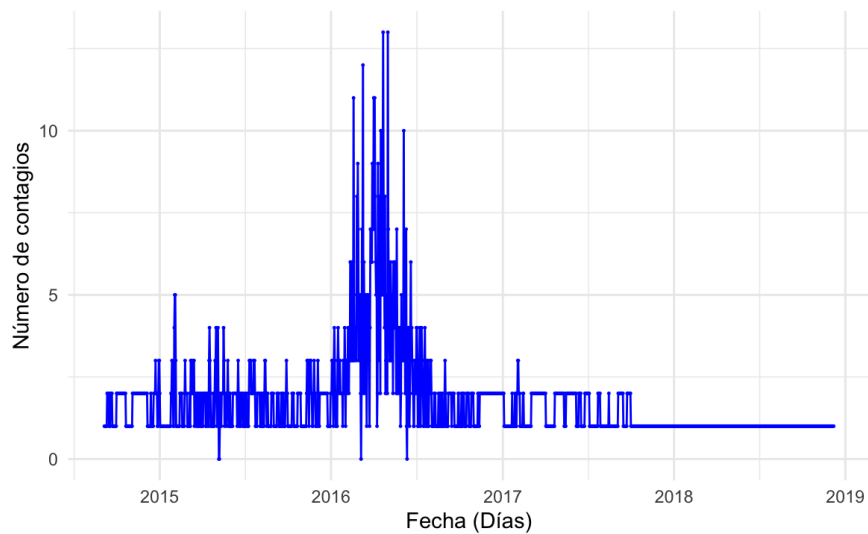


Figura 5. Casos diarios de chikungunya

Debido al comportamiento de la gráfica (5), se observa una alta variabilidad diaria y picos de notificación centrados en períodos específicos, es conveniente agrupar los datos por meses. Esta variabilidad puede atribuirse a factores como la menor asistencia a servicios médicos durante fines de semana o festivos, así como a la necesidad de atención especializada en casos extremos, lo que podría retrasar la notificación. Dado el contexto se dificulta identificar tendencias claras.

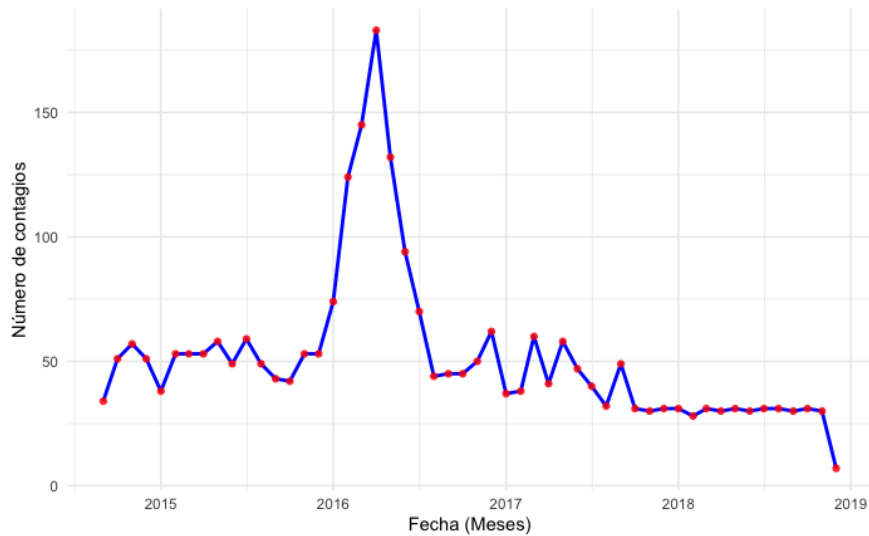


Figura 6. Casos mensuales de chikungunya

Para el análisis de series de tiempo, se utilizaron diversas funciones de las librerías de R:

- **TSA**⁷³, funciones como `periodogram()` para encontrar la frecuencia.
- **Forecast**⁷⁴, para ajustes automáticos y selección de parámetros `auto.arima`, `Box.test()` prueba de independencia de residuos y `forecast()` para generar predicciones para modelos ajustados.
- **Stats**⁷⁵, `shapiro.test()` prueba de normalidad de Shapiro Wilk y `ts()` para creación de objetos series de tiempo.

⁷³ David S. Stoffer: *TSA: Time Series Analysis*. R package version 1.3. 2023. <https://cran.r-project.org/web/packages/TSA/index.html>.

⁷⁴ Rob J Hyndman y Mitchell O'Hara-Wild: *forecast: Forecasting Functions for Time Series and Linear Models*. R package version 8.22. 2024. <https://cran.r-project.org/web/packages/forecast/index.html>.

⁷⁵ Kartikeya Bolar: *STAT: Interactive Document for Working with Basic Statistical Analysis*. R package version 0.1.0. 2019. DOI: 10.32614/CRAN.package.STAT. <https://CRAN.R-project.org/package=STAT>.

- **Zoo**⁷⁶, la función `na.spline` para interpolación de datos usando el método de spline.

Para modelar la serie temporal, se empleó la función `auto.arima`, la cual identifica de forma automática la mejor combinación de parámetros ARIMA, optimizando criterios como AIC.

Como parte del análisis exploratorio previo al ajuste del modelo, se construyó el periodograma, esta herramienta ilustra la distribución de la variabilidad de la serie según la frecuencia. Esto permite identificar los ciclos o patrones periódicos predominantes en los datos.

En este sentido, la frecuencia f se define como la fracción de un ciclo completo que ocurre en una unidad de tiempo y está dado por la fórmula:

$$f = \frac{1}{T}$$

⁷⁶ Achim Zeileis y Gabor Grothendieck: *zoo: S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations)*. R package version 1.8-12. 2024. <https://cran.r-project.org/web/packages/zoo/index.html>.

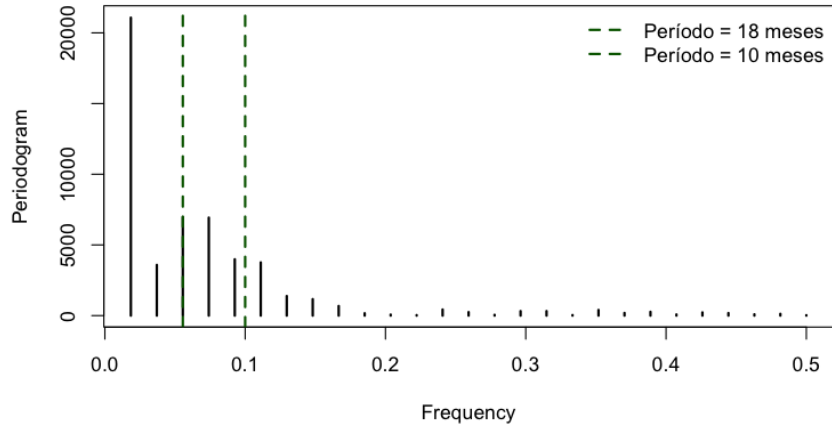


Figura 7. Periodograma.

