Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Físico Matemáticas

Aplicación de Técnicas de Series de Tiempo en el Estudio de la Esquizofrenia

Tesis presentada al

Colegio de Matemáticas

como requisito parcial para la obtención del grado de

Licenciado en Matemáticas Aplicadas

por

Yasmin Mauricio Muñoz

Asesorada por

Dr. Hugo Adán Cruz Suárez
Dra. Gladys Denisse Salgado Suárez

Puebla Pue.
14 de septiembre de 2025

Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Físico Matemáticas

Aplicación de Técnicas de Series de Tiempo en el Estudio de la Esquizofrenia

Tesis presentada al

Colegio de Matemáticas

como requisito parcial para la obtención del grado de

Licenciado en Matemáticas Aplicadas

por

Yasmin Mauricio Muñoz

Asesorada por

Dr. Hugo Adán Cruz Suárez
Dra. Gladys Denisse Salgado Suárez

Puebla Pue.
14 de septiembre de 2025

Título: Aplicación de Técnicas de Series de Tiempo en el Estudio de la Esquizofrenia
Estudiante: Yasmin Mauricio Muñoz

COMITÉ

Dr. Carlos Camilo Garay
Presidente

Dr. Rei Israel Ortega Gutiérrez
Secretario

Dr. Ruy Alberto López Ríos
Vocal

Dr. Hugo Adán Cruz Suárez
Dra. Gladys Denisse Salgado Suárez
Asesor

Índice general

Introducción

Las series de tiempo constituyen una herramienta estadística fundamental para analizar y predecir el comportamiento de fenómenos que evolucionan a lo largo del tiempo. Su aplicación es amplia y abarca campos como la economía, la meteorología y la ingeniería. Entre los modelos más utilizados en este tipo de análisis se encuentra el modelo $ARIMA$ (AutoRegressive Integrated Moving Average), reconocido por su capacidad para generar predicciones confiables a partir de datos históricos.

El modelo $ARIMA$ combina tres componentes: autorregresión (AR), integración o diferenciación (I) y medias móviles (MA), lo que permite modelar dependencias lineales, tendencias y estacionalidades en las series temporales. Esta tesis tiene como objetivo principal aplicar dicho modelo para analizar y predecir valores futuros de una serie relacionada con casos de esquizofrenia. A lo largo del documento se abordarán los procesos de identificación, estimación, validación y predicción de modelos $ARIMA$, así como un análisis detallado de la base de datos utilizada.

Aunque los modelos $ARIMA$ se han aplicado con frecuencia en contextos económicos y financieros, su utilidad no se limita a estos campos. Siempre que se disponga de datos temporales relevantes, es posible emplearlos en otras áreas, incluidas las ciencias de la salud. En este trabajo se ha optado por aplicar técnicas de series de tiempo al estudio de la esquizofrenia, un trastorno mental grave que afecta profundamente el pensamiento, las emociones y el comportamiento de quienes lo padecen.

En esta tesis se decidió trabajar con datos sobre casos de esquizofrenia debido a la naturaleza grave de este trastorno mental, que afecta profundamente el pensamiento, las emociones y el comportamiento de quienes lo padecen. La esquizofrenia puede generar una desconexión con la realidad, causando angustia tanto en la persona afectada como en su entorno cercano. Los síntomas, como alucinaciones, delirios y alteraciones cognitivas, dificultan el desarrollo de una vida cotidiana normal. Sin embargo, con un diagnóstico temprano y tratamiento adecuado, que incluye medicación, psicoterapia y apoyo social, muchos pacientes logran estabilizarse, lo que les permite continuar con sus actividades diarias, mantener su autonomía y establecer relaciones satisfactorias. El acceso a un tratamiento integral, que aborde tanto los síntomas como el bienestar general del paciente, es crucial para su integración plena en la sociedad.

La esquizofrenia es una enfermedad mental poco conocida y sobre la cual existen aún muchos mitos, además, muchas personas no consideran las matemáticas como una herramienta valiosa para abordar problemas sociales. De este modo, la presente tesis busca desafiar esta percepción al demostrar cómo las técnicas estadísticas, y en particular los modelos de series de tiempo, pueden contribuir al análisis de fenómenos complejos como las enfermedades mentales.

El análisis se llevará a cabo utilizando una base de datos proporcionada por el *Institute for Health Metrics and Evaluation* (IHME), basada en los resultados del Estudio sobre la Carga Glo-

bal de Enfermedades (*Global Burden of Disease*, GBD). Este conjunto de datos es reconocido a nivel mundial por su exhaustividad y precisión, y se utiliza para evaluar el impacto de diversas enfermedades, factores de riesgo y condiciones de salud en diferentes poblaciones. El análisis y la modelación se realizarán mediante la herramienta *RStudio*, donde se implementará el código necesario para generar las predicciones.

La estructura del trabajo es la siguiente: el Capítulo 1 presenta una introducción a la esquizofrenia y sus causas, incluyendo su definición, historia y desarrollo, así como conceptos fundamentales sobre su tratamiento, sin profundizar excesivamente en estos aspectos. El Capítulo 2 aborda los fundamentos teóricos de las series temporales, comenzando con los conceptos estadísticos básicos, las notaciones y definiciones necesarias para comprender adecuadamente este tema. Mientras que, el Capítulo 3 se centra en los modelos ARMA; el Capítulo 4 profundiza en el modelo $ARIMA$ y su proceso de construcción; finalmente, en el Capítulo 5 se presenta la aplicación del modelo $ARIMA$ a la base de datos referente a ezquizofrenia, y el Apéndice A contiene el código y las salidas correspondientes.

Además del objetivo técnico principal, la predicción mediante modelos $ARIMA$, esta tesis tiene como meta secundaria contribuir a la sensibilización sobre la gravedad de la esquizofrenia y su impacto humano. Este trabajo reconoce que, detrás de cada observación estadística, hay una realidad personal compleja, y por ello busca mantener un enfoque respetuoso y ético hacia las personas que conviven con este padecimiento y sus familias.

Capítulo 1

Preliminares

El presente capítulo ofrece una introducción general al trastorno de la esquizofrenia, con el propósito de contextualizar el objeto de estudio de esta tesis. La esquizofrenia es un trastorno psicótico severo que se manifiesta a través de una combinación de signos y síntomas que afectan múltiples procesos psicológicos, como la percepción (alucinaciones), la ideación, la verificación de la realidad (delirios), el pensamiento (desorganización), las emociones (afecto plano o inapropiado), la atención, la motivación y el juicio. Actualmente, no existe un síntoma único que permita diagnosticar la esquizofrenia con certeza, ya que las manifestaciones clínicas varían ampliamente entre individuos y se solapan con otros trastornos. Además, no todos los pacientes presentan la misma combinación de síntomas, lo que hace que la caracterización de la enfermedad dependa de múltiples factores [13]. A lo largo de este capítulo se revisan los elementos esenciales que definen este trastorno, incluyendo su sintomatología, las causas asociadas, las opciones de tratamiento y su impacto psicosocial. Esta revisión busca proporcionar al lector una base conceptual sólida sobre la esquizofrenia, que justifique la pertinencia de su análisis mediante técnicas estadísticas como los modelos de series de tiempo.

1.1. Síntomas de la Esquizofrenia

Según la American Psychiatry Association (2013), los síntomas de la esquizofrenia suelen clasificarse en tres categorías: **positivos, negativos** y **cognitivos**. Esta división nos permite una mejor comprensión de cómo se manifiesta la enfermedad y facilita tanto su diagnóstico como su tratamiento.

- **Síntomas positivos:** Son aquellos que implican una alteración en el funcionamiento mental normal, es decir, "añaden" experiencias que no deberían estar presentes. Estas son:
  - **Alucinaciones**, especialmente de tipo auditivo, como escuchar voces que otras personas no pueden oír. Estas voces pueden comentar las acciones del paciente, dar órdenes o mantener conversaciones entre sí.
  - **Delirios**, son creencias falsas o irracionales las cuales suelen ser mantenidas con firmeza, como creer que uno posee poderes sobrenaturales o que está siendo constantemente vigilado o perseguido.
  - **Trastornos del pensamiento**, los cuales son evidentes en un discurso desorganizado, es decir, el dar respuestas confusas o saltos ilógicos entre ideas, lo cual suele dificultar la comunicación efectiva.
- **Síntomas negativos:** Representan una disminución o pérdida de las funciones mentales normales. Algunos ejemplos de estos son:

  - **Anhedonia**, que es la incapacidad para experimentar placer en actividades que antes resultaban gratificantes.
  - **Aplanamiento afectivo**, esta se presenta como una reducción en la expresión emocional, ya sea en forma facial, gestual o verbal.
  - **Aislamiento social** es la disminución en la capacidad de establecer o mantener relaciones interpersonales.

- **Síntomas cognitivos**: Afectan el procesamiento mental de la persona, y aunque pueden ser más difíciles de detectar, tienen un gran impacto en el funcionamiento diario de las personas que los padecen. Entre ellos se encuentran:

  - **Dificultades para concentrarse**, prestar atención o mantener el enfoque en tareas simples.
  - **Déficits en la memoria de trabajo**, lo que afecta la capacidad de retener y manipular información a corto plazo.
  - **Problemas en la toma de decisiones y en el juicio**, lo que puede limitar la autonomía del individuo.

1.2. Causas de la Esquizofrenia

Según lo expuesto en [12], la esquizofrenia se considera un trastorno de origen multifactorial, lo que significa que no hay una sola causa específica que explique su aparición. En lugar de ello, se entiende que el desarrollo de la enfermedad es el resultado de la interacción entre diversos factores, los cuales son: **genéticos, neurobiológicos** y **ambientales**.

- **Factores genéticos:** La predisposición genética juega un papel importante en el desarrollo de la esquizofrenia. En algunos estudios se ha observado que los individuos que tienen un familiar cercano con el trastorno tienen un riesgo considerablemente mayor de desarrollar la enfermedad en comparación con aquellos sin antecedentes familiares. Sin embargo, la presencia de estos genes no garantiza el desarrollo de la esquizofrenia, lo cual sugiere la influencia de otros o más factores.

- **Factores neurobiológicos:** Investigaciones han revelado que las personas con esquizofrenia suelen presentar alteraciones en su estructura cerebral, especialmente en áreas relacionadas con el procesamiento de la información. Además, se ha identificado un desequilibrio en los neurotransmisores clave, como la dopamina y el glutamato, que están relacionados con los síntomas psicóticos de la enfermedad. Estos desajustes químicos contribuyen a la disfunción de las redes neuronales involucradas en el pensamiento y la percepción.

- **Factores ambientales:** El contexto en el que una persona se desarrolla también juega un papel importante en la aparición de la esquizofrenia. El estrés prenatal, el consumo de sustancias psicoactivas, así como experiencias traumáticas durante la infancia, se han identificado como factores de riesgo significativos. Estos elementos pueden desencadenar o incrementar la aparición de los primeros episodios psicóticos en individuos con una predisposición genética.

La interacción entre los factores mencionados puede variar considerablemente entre un individuo a otro, lo que dificulta aún más el diagnóstico temprano en pacientes con esquizofrenia.

1.3. Tratamiento de la Esquizofrenia

Podemos observar en [8] que el tratamiento de la esquizofrenia suele tratarse desde un enfoque más amplio, en donde, se trata de combinar distintas estrategias terapéuticas, con el objetivo de reducir los síntomas, mejorar el funcionamiento más completo del paciente y potenciar una mejor calidad de vida. Este tratamiento se basa principalmente en:

- **Medicamentos antipsicóticos:** Estos son esenciales para controlar los síntomas positivos de la enfermedad, como lo son las alucinaciones y los delirios. Se clasifican en antipsicóticos de primera y segunda generación. Estos últimos suelen ser preferidos en la práctica clínica, ya que tienden a causar menos efectos secundarios, lo cual mejora el seguimiento del tratamiento y el bienestar general de los pacientes.

  Es importante también considerar que cuando haya una mala respuesta en la medicación, se debe evaluar la consistencia de la medicación y el posible uso de otras sustancias antes de que se pueda establecer definitivamente la falta de respuesta. Si no llega a haber una respuesta a la medicación después unas 4 o 6 semanas, a pesar de la optimización de la dosis, se debe considerar un cambio en el antipsicótico.