Frecuencia	Espectro
0.01851852	21073.422
0.05555556	6966.209
0.07407407	6929.996
0.09259259	3983.397
0.11111111	3762.661

Tabla 2. Valores de frecuencia y espectro.

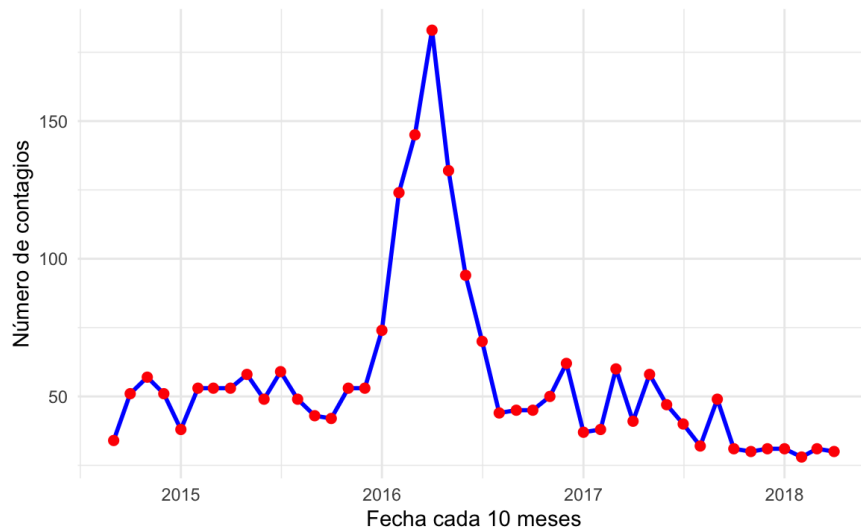


Figura 8. Serie temporal de casos de chikungunya con período de 10 meses

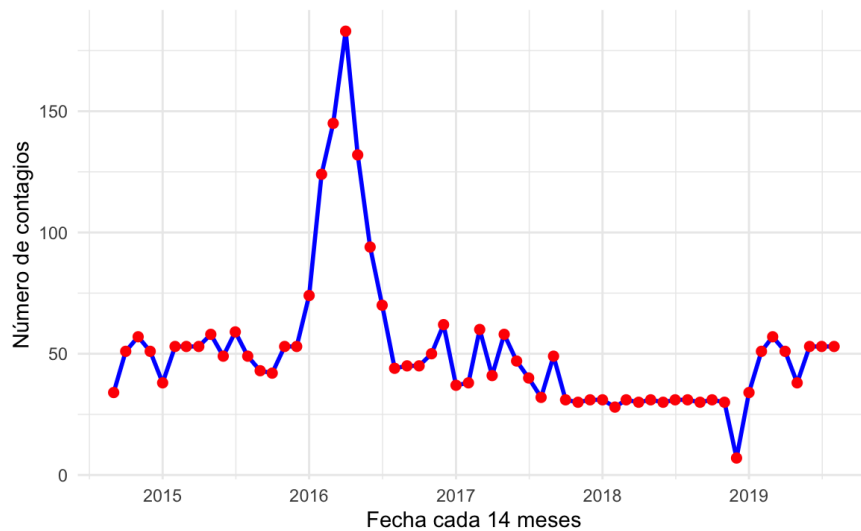


Figura 9. Serie temporal de casos de chikungunya con período de 14 meses

A partir de los resultados mostrados en la figura (7) y cuadro anterior (2), se identificaron múltiples frecuencias en los datos, con períodos significativos de 18, 14 y 10 meses. Sin embargo, estos no se incluyeron en el análisis final para evitar ruido en la identificación de patrones predominantes. En su lugar, se optó por un período de 12 meses, alineado

con la evidencia científica que demuestra que este período refleja de forma consistente los ciclos climáticos y epidemiológicos anuales en el país.

Desde una perspectiva climática, el comportamiento estacional de Colombia está dominado por la influencia de la zona de convergencia intertropical (ZCIT), una franja de baja presión atmosférica donde convergen los vientos alisios del noreste (hemisferio norte) y del sureste (hemisferio sur). Este fenómeno regula los patrones de precipitaciones en el país, aunque con variaciones regionales debido a factores como la altitud, la topografía y fenómenos macroclimáticos como El Niño-Oscilación del Sur (ENOS)⁷⁷. Esta dinámica atmosférica también condiciona la ecología de los vectores, modulando su capacidad reproductiva y distribución⁷⁸.

En el contexto epidemiológico, diversos estudios han demostrado como los factores climáticos influyen en la dinámica de transmisión de enfermedades como el chikungunya, dengue y el Zika. Por ejemplo, Ortega et. al.⁷⁹, identificaron que aumentos en la temperatura promedio en el Valle del Cauca, Colombia, se asocian con un incremento de hasta el 40% en el riesgo de dengue 0 a 2 meses después. Asimismo, altos niveles de precipitación y humedad conducían a un mayor riesgo de transmisión. De manera

⁷⁷ Asociación Nacional de Empresarios de Colombia (ANDI): *Variabilidad y cambio climático en Colombia: Impactos y desafíos*. Consultado el 23 de marzo de 2025. 2023. <https://www.andi.com.co/uploads/variabilidad.pdf>.

⁷⁸ Erin A. Mordecai et al.: *Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models*. En: *PLOS Neglected Tropical Diseases* 11.10 (2017), e0006451. DOI: 10.1371/journal.pntd.0006451. <https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0006451>.

⁷⁹ Aritra Sinha et al.: *Dengue dynamics and El Niño Southern Oscillation: A modeling study in Valle del Cauca, Colombia*. En: *PLOS Neglected Tropical Diseases* 18.1 (2024). PMID: 39378213, e0011856. DOI: 10.1371/journal.pntd.0011856. <https://pubmed.ncbi.nlm.nih.gov/39378213/>.

similar, Kabir et. al.⁸⁰ señalan que estas variables modulan la abundancia del vector y la propagación del virus del chikungunya, particularmente en aquellas regiones con climas tropicales húmedos como es el 60 % del territorio colombiano.

En ese sentido, la elección de un período de 12 meses constituye la unidad temporal óptima para estudiar la relación entre las variables climáticas y la incidencia de la enfermedad. Esta ventana temporal permite integrar los factores climáticos predominantes y su sincronía con los ciclos epidemiológicos.

Para describir el comportamiento climático de la región en cuestión, se elaboraron las series de tiempo mensuales de temperatura, precipitación y humedad relativa, basándose en los datos disponibles durante el mismo período de tiempo. Estas series fueron construidas mediante la función $ts()$ del software R, especificando una frecuencia de 12 observaciones por año, que corresponden a los registros mensuales. Con base en estas series, se aplicó una descomposición aditiva que permitió separar las componentes principales: tendencia, patrón estacional y fluctuaciones aleatorias. De esta manera, se puede proporcionar una explicación minuciosa del comportamiento climático y su potencial relación con la dinámica de la enfermedad.

⁸⁰ Mohammad Jahidul Kabir y Viroj Wiwanitkit: *Climatic factors influence dengue epidemiology: A viewpoint*. En: *Atmosphere* 12.7 (2021), pág. 905. DOI: 10.3390/atmos12070905. <https://www.mdpi.com/2073-4433/12/7/905>.

Datos climáticos e incidencia del Chikungunya en Bucaramanga					
Clasificación climática: Cálido semiárido					
Elevación: 959 msnm					
	Año				
	2014	2015	2016	2017	2018
Precipitación anual (mm)	1195.86	1021.83	1241.90	1550.40	1625.55
Temperatura media (°C)	20.19	20.88	20.70	19.66	19.67
Humedad relativa media (%)	75.05	72.22	75.07	80.18	78.84
Población (habitantes)	527451	527913	528269	528497	528610
Incidencia de enfermedad (por 100000 hab.)	36.59	114.22	202.16	93.47	64.50

Tabla 3. Datos climáticos e incidencia del Chikungunya en Bucaramanga

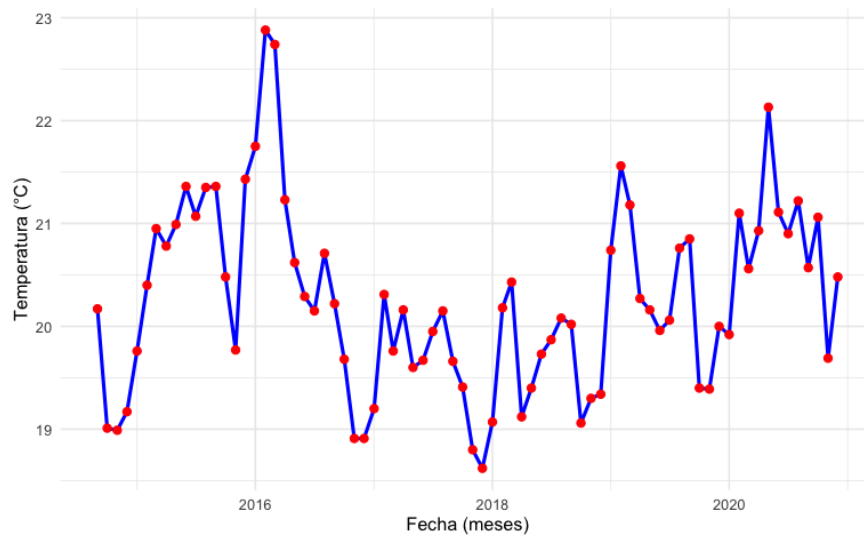


Figura 10. Temperatura promedio en la región de estudio

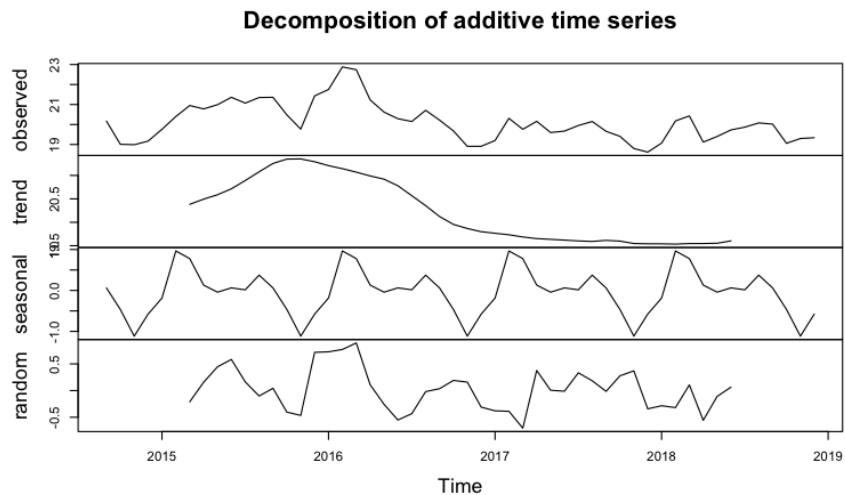


Figura 11. Descomposición de la serie temporal de temperatura

6.3.0.1. Temperatura La figura (11) muestra un marcado incremento térmico durante 2016, seguido de un descenso progresivo hasta 2018 y una posterior estabilización. Este patrón coincide con los efectos documentados del fenómeno ENOS (El Niño-Oscilación del Sur), en particular con el evento de El Niño 2014-2016, uno de los más intensos registrados. Como demuestra Peña et al.⁸¹, estas anomalías climáticas no solo alteran las condiciones atmosféricas, sino que impactan directamente en la salud pública, modificando la dinámica de vectores como el *Aedes aegypti* y, consecuentemente, la incidencia de enfermedades como el dengue.

Además, el estudio muestra un componente estacional evidente con fluctuaciones regulares y periódicas, mostrando ciclos predecibles en la variabilidad de la temperatura.

⁸¹ Ximena Peña y otros: *Impacto del fenómeno El Niño-Oscilación del Sur sobre las condiciones climáticas y su relación con enfermedades transmitidas por vectores en Colombia*. En: *Revista Colombiana de Geografía* 15.1 (2001). Disponible en: <https://revistas.unal.edu.co/index.php/rcg/article/view/41090>, págs. 47-60.

Cambios climáticos rítmicos, combinados con los extremos asociados a ENOS, crean escenarios epidemiológicos diferenciados donde las variables climáticas (particularmente la temperatura) modulan la transmisión del dengue de forma regionalmente específica.