- **Psicoterapia:** Especialmente la terapia cognitivo-conductual, ha demostrado ser más eficaz para ayudar a los pacientes a identificar y manejar pensamientos desorganizados, disminuir la angustia relacionada con los síntomas psicóticos y desarrollar habilidades de adaptación. Además, puede fomentar una mayor conciencia de la enfermedad, es decir, cuando la persona acepta que tiene un problema de salud mental y comprende que necesita ayuda médica, lo cual es clave para la continuidad del tratamiento.

- **Intervenciones psicosociales:** Incluyen programas de rehabilitación psicosocial, entrenamiento en habilidades sociales, apoyo familiar, participación laboral y servicios comunitarios. Estas intervenciones son esenciales para facilitar la integración social del paciente, mejorar su autonomía y prevenir recaídas, contribuyendo así a una vida lo más funcional e independiente posible.

1.4. Datos Estadísticos Sobre la Esquizofrenia en México

- Se calcula que entre el 1 % y el 2 % de la población adulta urbana padece esquizofrenia, lo que equivaldría aproximadamente a 500 000–700 000 personas [14].

- Un informe oficial de la Secretaría de Salud estimaba que en México hay entre 500 000 y 600 000 personas con esquizofrenia, aunque solo el 10 % recibe atención hospitalaria [21].

- En 2016, la tasa de consultas ambulatorias por esquizofrenia fue de 13.44 por cada 100 000 habitantes, una disminución del 19 % respecto al año 2000 [20].

- En el ámbito hospitalario, en 2016 se registraron 8.51 egresos hospitalarios por esquizofrenia por cada 100 000 habitantes, frente a 4.56 en el año 2000 [20].

- La prevalencia estimada para esquizofrenia fue de 202 por cada 100 000 habitantes. Se calcula que solo el 10 % de los pacientes requerirían hospitalización anual, lo que muestra una brecha significativa entre demanda y atención disponible [20].

- En 2020, la esquizofrenia ocasionó 127 muertes, equivalentes al 0.02 % del total de defunciones en México, con una tasa de mortalidad ajustada por edad de 0.10 por cada 100 000 habitantes. Esto sitúa al país en el puesto 46 a nivel mundial para esta causa [27].

- Se estima que alrededor del 85 % de las personas con esquizofrenia no pueden acceder a tratamiento adecuado [15].
- Además, solo el 19 % de los casos son realmente tratados [15].
- En cuanto a recursos humanos especializados, en 2020 México contaba con apenas 0.2 psiquiatras y 3 psicólogos por cada 100 000 habitantes, cifras muy bajas en comparación internacional [16].
- Se reporta una escasez grave de medicamentos clave como clozapina y carbonato de litio, esenciales para el tratamiento de esquizofrenia. Esta situación ha provocado recaídas, síntomas de abstinencia e incluso riesgos médicos graves [16].

1.5. Impacto Social y Calidad de Vida

La esquizofrenia tiene un impacto significativo en la vida diaria de las personas que la padecen, afectando no solo su bienestar personal, sino también su capacidad para mantener relaciones sociales y laborales. Los individuos con esquizofrenia a menudo enfrentan dificultades para interactuar en entornos sociales, lo que ocasiona que haya un aumento en el aislamiento. Este aislamiento social, combinado con el rechazo social que usualmente sufren de parte de las demás personas, puede contribuir a una mayor exclusión de la sociedad. Además, la enfermedad se asocia con una tasa elevada de desempleo, ya que las personas con esquizofrenia tienen dificultades para cumplir con los requisitos laborales debido a sus síntomas cognitivos, emocionales y comportamentales. A todo esto, también se le suman los obstáculos que tienen para acceder a una atención médica adecuada y oportuna, lo cual limita aún más sus posibilidades de recibir un tratamiento efectivo. Estos factores en conjunto no solo agravan el sufrimiento individual, sino que también tienen un impacto negativo en la calidad de vida de las personas que padecen esquizofrenia, dificultando la rehabilitación social y la integración plena en la comunidad [8].

1.6. Perspectivas Futuras

A pesar de todos los avances obtenidos a lo largo del tiempo en el estudio y la comprensión de la esquizofrenia, aún existen muchas incógnitas, especialmente en lo que respecta a las causas exactas que pueden llevar a una persona a desarrollar este trastorno. Asimismo, aunque los tratamientos actuales permiten controlar algunos de los síntomas más leves, no existe ningún tratamiento que garantice un control total de la enfermedad, ni mucho menos una cura definitiva. Además, aún no se cuenta con un conocimiento profundo sobre cómo mejorar la calidad de vida de los pacientes, quienes no solo deben enfrentarse a los síntomas de la esquizofrenia, sino también al rechazo social. Sin embargo, la investigación continúa avanzando en áreas importantes como lo son: la genética, la neurobiología y las terapias innovadoras, como la estimulación cerebral profunda, lo que ofrece nuevas esperanzas para el tratamiento y la mejora de la calidad de vida de quienes padecen esta enfermedad [1, 8].

En resumen, la esquizofrenia representa un desafío clínico y social de gran complejidad, debido a la diversidad de sus manifestaciones, la multiplicidad de factores que intervienen en su desarrollo y el impacto que tiene sobre la vida de los pacientes y su entorno. Esta complejidad justifica la necesidad de emplear herramientas analíticas que permitan comprender su evolución en el tiempo y anticipar posibles patrones en su comportamiento. En este contexto, el análisis de series de tiempo emerge como una estrategia metodológica adecuada para modelar los datos históricos relacionados con este trastorno. El siguiente capítulo introduce los fundamentos teóricos de las series temporales, sentando las bases estadísticas necesarias para su aplicación en el estudio de la esquizofrenia.

Capítulo 2

Series de Tiempo

Con el propósito de analizar la evolución temporal de los casos de esquizofrenia, es necesario introducir el marco teórico y metodológico de las series de tiempo. Este capítulo presenta los conceptos fundamentales relacionados con este tipo de datos, incluyendo su definición formal, principales componentes, tipos de modelos y áreas de aplicación. Las series de tiempo permiten estudiar fenómenos que varían a lo largo del tiempo, y proporcionan herramientas estadísticas para identificar patrones, evaluar comportamientos pasados y generar predicciones futuras. Esta base teórica será empleada para comprender los modelos que se utilizarán posteriormente en la modelación de los datos asociados a la esquizofrenia.