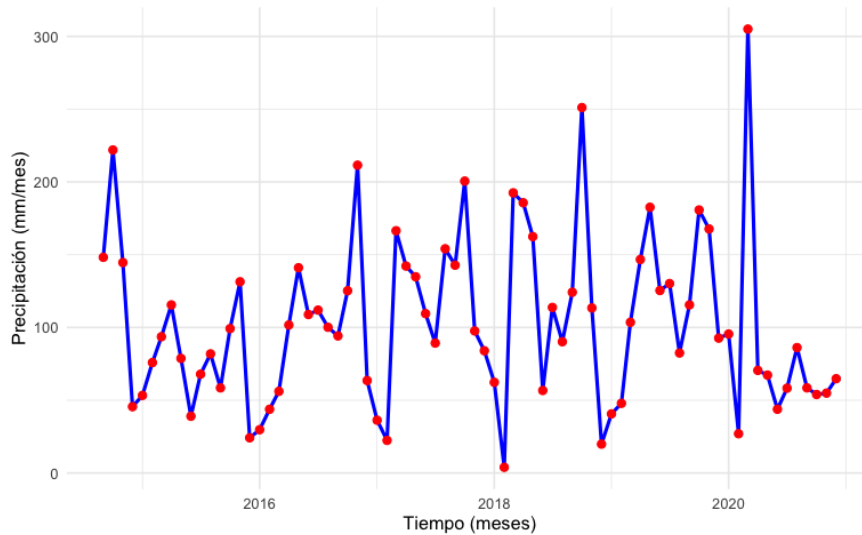


Figura 12. Precipitación promedio en la región de estudio

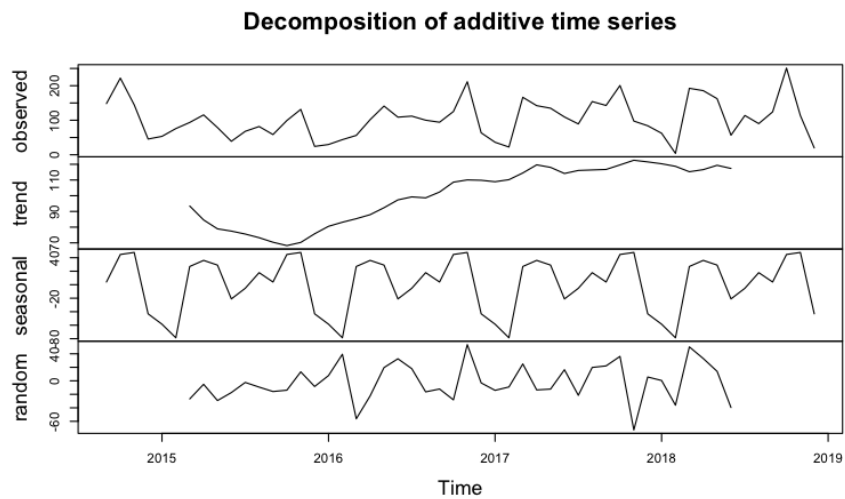


Figura 13. Descomposición de la serie temporal de precipitación

6.3.0.2. Precipitación Según la figura (13), la precipitación experimentó un incremento progresivo a partir de 2015, que continuó hasta 2017, momento en que los valores tienden a estabilizarse. La descomposición de la serie temporal revela que la componente estacional muestra oscilaciones periódicas que se repiten de manera regular, con un patrón claramente anual.

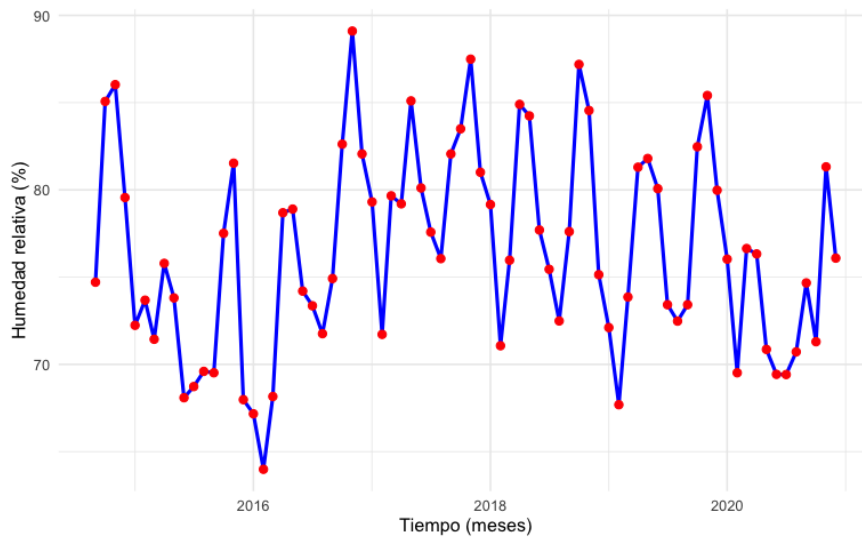


Figura 14. Humedad relativa promedio en la región de estudio

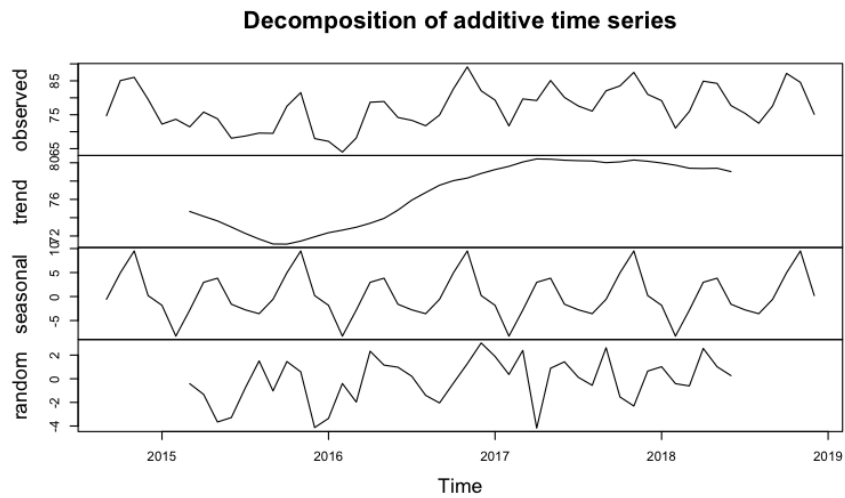


Figura 15. Descomposición de la serie temporal de humedad relativa

6.3.0.3. Humedad relativa Finalmente, la descomposición de la serie de tiempo de la humedad relativa (15), sugiere una tendencia creciente desde 2016 hasta mediados de 2017, seguida de una fase de estabilización. Asimismo, la componente estacional muestra oscilaciones periódicas que se repiten a lo largo del tiempo, lo que evidencia un comportamiento cíclico característico de este tipo de variable climática.

El análisis de las gráficas anteriores (11, 13, 15) revela diversas variaciones significativas asociadas a eventos climáticos globales, como El Niño 2014-2016. A continuación, se resumen los patrones observados.

- **Temperatura:** Aumento pronunciado durante 2016, seguido de descenso progresivo y estabilización hacia 2018, con una marcada componente estacional.
- **Precipitación:** Incremento sostenido entre 2015 y 2017, con posterior estabilización.

- **Humedad relativa:** Tendencia creciente hasta mediados de 2017, seguida de estabilidad, con oscilaciones periódicas.