2.1. Definición de una Serie de Tiempo

Antes de definir una serie de tiempo, es importante considerar ciertos conceptos estadísticos fundamentales que serán necesarios posteriormente.

2.1.1. Conceptos Estadísticos Básicos

Definición 1. Un *proceso estocástico* es una colección de variables aleatorias indexadas por el tiempo, denotada por $\{y_t : t \in \mathbb{Z}\}$ y lo representamos abreviadamente como $\{y_t\}$ [26]. A cada instante t le corresponde una variable aleatoria y_t, y el análisis se enfoca en estudiar sus propiedades estadísticas. Se supone que dichas variables se enuentran definidas en un espacio común de probabilidad $(\Omega, \mathcal{F}, P)$. Para dicho proceso se definen las siguientes medidas:

1. $E(y_t) = \mu_t$, *esperanza del proceso* en el tiempo t.
2. $Var[y_t] = \gamma_{t,t}$, *varianza* en el tiempo t.
3. $Cov[y_t, y_s] = E[y_t - E(y_t)][y_s - E(y_s)] = \gamma_{t,s}$, *covarianza* entre dos tiempos.
4. $Corr[y_t, y_s] = \frac{Cov[y_t, y_s]}{\sqrt{Var[y_t]Var[y_s]}} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}} = \rho_{t,s}$, *correlación* de y entre el tiempo t y s.

Definición 2. Se dice que el proceso estocástico $\{y_t\}$ es un *proceso de ruido blanco* con varianza σ^2 si es un proceso estacionario tal que:

1. $E(y_t) = 0, \quad t = 0, 1, 2, \ldots$
2. $Var[y_t] = \sigma^2, \quad t = 0, 1, 2, \ldots$
3. $Cov[y_t, y_{t-k}] = 0, \quad k = \pm 1, \pm 2, \ldots$

Es decir, la esperanza del proceso siempre es cero por lo tanto, no depende de t, la varianza del proceso es σ^2 constante. Así que, la varianza tampoco depende del tiempo y por último la tercera condicional implica que las variables del proceso están no correlacionadas para todos los retardos.

Recordando que, un retardo (o rezago, en algunos textos) es cuando se toma el valor pasado de una variable para usarlo en el presente.

De Definición 2 se deduce que un proceso de ruido blanco satisface:

$$Corr[y_t, y_{t-k}] = \frac{Cov[y_t, y_{t-k}]}{\sqrt{Var[y_t]Var[y_{t-k}]}} = 0, \;\; k = \pm 1, \pm 2, \ldots$$

Por lo tanto, según Wooldridge [26], el ruido blanco es trivialmente asintóticamente incorrelado, es decir, satisface la propiedad de dependencia débil (lo cual se definirá más adelante).

2.1.2. Caracterización y Clasificación de Series de Tiempo

Una serie de tiempo es un conjunto de observaciones cuantitativas registradas secuencialmente en el tiempo. Estas observaciones pueden exhibir diversos patrones, como tendencias, ciclos, estacionalidades o simplemente comportamientos aleatorios. Formalmente, de acuerdo con Box [3] una serie temporal se denota como $\{y_t\}$, donde cada y_t representa el valor observado en el instante t.

Los modelos de series de tiempo se enfocan en la predicción a partir del comportamiento histórico de la variable en estudio [3]. En general, se distinguen dos enfoques principales:

- **Modelos deterministas:** se basan en métodos de extrapolación, sin considerar la aleatoriedad inherente a los datos. Son más simples, pero tienden a ser menos precisos. Un ejemplo es el uso de promedios móviles como técnica de suavizamiento, que calcula el valor futuro como el promedio de observaciones recientes.
- **Modelos estocásticos:** describen los datos como realizaciones de un proceso aleatorio subyacente. Dado que las distribuciones conjuntas de estos procesos son generalmente desconocidas o complejas, se construyen modelos aproximados que capturen sus propiedades esenciales y permitan realizar inferencias y pronósticos.

Una serie de tiempo $\{y_t\}$ puede ser estacionaria o no estacionaria. Para caracterizar estas propiedades, es necesario introducir algunas definiciones básicas de estadística de procesos.

Definición 3. Se dice que un proceso estocástico (serie de tiempo) es *estacionario en sentido débil*, o débilmente estacionario, si se cumple que:

1. $E(y_t) = \mu$, constante, para $t = 0, 1, 2, \ldots$.
2. $Var[y_t] = \gamma_0$, constante, para $t = 0, 1, 2, \ldots$.
3. $Cov[y_t, y_{t-k}] = E[y_t - \mu][y_{t-k} - \mu] = \gamma_k$, $k = \pm 1, \pm 2, \ldots$, depende solo de la distancia temporal k, no del tiempo t en sí.

Para simplificar, nos referimos a estos procesos solamente como "estacionarios".

Consideremos s, t dos instantes de tiempo. Entonces si un proceso es no estacionario, tanto $E(y_t)$ como $Var[y_t]$ dependen de t. Si el proceso es estacionario, estos momentos son finitos y no dependen de t. Además, si el proceso no es estacionario la covarianza entre cualquier par de variables aleatorias dependen tanto de t como de s, mientras que si el proceso es estacionario la covarianza solo depende de la distancia temporal entre ambas, es decir, la covarianza entre y_t y y_s solo depende de $|t - s| = k > 0$, [26].

De la definición de estacionariedad podemos deducir que también las correlaciones entre dos variables del proceso dependen únicamente de la distancia temporal entre ambas, es decir, para todo $k = 1, 2, \ldots$:

$$Corr[y_t, y_{t-k}] = \frac{Cov[y_t, y_{t-k}]}{\sqrt{Var[y_t]Var[y_{t-k}]}} = \frac{\gamma_k}{\gamma_0} = \rho_k.$$

Otro concepto fundamental en el estudio de series de tiempo es el de *dependencia débil*, el cual impone restricciones sobre la intensidad de la relación estadística entre observaciones del proceso separadas en el tiempo. En particular, en el caso de procesos estocásticos *estacionarios*, dicha relación puede analizarse a través de su función de autocorrelación. Un proceso estacionario se considera débilmente dependiente si la correlación entre dos observaciones y_t y y_{t-k} disminuye conforme crece el desfase k, y tiende a cero en el límite. Esta propiedad se formaliza mediante la condición:

$$\text{Corr}[y_t, y_{t-k}] \xrightarrow[k\to\infty]{} 0.$$

Este comportamiento implica que las observaciones del proceso se vuelven asintóticamente incorrelacionadas a medida que aumenta la distancia temporal, lo que justifica la aplicabilidad de métodos que asumen independencia débil entre periodos lejanos. Cabe señalar que existen diversas formas de dependencia débil —como la mezcla fuerte o la correlación absolutamente sumable— que permiten extender este análisis a procesos no estacionarios o con estructuras más complejas [26].