Estas variaciones climáticas evidencian el marcado efecto de fenómenos climáticos globales sobre las condiciones locales, con repercusiones potenciales en la ecología de vectores como el *Aedes aegypti*. La detección sistemática de las características estacionales y tendenciales en las variables climáticas ofrecen un marco de referencia para el desarrollo modelos dinámicos de riesgo epidemiológico asociado a arbovirus.

De acuerdo con el esquema metodológico presentado en la figura, se utilizó la función `auto.arima()` del paquete `forecast` de R para identificar automáticamente los órdenes $(p, d, q) \times (P, D, Q)_{12}$ de un modelo ARIMA estacional. Este algoritmo evalúa exhaustivamente distintas configuraciones mediante criterios de información (AIC), seleccionando aquella con mejor ajuste.

Para la incorporación de variables exógenas, se construyó una matriz a partir de las series climáticas (temperatura, precipitación y humedad relativa). Con el objetivo de asegurar la correspondencia temporal con la serie de casos mensuales, se eliminaron los últimos 24 registros de dicha matriz. Luego, las variables fueron normalizadas mediante escalamiento estándar (`scale()`), lo que permitió mejorar la comparabilidad y estabilidad numérica del modelo.

Por último, se utilizó la función `auto.arima()` incluyendo la matriz de variables exógenas a través del argumento `xreg`. Como resultado, fue posible establecer el modelo ARIMA que mejor representa el vínculo entre las condiciones climáticas y la evolución temporal de casos de chikungunya.

```

# Matriz de variables exógenas
#Combinación de variables exógenas: precipitación, humedad y temperatura
var_exogenas = cbind(PP_data$PP, Hum_data$Humedad_valor, Temp_data$Temperatura_valor)

# Verificar dimensiones
cat("Longitud de la serie:", length(serie_mensual), "\n")
cat("Filas de xreg antes de recorte:", nrow(var_exogenas), "\n")

# Eliminar los últimos 24 registros
var_exogenas = var_exogenas[1:(nrow(var_exogenas) - 24), ]

# Asegurar que las dimensiones coincidan
if (length(serie_mensual) != nrow(var_exogenas)) {
  stop("Error: La serie de tiempo y las variables exógenas tienen diferente longitud")
}

# Normalizar las variables exógenas
var_exogenas = scale(var_exogenas)

# Modelo ARIMA con variables exógenas
modelo_arima = auto.arima(serie_mensual,
                          seasonal = TRUE,
                          stepwise = FALSE,
                          approximation = FALSE,
                          trace = TRUE,
                          xreg = var_exogenas)

# Resumen del modelo
summary(modelo_arima)

```

Modelo ARIMA	Error	AIC	BIC
ARIMA(1,0,2)(0,0,1)[12]	441.3607	437.07	454.64
ARIMA(1,0,2)(1,0,0)[12]	441.362	437.08	454.64
ARIMA(1,0,0)(0,0,1)[12]	441.6485	439.1	452.76
ARIMA(1,0,0)(1,0,0)[12]	441.6638	439.12	452.78

Tabla 4. Mejores modelos ARIMA

El modelo $ARIMA(1, 0, 2) \times (0, 0, 1)_{12}$ fue seleccionado mediante criterios de información (AIC) y validación de residuos. Los términos AR(1) y MA(2) capturan tendencias temporales, el término autoregresivo refleja una influencia significativa y positiva de los valores pasados, mientras que los términos de medias móviles proponen un impacto moderado.

Parámetro	Valor	Error estándar
ar1	0.7488	0.1070
ma1	0.1152	0.1419
ma2	0.4985	0.1800
sma1	0.0180	0.1652
Intercepto	49.0163	11.4059
xreg1 (Precipitación estandarizada)	-1.0377	2.3774
xreg2 (Humedad relativa estandarizada)	14.7158	4.8459
xreg3 (Temperatura estandarizada)	10.9587	6.3925

Tabla 5. Coeficientes del modelo ARIMA con variables exógenas.

A pesar de que el error estándar sugiere un cierto grado de incertidumbre en la estimación del coeficiente asociado a la variable exógena temperatura, este parámetro mantiene su relevancia estadística dentro del modelo, como se evidencia en el resumen del ajuste (summary).

Métrica	Valor
Varianza de errores (σ^2)	210.5
Log-likelihood	-209.54
AIC	437.07
AICc	441.36
BIC	454.64

Tabla 6. Indicadores de ajuste del modelo.

6.4. Validación del modelo

Además de emplear el criterio de Akaike (AIC) para la selección del modelo, se realizó una validación rigurosa mediante un diagnóstico completo, considerando un nivel de significancia de $\alpha = 0,05$. Este análisis verificó el cumplimiento de supuestos esenciales:

- **Independencia:** Mediante la prueba de Ljung-Box se evidenció la ausencia de autocorrelación significativa, garantizando que la estructura temporal fue adecuadamente modelada.
- **Homocedasticidad:** La prueba de McLeod-Li sobre los residuos al cuadrado descartó heterocedasticidad condicional.
- **Normalidad:** La distribución de los residuos se validó con pruebas estadísticas Shapiro-Wilk y métodos gráficos (ver anexos para el método gráfico).

Con base en lo anterior, la evaluación de independencia de los residuos se realizó mediante la prueba estadística de Ljung-Box el cual mostró un valor p de 0.3625. Este resultado indica que no existe evidencia suficiente para afirmar la presencia de autocorrelación en los residuos, lo cual es favorable, ya que demuestra que el modelo ha capturado adecuadamente los patrones temporales de la serie y que los residuos se comportan de manera independiente.

Para la evaluación de homocedasticidad, la prueba de McLeod-Li mostró valores significativos en algunos rezagos iniciales, lo que sugiere episodios puntuales de heterocedasticidad condicional. Sin embargo, la inconsistencia de este patrón indica que no se trata de un fenómeno persistente. Por tanto, a pesar de que el modelo podría mostrar cierta variabilidad no uniforme en momentos concretos, no hay evidencia de un problema generalizado de heterocedasticidad.

En términos generales, el modelo presenta homocedasticidad, aunque se observaron indicios puntuales de posible heterocedasticidad en algunos rezagos específicos. No obstante, la ausencia de un patrón sistemático sugiere que la varianza de los residuos se mantiene relativamente constante a lo largo del tiempo.

Por otro lado, respecto al supuesto de normalidad, la prueba de Shapiro-Wilk arrojó un valor-p de 0.1259, lo cual indica que no se rechaza la hipótesis nula de normalidad. En consecuencia, se concluye que los residuos presentan un comportamiento compatible con una distribución normal, cumpliendo así con el último supuesto requerido para la validación del modelo.