2.2. Aplicaciones Multidisciplinarias de las Series Temporales

En la actualidad, muchas organizaciones necesitan anticipar el comportamiento futuro de diversos fenómenos con el fin de planificar estratégicamente o prevenir posibles contingencias. Para ello, las series de tiempo constituyen una herramienta básica, ya que permiten predecir el valor futuro de una variable a partir de sus patrones observados en el pasado. Las proyecciones a corto, mediano y largo plazo resultan especialmente útiles en contextos organizacionales, donde se requiere estimar, por ejemplo, la demanda de un producto, las ventas esperadas o la gestión eficiente del inventario. Algunas áreas de aplicación en las que se pueden utilizar las series de tiempo son:

1. **Finanzas y Economía:**
   - Predicción de precios de activos: modelar y predecir precios de acciones, tipos de cambio, tasas de interés y otros activos financieros.
   - Análisis macroeconómico: predecir indicadores económicos como el PIB, la inflación y el desempleo.
   - Riesgo financiero: evaluar y predecir la volatilidad del mercado y gestionar el riesgo en carteras de inversión.
2. **Salud:**
   - Epidemiología: seguimiento y predicción de la propagación de enfermedades infecciosas.
   - Monitorización de pacientes: análisis de datos de salud en tiempo real para detectar anomalías en señales vitales.
   - Planificación de recursos: previsión de la demanda de servicios de salud, como lo son las camas en hospitales y el personal médico.
3. **Meteorología y Medio Ambiente:**
   - Predicción del clima: predecir el clima a corto y largo plazo.

- Análisis ambiental: monitorear la calidad del aire, niveles de contaminación y cambios en el ecosistema.

4. **Marketing y Ventas:**
   - Previsión de ventas: predecir la demanda de productos y servicios.
   - Análisis de comportamiento del consumidor: seguimiento de tendencias y patrones en las compras de los clientes.
5. **Manufactura y Gestión de Operaciones:**
   - Mantenimiento predictivo: predicción de fallos en equipos y maquinaria para realizar mantenimiento preventivo.
   - Gestión de inventarios: optimizar la disponibilidad de productos y minimizar costos.
6. **Energía:**
   - Predicción de demanda energética: estimar la demanda futura de electricidad y otros recursos energéticos.
7. **Seguridad y Defensa:**
   - Análisis de incidentes: seguimiento y predicción de incidentes de seguridad.
   - Planificación estratégica: evaluar riesgos y planificar respuestas basadas en datos pasados.
8. **Transporte y Logística:**
   - Planificación de rutas: optimización de rutas de transporte basadas en patrones pasados de tráfico.
9. **Deportes:**
   - Análisis de rendimiento: seguimiento del rendimiento de atletas y equipos a lo largo del tiempo.
   - Estrategias de juego: análisis de patrones y tendencias en el desempeño deportivo para mejorar estrategias.
10. **Ciencias Sociales:**
    - Estudios demográficos: análisis y previsión de cambios en la población y tendencias migratorias.
    - Investigación sociológica: seguimiento de patrones en comportamiento social y cultural.

Éstas representan solo algunas de las múltiples áreas en las que pueden aplicarse las series de tiempo. Siempre que se disponga de observaciones registradas a lo largo del tiempo, es posible construir modelos de predicción que permitan analizar la evolución de las variables y anticipar su comportamiento futuro.

2.3. Componentes de una Serie de Tiempo

Una serie de tiempo puede descomponerse en cuatro componentes fundamentales (o cinco si se incluye una constante denominada *nivel*) los cuales no son directamente observables y deben ser estimados a partir de los datos. Estos componentes son:

a) **Tendencia ($T(t)$):** corresponde al comportamiento sistemático y persistente que presenta la serie a lo largo del tiempo, reflejando su dirección general en el largo plazo. Puede manifestarse como un crecimiento sostenido, una disminución progresiva o seguir un patrón más complejo. En términos generales, la tendencia representa el cambio estructural de los niveles de la serie a lo largo del tiempo. Formalmente, es una función determinista $T(t)$ que captura la evolución sistemática de los niveles medios de la serie a lo largo del tiempo. Si $Y(t)$ representa una serie temporal, puede expresarse como:

$$Y(t) = T(t) + \text{otros componentes},$$

donde $T(t)$ describe la dirección general del cambio en el tiempo. Esta función puede adoptar diferentes formas según el comportamiento observado, por ejemplo:

- **Lineal:** $T(t) = \beta_0 + \beta_1 t$,
- **Cuadrática:** $T(t) = \beta_0 + \beta_1 t + \beta_2 t^2$,
- **Exponencial:** $T(t) = \beta_0 e^{\beta_1 t}$,

entre otras, dependiendo de la naturaleza del fenómeno analizado.

b) **Estacionalidad ($E(t)$):** se refiere a las fluctuaciones periódicas y sistemáticas que ocurren en la serie con una frecuencia constante, generalmente dentro de un intervalo anual, trimestral, mensual u otro ciclo predefinido. Estas variaciones están asociadas a factores recurrentes como las estaciones del año, días festivos o eventos específicos del calendario. El componente estacional puede representarse mediante una función periódica:

$$E(t) = E(t+s), \quad \text{para todo } t \in \mathbb{Z},$$

donde s es el periodo de estacionalidad (por ejemplo, $s = 12$ para series mensuales con estacionalidad anual).

c) **Aleatoriedad ($A(t)$):** representa las variaciones impredecibles o residuales de una serie de tiempo que no pueden ser explicadas por la tendencia, la estacionalidad ni los ciclos. Este componente refleja el efecto de perturbaciones externas, errores de medición u otros factores no modelados, y se asume como un proceso estocástico con media cero y varianza constante.

Matemáticamente, suele modelarse como un término de error $A(t)$ que sigue un proceso de **ruido blanco**, denotado comúnmente como $\text{WN}(0, \sigma^2)$ (WN(White Noise)). Esto significa que:

$$\mathbb{E}[y_t] = 0, \quad \text{Var}(y_t) = \sigma^2, \quad \text{Cov}(y_t, y_{t-k}) = 0 \text{ para } k \neq 0.$$

El componente aleatorio es empleado para capturar la incertidumbre inherente al fenómeno observado y constituye la parte no sistemática del modelo.

d) **Ciclo ($C(t)$):** representa las fluctuaciones recurrentes pero no necesariamente periódicas en una serie de tiempo, asociadas con fases de expansión y contracción de fenómenos económicos, políticos o sociales. A diferencia de la estacionalidad, los ciclos tienen una duración variable y no están ligados a intervalos regulares del calendario. El componente cíclico suele modelarse

como una función suave de largo plazo que capta patrones de comportamiento que se repiten, pero con frecuencia y amplitud variables. Los ciclos, puede expresarse como una función suavizada:

$$C(t) = Y(t) - T(t) - E(t) - A(t),$$

cuando los demás componentes (tendencia, estacionalidad y ruido aleatorio) han sido identificados y extraídos. También puede estimarse mediante métodos como filtros de suavizamiento (por ejemplo, el filtro de Hodrick–Prescott) o descomposición STL (un método que descompone una serie en tendencia + estacionalidad + residuo) [11].

El poder identificar y modelar estas componentes es fundamental para comprender y predecir el comportamiento de una serie de tiempo.

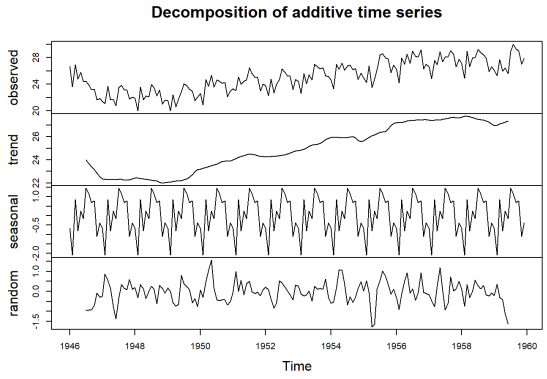


Figura 2.1: Componentes de una Serie de Tiempo. Fuente: [19].

En la **Figura 2.1** se muestra una representación gráfica que ilustra de manera clara los principales componentes de una serie de tiempo. Esta visualización facilita la identificación y análisis de elementos como la tendencia, el ciclo, la estacionalidad y la variabilidad aleatoria, proporcionando una comprensión estructurada de cómo se descomponen los datos observados.

Por ejemplo, en una serie que describe las ventas mensuales de una tienda, la **tendencia** puede reflejar un crecimiento sostenido a lo largo de los años; el **ciclo** puede capturar fluctuaciones asociadas a condiciones económicas generales; la **estacionalidad** puede evidenciar incrementos sistemáticos en diciembre debido a las compras navideñas; y el **componente aleatorio** recogería variaciones impredecibles causadas por factores externos, como interrupciones logísticas o eventos inesperados.

Desde un enfoque clásico, una serie de tiempo $\{y_t\}$ puede descomponerse mediante uno de los siguientes modelos:

- **Modelo aditivo:** $Y(t) = T(t) + C(t) + E(t) + A(t)$.
- **Modelo multiplicativo:** $Y(t) = T(t) \times C(t) \times E(t) \times A(t)$.

En resumen, la descomposición de una serie de tiempo en sus componentes estructurales permite comprender la dinámica subyacente del fenómeno analizado. Esta descomposición proporciona una base conceptual para la formulación de modelos estadísticos que capturen y aprovechen estos patrones con fines descriptivos y predictivos. En el siguiente capítulo se presentarán los modelos clásicos de series temporales, con énfasis en los modelos autorregresivos y de medias móviles, los

cuales permiten modelar la dependencia temporal entre observaciones y constituyen la base para modelos más generales como el $ARIMA$.

Capítulo 3

Modelo ARMA

Este capítulo está dedicado al estudio de los modelos lineales clásicos para series estacionarias: los modelos autorregresivos (AR), los modelos de medias móviles (MA) y su combinación, conocida como modelo autorregresivo de medias móviles (ARMA). Se presentarán sus definiciones formales, propiedades básicas, condiciones de estacionariedad e invertibilidad, así como métodos para su identificación, estimación e interpretación.

3.1. Procesos Autorregresivos (AR)

Los modelos más sencillos de procesos estacionarios utilizados para representar la dependencia de los valores de una serie temporal respecto a su pasado son los modelos autorregresivos, los cuales extienden el concepto de regresión lineal entre dos variables aleatorias.

Los procesos autorregresivos, nombrados así por su relación con la regresión, fueron los primeros procesos estacionarios en ser estudiados [9]. Estos modelos suponen que el valor actual de la serie puede expresarse como una combinación lineal de sus propios valores pasados, más un término aleatorio que representa perturbaciones no explicadas. Su simplicidad matemática y la claridad en la interpretación de sus parámetros los convierten en herramientas fundamentales para el análisis de series temporales.

3.1.1. AR(1)

Una serie temporal $\{y_t\}$ sigue un proceso autorregresivo de orden uno (AR(1)) si está definida por la siguiente ecuación:

$$y_t = c + \phi y_{t-1} + a_t, \qquad t = 1, 2, \ldots \tag{3.1}$$

donde c y ϕ son constantes y $\{a_t\}$ es un proceso de ruido blanco con media cero, varianza constante σ^2 y sin correlación serial [4].

En este modelo, el valor presente de la serie y_t depende linealmente de su valor inmediato anterior y_{t-1} y de un componente aleatorio a_t. El proceso recibe el nombre de autorregresivo debido a su estructura similar a una regresión lineal, en la que la variable explicativa es el propio pasado de la serie. Fue uno de los primeros procesos estacionarios en ser estudiados formalmente [9].