En conjunto, los resultados de las pruebas estadísticas aplicadas para la validación del modelo confirman el cumplimiento de los supuestos fundamentales. Específicamente, la prueba de Ljung-Box para independencia, McLeod-Li para homocedasticidad y Shapiro-Wilk para normalidad coinciden en un mismo veredicto: los residuos del modelo se comportan de manera adecuada para confiar en la validez de las inferencias y predicciones realizadas.

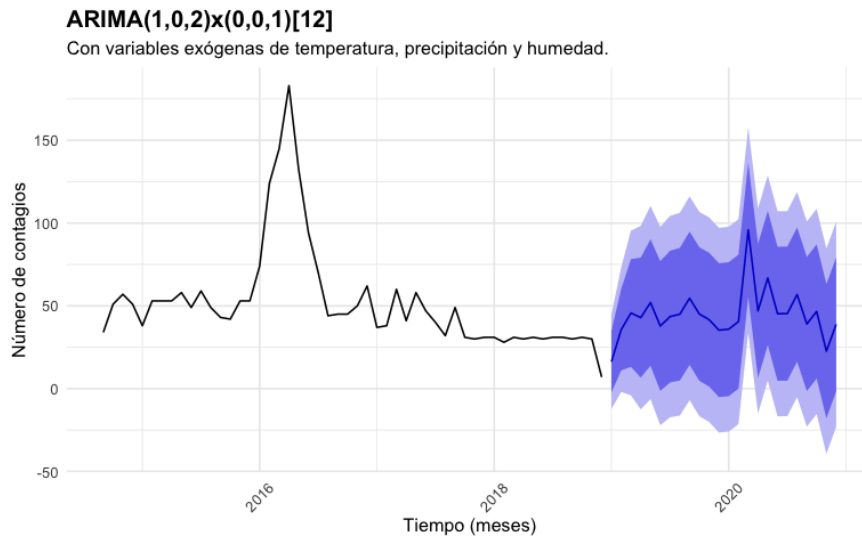


Figura 16. Gráfica del mejor modelo con mejor criterio AIC

La figura (16) compara la serie histórica con las predicciones generadas por el modelo $ARIMA(1, 0, 2) \times (0, 0, 1)_{12}$, el cual incorpora variables climáticas (temperatura, precipitación y humedad). El salto observado entre la serie y la predicción puede atribuirse a múltiples causas: por un lado, marca el punto donde terminan los datos reales y comienzan las estimaciones; por otro, podría indicar cambios estructurales entre el patrón histórico y las dinámicas futuras del modelo que intenta capturar. Más allá de este punto, los intervalos de confianza representados como bandas laterales se ensanchan de manera gradual, evidenciando el aumento natural de la incertidumbre en pronósticos a largo plazo.

La predicción sugiere que, en el corto plazo, no se anticipan incrementos abruptos en el número de casos. Desde una perspectiva epidemiológica, la interacción entre las variables climáticas y la serie temporal analizada permitió identificar patrones relevantes para la comprensión y gestión de la enfermedad. En particular, las condiciones climáticas posteriores al periodo observado no parecen favorecer la reproducción y actividad del

mosquito vector, lo cual podría explicar un escenario de relativa estabilidad o “calma” en la incidencia proyectada.

Es importante resaltar que el período de predicción coincide con la emergencia sanitaria por el virus SARS-CoV-2 (COVID-19), lo cual pudo haber influido significativamente la dinámica de notificación de casos de chikungunya. Las restricciones de movilidad, la reorientación de recursos del sistema de salud hacia la atención de la creciente pandemia y la disminución de actividades de vigilancia, probablemente condujeron a una subnotificación de casos durante ese tiempo.

Este contexto extraordinario sugiere que, aunque el modelo tenga un buen ajuste entre los datos históricos y las variables exógenas incluidas, las proyecciones deben interpretarse con cautela, considerando las particularidades de la dinámica epidemiológica durante ese periodo.

Finalmente, en este contexto de epidemias concurrentes (dengue, Zika y chikungunya) y una pandemia emergente, la coyuntura epidemiológica del chikungunya revela una compleja interacción de factores climáticos, sociales, económico. Estas predicciones se consideran como el umbral mínimo que debería considerarse para la planificación de estrategias de prevención y control.

7. Conclusiones

En esta investigación se analizó la dinámica de los casos confirmados de chikungunya (CHIKV) en Bucaramanga, Santander, mediante un enfoque de series de tiempo con regresión dinámica. El modelo seleccionado, un $ARIMA(1, 0, 2) \times (0, 0, 1)_{12}$ con variables exógenas (temperatura, precipitación y humedad relativa), permitió predecir la evolución de la serie temporal y evaluar la influencia de factores climáticos en la transmisión del virus.

Los resultados evidenciaron que las variables meteorológicas contribuyen significativamente a mejorar la capacidad predictiva del modelo, sugiriendo que el comportamiento del virus se ve parcialmente modulado por las condiciones ambientales. Sin embargo, la persistencia de transmisión también refleja la interacción con otros determinantes no climáticos.

Estos hallazgos son coherentes con estudios previos que relacionan la incidencia de enfermedades transmitidas por vectores con variaciones climáticas, las cuales afectan la reproducción y supervivencia del mosquito vector. Se recomienda ampliar futuras investigaciones, incluyendo otras variables climatológicas, así como fenómenos a gran escala como El Niño y la Niña, y factores sociales como la movilidad poblacional, con el fin de comprender integralmente la dinámica de transmisión.

Los resultados de la validación confirmaron el cumplimiento de los supuestos del modelo. La prueba de Ljung-Box no evidenció autocorrelación significativa en los residuos; la prueba de McLeod-Li concluyó homocedasticidad general; y la prueba de

Shapiro-Wilk respaldó la hipótesis de normalidad. El salto observado entre los datos reales y la predicción marca el inicio del período estimado y puede deberse tanto al cambio estructural en la dinámica como a cambios externos no capturados, como fue la pandemia de COVID-19.

Una de las principales limitaciones del estudio es la posible subnotificación de casos, particularmente en contextos de alta presión sobre el sistema de salud, como durante la pandemia de COVID-19. Adicionalmente, la cocirculación de otros virus transmitidos por el mismo vector, como dengue y Zika, pudo generar confusión diagnóstica o priorización selectiva en la atención y el reporte, afectando la precisión de los registros disponibles. Estas condiciones deben tenerse en cuenta al interpretar los resultados del modelo. Se recomienda que futuras investigaciones incorporen estrategias de ajuste por subregistro, utilicen fuentes de información complementarias (como registros hospitalarios o bases de datos independientes) y consideren explícitamente el contexto de cocirculación viral para mejorar la validez y robustez de las estimaciones.