Para determinar si un proceso AR(1) es estacionario, es necesario verificar las siguientes condiciones [9]:

a) Estacionariedad en media: Evaluamos la esperanza matemática del proceso:

$$\mathbb{E}(y_t) = \mathbb{E}(c + \phi y_{t-1} + a_t).$$

Aplicando la linealidad de la esperanza:

$$\mathbb{E}(y_t) = c + \phi\mathbb{E}(y_{t-1}) + \mathbb{E}(a_t).$$

Dado que $\mathbb{E}(a_t) = 0$, se tiene:

$$\mathbb{E}(y_t) = c + \phi\mathbb{E}(y_{t-1}).$$

Si el proceso es estacionario, entonces $\mathbb{E}(y_t) = \mu$ constante para todo t, por lo que:

$$\mu = c + \phi\mu \quad \Rightarrow \quad (1-\phi)\mu = c \quad \Rightarrow \quad \mu = \frac{c}{1-\phi}.$$

Para que μ sea finita, es necesario que $|\phi| < 1$.

b) Estacionariedad en covarianza: Además, de una media constante, la varianza y la covarianza del proceso deben ser constantes y no depender del tiempo. Comenzamos con el cálculo de la varianza:

$$\gamma_0 = \mathrm{Var}(y_t) = \mathbb{E}\left[(y_t - \mu)^2\right],$$

y usando la forma centrada del modelo:

$$y_t - \mu = \phi(y_{t-1} - \mu) + a_t,$$

obtenemos:

$$\gamma_0 = \mathbb{E}[(\phi(y_{t-1} - \mu) + a_t)^2] = \phi^2\gamma_0 + \sigma^2 + 2\phi\mathbb{E}[(y_{t-1} - \mu)a_t].$$

Dado que a_t es ruido blanco, sin relación lineal con y_{t-1}, el término cruzado se anula:

$$\mathbb{E}[(y_{t-1} - \mu)a_t] = 0.$$

Así, la varianza satisface:

$$\gamma_0 = \phi^2\gamma_0 + \sigma^2 \quad \Rightarrow \quad (1-\phi^2)\gamma_0 = \sigma^2 \quad \Rightarrow \quad \gamma_0 = \frac{\sigma^2}{1-\phi^2}.$$

De igual manera, la **función de autocovarianza** de orden k, denotada $\gamma_k = \mathrm{Cov}(y_t, y_{t-k}) = E[(y_t - \mu)(y_{t-k} - \mu)]$, donde $\mu = E(y_t)$, debe depender únicamente del desfase k, y no del tiempo t. Para el modelo AR(1), se cumple la siguiente relación recursiva:

$$\gamma_k = \phi\gamma_{k-1}, \quad \text{para } k \geq 1.$$

Aplicando recursión:

$$\gamma_1 = \phi\gamma_0, \quad \gamma_2 = \phi\gamma_1 = \phi^2\gamma_0, \quad \gamma_k = \phi^k\gamma_0.$$

La **función de autocorrelación** (FAC), definida como $\rho_k = \gamma_k/\gamma_0$, toma la forma:

$$\rho_k = \phi^k, \quad \text{para } k = 0, 1, 2, \ldots$$

Esta función decrece exponencialmente cuando $|\phi| < 1$, lo cual es característico de procesos AR(1) estacionarios. En otras palabras, la dependencia entre valores separados en el tiempo se debilita progresivamente a medida que el desfase k aumenta, y $\rho_k \to 0$ cuando $k \to \infty$, aunque nunca se anula exactamente [18].

3.1.2. AR(p)

Una serie temporal estacionaria $\{y_t\}$ sigue un proceso autorregresivo de orden p, denotado como $AR(p)$, si puede expresarse como una combinación lineal de sus p valores pasados, más un término de innovación a_t, es decir:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + a_t, \qquad t = 1, 2, \ldots \tag{3.2}$$

donde $\phi_1, \phi_2, \ldots, \phi_p$ son los parámetros autorregresivos y $\{a_t\}$ es un proceso de ruido blanco con media cero y varianza constante σ^2.

Utilizando el operador de rezago L, definido por $Ly_t = y_{t-1}$, la ecuación (3.2) puede escribirse de forma compacta como:

$$\left(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p\right) y_t = a_t,$$

o bien,

$$\phi_p(L) y_t = a_t,$$

donde $\phi_p(L)$ es el *polinomio autorregresivo* de orden p, definido como:

$$\phi_p(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p,$$

y el vector $(\phi_1, \phi_2, \ldots, \phi_p)$ contiene los coeficientes autorregresivos del modelo [9].

Para que un proceso $AR(p)$ sea estacionario, es necesario verificar ciertas condiciones sobre los parámetros del modelo. Mientras que en el caso del modelo $AR(1)$ esta condición es sencilla ($|\phi| < 1$), en modelos de orden superior el análisis es más complejo, ya que los parámetros deben satisfacer restricciones más estrictas. A continuación, se presenta un teorema fundamental que establece las condiciones necesarias y suficientes para que un proceso $AR(p)$ sea estacionario [9, 10].

Teorema 1. Un proceso autorregresivo finito $AR(p)$ es estacionario si y solo si todas las raíces del polinomio autorregresivo $\phi_p(L)$ tienen módulo estrictamente mayor que uno; es decir, se encuentran fuera del círculo unidad en el plano complejo.

Además de la condición de estacionariedad, un modelo lineal general —incluido el proceso $AR(p)$— debe satisfacer dos propiedades adicionales: debe ser *no anticipante* e *invertible*.

Según Hamilton [10], las principales características de un proceso $AR(p)$ estacionario son las siguientes:

a) **Media:** Sea el modelo:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + a_t.$$

Tomando esperanza:

$$\mathbb{E}(y_t) = \phi_1 \mathbb{E}(y_{t-1}) + \phi_2 \mathbb{E}(y_{t-2}) + \cdots + \phi_p \mathbb{E}(y_{t-p}) + \mathbb{E}(a_t).$$

Como $\mathbb{E}(a_t) = 0$ y, bajo estacionariedad, $\mathbb{E}(y_t) = \mu$ constante, se obtiene:

$$\mu = \phi_1 \mu + \phi_2 \mu + \cdots + \phi_p \mu = \mu(\phi_1 + \phi_2 + \cdots + \phi_p),$$

por lo tanto:

$$(1 - \phi_1 - \phi_2 - \cdots - \phi_p)\mu = 0 \quad \Rightarrow \quad \mu = 0.$$

Este resultado es válido solo cuando no se incluye término constante. Si se generaliza el modelo para incorporar una constante δ:

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

entonces la media del proceso es:

$$\mu = \mathbb{E}(y_t) = \frac{\delta}{1 - \phi_1 - \phi_2 - \cdots - \phi_p},$$

siempre que se cumpla la condición de estacionariedad.

b) **Función de autocorrelación:** La función de autocorrelación (FAC), denotada ρ_k para $k = 0, 1, 2, \ldots$, de un proceso $AR(p)$ estacionario decrece hacia cero conforme aumenta el desfase k. Si bien en el caso del modelo $AR(1)$ esta función tiene forma exponencial simple, en modelos de orden superior ($p > 1$) la estructura de la FAC es más compleja y puede presentar distintos comportamientos, incluyendo oscilaciones o signos alternantes, dependiendo de los valores de los coeficientes ϕ_i.