En conjunto, el modelo propuesto permite capturar la influencia de factores climáticos sobre la dinámica de transmisión del chikungunya. Esto resulta fundamental para la previsión de casos, facilitando el diseño de estrategias preventivas, particularmente aquellas orientadas al control del mosquito vector. La inclusión de variables climáticas mejoraron la precisión de las proyecciones, cumpliendo con el objetivo de identificar y cuantificar el papel del entorno climático en la propagación del virus.

BIBLIOGRAFÍA

Alharbi, Fahad R. y Denes Csala: *A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) Forecasting Model-Based Time Series Approach*. En: *Inventions* 7.4 (2022), pág. 94. DOI: 10.3390/inventions7040094. <https://www.mdpi.com/2411-5134/7/4/94>.

Banu, Shamima et al.: *Space-time clusters of dengue fever in Bangladesh*. En: *Tropical Medicine and International Health* 20.8 (2015), págs. 1086-1091. DOI: 10.1111/tmi.12521.

Bolar, Kartikeya: *STAT: Interactive Document for Working with Basic Statistical Analysis*. R package version 0.1.0. 2019. DOI: 10.32614/CRAN.package.STAT. <https://CRAN.R-project.org/package=STAT>.

Box, George EP, Gwilym M Jenkins y Gregory C Reinsel: *Time Series Analysis: Forecasting and Control*. 4th. Wiley, 2008.

Brockwell, Peter J. y Richard A. Davis: *Introduction to Time Series and Forecasting*. 3rd. Springer Texts in Statistics. Springer, 2016. ISBN: 978-3-319-29852-8. DOI: 10.1007/978-3-319-29854-2. <https://doi.org/10.1007/978-3-319-29854-2>.

Carrillo, M. A. et al.: *Risk of dengue, Zika, and chikungunya transmission in the metropolitan area of Cucuta, Colombia: cross-sectional analysis, baseline for a cluster-randomised controlled trial of a novel vector tool for water containers*. En: *BMC Public Health* 23.1 (2023). DOI: <https://doi.org/10.1186/s12889-023-15893-4>.

Casella, George y Roger L. Berger: *Statistical Inference*. 2nd. Pacific Grove, CA: Duxbury, 2002.

Catalunya (UOC), Universitat Oberta de: *Econometría - Módulo 5: Modelos de regresión dinámicos y multiecuacionales*. Último acceso: 19 de marzo de 2025. 2018. https://openaccess.uoc.edu/bitstream/10609/71665/1/Econometr%C3%ADa_M%C3%B3dulo%205_Modelos%20de%20regresi%C3%B3n%20din%C3%A1micos%20y%20multiecuacionales.pdf.

Cavanaugh, Joseph E.: *Unifying Derivations for the Akaike and Corrected Akaike Information Criteria*. En: *Statistics and Probability Letters* 33.2 (1997), págs. 201-208. DOI: 10.1016/S0167-7152(96)00128-9.

Desconocido: *Modelos ARIMA: Metodología Box-Jenkins*. Último acceso: 17 de marzo de 2025. s.f. <https://www.estadistica.net/ECONOMETRIA/SERIES-TEMPORALES/modelo-arima.pdf>.

DiCYT: *El Niño intensifica la infestación del mosquito transmisor del dengue en el estado de São Paulo*. Noticia consultada el 23 de marzo de 2025. 2024. <https://www.dicyt.com/noticias/el-nino-intensifica-la-infestacion-del-mosquito-transmisor-del-dengue-en-el-estado-de-sao-paulo>.

Empresarios de Colombia (ANDI), Asociación Nacional de: *Variabilidad y cambio climático en Colombia: Impactos y desafíos*. Consultado el 23 de marzo de 2025. 2023. <https://www.andi.com.co/uploads/variabilidad.pdf>.

González-Pereira, José Ramón y María de los Ángeles Gil Álvarez: *Predicción mediante modelos ARIMA: Una aplicación al consumo de cemento en España*. Último acceso: 18 de marzo de 2025. Universidad del País Vasco (UPV/EHU), 2009. <https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf>.

Goyal, Manu et al.: *Recent development in the strategies projected for chikungunya vaccine in humans*. En: *Drug Design, Development and Therapy* 12 (2018). Accessed:

2025-07-02, págs. 4195-4206. DOI: 10.2147/DDDT.S181574. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6292406/>.

Hyndman, Rob J y George Athanasopoulos: *Forecasting: principles and practice*. Capítulo 2: Introducción a las series temporales. OTexts, 2018. <https://otexts.com/fpp3/>.

Hyndman, Rob J y Mitchell O'Hara-Wild: *forecast: Forecasting Functions for Time Series and Linear Models*. R package version 8.22. 2024. <https://cran.r-project.org/web/packages/forecast/index.html>.

Hyndman, Rob J. y Yeasmin Khandakar: *Automatic Time Series Forecasting: The forecast Package for R*. En: *Journal of Statistical Software* 27.3 (2008), págs. 1-22. <http://www.jstatsoft.org/v27/i03/>.

IBM: *Autocorrelation and Partial Autocorrelation Functions*. Accessed: March 23, 2025. 2025. <https://www.ibm.com/docs/es/spss-modeler/saas?topic=data-autocorrelation-partial-autocorrelation-functions>.

Instituto de Estadísticas de Puerto Rico: *Análisis de Series de Tiempo: Fundamentos y Aplicaciones*. Inf. téc. Consultado el [Fecha] en [Hora]. Gobierno de Puerto Rico, 2023, pág. 4.54. http://www.estadisticas.gobierno.pr/iepr/LinkClick.aspx?fileticket=4_BxecUaZmg%3D.

Instituto Nacional de Salud (INS): *Manual del Usuario SIVIGILA 4.0*. Disponible en: <https://www.ins.gov.co/BibliotecaDigital/manual-sivigila.pdf>. Instituto Nacional de Salud de Colombia. 2023.

Kabir, Mohammad Jahidul y Viroj Wiwanitkit: *Climatic factors influence dengue epidemiology: A viewpoint*. En: *Atmosphere* 12.7 (2021), pág. 905. DOI: 10.3390/atmos12070905. <https://www.mdpi.com/2073-4433/12/7/905>.

Karasinghe, Sanjaya, Ruwan Silva y Tharaka Perera: *Short-term forecasting of dengue outbreaks using ARIMA models: A case study in Colombo, Sri Lanka*. En: *Infectious Disease Modelling* 9 (2024), págs. 15-27. DOI: 10.1016/j.idm.2024.01.002.

Lowe, Rachel et al.: *Nonlinear and delayed impacts of climate on dengue risk in Barbados: A modelling study*. En: *PLOS Medicine* 15.7 (2018), e1002613. DOI: 10.1371/journal.pmed.1002613.

Medina, Juliana, Andrés Pérez y Laura Gómez: *Predicción de casos de dengue en Córdoba, Colombia usando modelos SARIMAX con variables climáticas*. En: *Revista Colombiana de Salud Pública* 26.1 (2024). Artículo en prensa, págs. 45-58.

Mordecai, Erin A. et al.: *Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models*. En: *PLOS Neglected Tropical Diseases* 11.10 (2017), e0006451. DOI: 10.1371/journal.pntd.0006451. <https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0006451>.

Morgan, J., C. Strode y J. E. Salcedo-Sora: *Climatic and socio-economic factors supporting the co-circulation of dengue, Zika and chikungunya in three different ecosystems in Colombia*. En: *PLOS Neglected Tropical Diseases* 15.3 (2021). DOI: <https://doi.org/10.1371/journal.pntd.0009259>.

Organización Panamericana de la Salud (OPS): *Preparación y respuesta ante la introducción del virus chikungunya en las Américas*. Inf. téc. Disponible en: <https://www.paho.org>. Organización Panamericana de la Salud, Diciembre de 2011. <https://www.paho.org/sites/default/files/2024-02/preparacion-respuesta-introduccion-virus-chikungunya-americas-2011.pdf> visitado feb. de 2024.

Peña, Ximena y otros: *Impacto del fenómeno El Niño-Oscilación del Sur sobre las condiciones climáticas y su relación con enfermedades transmitidas por vectores*

en Colombia. En: *Revista Colombiana de Geografía* 15.1 (2001). Disponible en: <https://revistas.unal.edu.co/index.php/rcg/article/view/41090>, págs. 47-60.

Poveda, Germán, Diana M. Álvarez y Óscar A. Rueda: *Hydro-climatic variability over the Andes of Colombia associated with ENSO: a review of climatic processes and their impact on one of the Earth's most important biodiversity hotspots*. En: *Climate Dynamics* 36.11–12 (2011), págs. 2233-2249. DOI: 10.1007/s00382-010-0931-y. <https://link.springer.com/article/10.1007/s00382-010-0931-y>.

Poveda, Germán et al.: *La hidroclimatología de Colombia: una síntesis desde la escala inter-decadal hasta la escala diurna*. En: *Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales* 28.107 (2004), págs. 201-222. https://www.accefyn.com/revista/Vol_28/107/201-222.pdf.

Rico-Mendoza, Alejandro et al.: *Co-circulation of dengue, chikungunya, and Zika viruses in Colombia from 2008 to 2018*. En: *Revista Panamericana de Salud Pública* 43 (2019), e49. DOI: 10.26633/RPSP.2019.49. <https://iris.paho.org/handle/10665.2/51077>.

Rincón, Luis: *Conceptos Elementales de la Teoría Matemática de la Probabilidad*. 2007. <https://www.cimat.mx/~pabreu/LuisRinconI.pdf>.

Romero Moreno, Joaquín Santiago: *Un acercamiento a los modelos ARIMAX y lineal dinámico para estimar natalidad en Colombia en el periodo entre enero 2011 a diciembre 2017*. Especialización en Estadística Aplicada. 2019. <https://repository.libertadores.edu.co/server/api/core/bitstreams/48d35dc8-aaf0-4157-b638-b763a2a1cf81/content>.

Rueda, J. C. et al.: *Demographic and clinical characteristics of chikungunya patients from six Colombian cities, 2014–2015*. En: *Emerging Microbes and Infections* 8.1 (2019), págs. 1490-1500. DOI: <https://doi.org/10.1080/22221751.2019.1678366>.

Salomon, Joshua A. et al.: *Forecasting epidemiological transitions: multilevel modeling of time-series and contextual data for health policy planning*. En: *The Lancet* 380.9859 (2012), págs. 2063-2071. DOI: 10.1016/S0140-6736(12)61719-7.

Salud, Organización Mundial de la: *Partes sobre brotes epidémicos; Expansión geográfica de los casos de dengue y chikungunya más allá de las áreas históricas de transmisión en la Región de las Américas*. 2023. <https://www.who.int/es/emergencies/disease-outbreak-news/item/2023-DON448>.

Salud de Colombia, Instituto Nacional de: *Portal SIVIGILA - Sistema de Vigilancia en Salud Pública*. <https://portalsivigila.ins.gov.co/>.

Salud de Colombia, Instituto Nacional de: *Protocolo de Vigilancia en Salud Pública: Chikungunya 2024*. 2024. https://www.ins.gov.co/buscador-eventos/Lineamientos/Pro_Chikungunya%202024.pdf.

Shapiro, S. y M. B. Wilk: *An analysis of variance test for normality*. En: *Biometrika* 52.3 (1965), págs. 591-611.

Sinha, Aritra et al.: *Dengue dynamics and El Niño Southern Oscillation: A modeling study in Valle del Cauca, Colombia*. En: *PLOS Neglected Tropical Diseases* 18.1 (2024). PMID: 39378213, e0011856. DOI: 10.1371/journal.pntd.0011856. <https://pubmed.ncbi.nlm.nih.gov/39378213/>.

Stoffer, David S.: *TSA: Time Series Analysis*. R package version 1.3. 2023. <https://cran.r-project.org/web/packages/TSA/index.html>.

Tjaden, Nils B. et al.: *Modelling the effects of global climate change on Chikungunya transmission in the 21st century*. En: *Scientific Reports* 7.1 (2017), pág. 3813. DOI: 10.1038/s41598-017-03566-3. <https://www.nature.com/articles/s41598-017-03566-3>.

Vega, Benjamín: *Notas de Clase: Series de Tiempo*. Último acceso: 18 de marzo de 2025. 2023. <https://github.com/benjov/Series-Tiempo/blob/main/docs/Notas-Series-Tiempo.pdf>.

Walsh, Bryan: *1 Million People Have a Disease You've Never Heard Of*. Accessed: 2025-07-02. 2014. <https://time.com/3631711/chikungunya-epidemic-2014/>.

World Health Organization: *El Niño–Southern Oscillation (ENSO)*. Consultado el 23 de marzo de 2025. 2024. [https://www.who.int/es/news-room/fact-sheets/detail/el-nino-southern-oscillation-\(enso\)](https://www.who.int/es/news-room/fact-sheets/detail/el-nino-southern-oscillation-(enso)).

Worldwide Energy Resources (POWER), NASA Prediction of: *Nasa POWER Data Access Viewer*. <https://power.larc.nasa.gov/>.

Yakob, Laith: *Modeling the transmission of chikungunya in Brazil using time series methods*. En: *PLOS Neglected Tropical Diseases* 16.5 (2022), e0010472. DOI: 10.1371/journal.pntd.0010472.

Zeileis, Achim y Gabor Grothendieck: *zoo: S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations)*. R package version 1.8-12. 2024. <https://cran.r-project.org/web/packages/zoo/index.html>.