3.2. Procesos de Medias Móviles (MA)

Los procesos de medias móviles (MA) y los procesos autorregresivos (AR) difieren significativamente en la forma en que incorporan la información pasada, lo cual se traduce en una diferencia en la duración del impacto de las perturbaciones a_t sobre el proceso.

Los modelos MA son conocidos como *procesos de memoria corta*, ya que una perturbación a_t afecta únicamente a un número limitado de observaciones futuras. Por ejemplo, en un proceso $MA(q)$, el efecto de una innovación desaparece por completo después de q periodos. En particular, en un $MA(1)$, la perturbación a_t incide únicamente en y_t y y_{t+1}, sin influir en observaciones posteriores. Esta característica hace que las funciones de autocorrelación de los modelos MA se anulen a partir del desfase $q + 1$, lo que refleja su estructura de dependencia limitada en el tiempo.

Por el contrario, los modelos autorregresivos $AR(p)$ son considerados *procesos de memoria larga*, ya que el efecto de una perturbación a_t persiste indirectamente en el sistema a través de su influencia sobre los valores futuros de la serie. Es decir, la innovación afecta a y_t, pero como y_t entra en la determinación de y_{t+1}, y así sucesivamente, su impacto se propaga a lo largo del tiempo. Aunque la magnitud de dicha influencia tiende a disminuir, en los modelos AR estacionarios ésta decae de forma exponencial pero nunca desaparece completamente. Por esta razón, la función de autocorrelación de un proceso AR no se anula para valores grandes de k, sino que decrece asintóticamente hacia cero.

Este contraste entre procesos de memoria corta y larga es importante para entender la estructura de dependencia temporal en las series y tiene implicaciones tanto para la modelación como para la predicción [6].

Sin embargo, para estudiar sus características más a fondo, empezaremos con el modelo más simple: el modelo de medias móviles de primer orden, $MA(1)$.

3.2.1. MA(1)

Un proceso de medias móviles de orden uno, denotado $MA(1)$, determina el valor de y_t en función de la innovación contemporánea a_t y su primer retardo a_{t-1}. El modelo está dado por:

$$y_t = a_t - \theta a_{t-1}, \qquad t = 1, 2, \ldots \tag{3.3}$$

donde θ es un parámetro real y $\{a_t\}$ es un proceso de ruido blanco con media cero y varianza constante σ^2. Aunque θ actúa como coeficiente del término rezagado de la innovación, no es un parámetro autorregresivo, ya que el modelo no incluye regresión de y_t sobre sus propios valores pasados.

Para verificar la estacionariedad del modelo $MA(1)$ para cualquier valor del parámetro θ, se deben analizar las siguientes condiciones [9]:

a) **Estacionariedad en media:** Como $\mathbb{E}(a_t) = \mathbb{E}(a_{t-1}) = 0$, se tiene:

$$\mathbb{E}(y_t) = \mathbb{E}(a_t - \theta a_{t-1}) = 0.$$

Por lo tanto, el modelo $MA(1)$ es estacionario en media para todo valor de θ.

b) **Estacionariedad en covarianza:**
Varianza:

$$\begin{aligned}\gamma_0 = \mathrm{Var}(y_t) &= \mathbb{E}[(a_t - \theta a_{t-1})^2] \\ &= \mathbb{E}(a_t^2) + \theta^2 \mathbb{E}(a_{t-1}^2) - 2\theta \mathbb{E}(a_t a_{t-1}) \\ &= \sigma^2 + \theta^2 \sigma^2 + 0 = (1 + \theta^2)\sigma^2.\end{aligned}$$

Esta es finita para todo $\theta \in \mathbb{R}$.
Autocovarianza de orden 1:

$$\begin{aligned}\gamma_1 &= \mathbb{E}[(a_t - \theta a_{t-1})(a_{t-1} - \theta a_{t-2})] \\ &= \mathbb{E}(a_t a_{t-1}) - \theta \mathbb{E}(a_{t-1}^2) - \theta \mathbb{E}(a_t a_{t-2}) + \theta^2 \mathbb{E}(a_{t-1} a_{t-2}) \\ &= 0 - \theta\sigma^2 - 0 + 0 = -\theta\sigma^2.\end{aligned}$$

Autocovarianza de orden mayor a 1:

$$\gamma_k = 0 \quad \text{para todo } k > 1,$$

ya que no hay superposición entre los términos a_t y a_{t-k} cuando $k > 1$ debido a la independencia del ruido blanco.

En consecuencia, el modelo $MA(1)$ presenta autocovarianzas finitas que dependen únicamente del desfase k y no del tiempo t, cumpliendo así con la condición de estacionariedad en covarianza para todo $\theta \in \mathbb{R}$.

Función de autocorrelación. La función de autocorrelación (FAC), definida como $\rho_k = \gamma_k/\gamma_0$, para un proceso $MA(1)$ es:

$$\rho_k = \begin{cases} 1 & \text{si } k = 0, \\ \dfrac{-\theta}{1 + \theta^2} & \text{si } k = 1, \\ 0 & \text{si } k > 1. \end{cases}$$

La FAC de un proceso $MA(1)$ se caracteriza por ser una función *truncada*, ya que se anula a partir del segundo desfase. Este comportamiento contrasta con el de los procesos autorregresivos, cuya autocorrelación decrece gradualmente [4].

Invertibilidad y no anticipación. Todo modelo lineal debe ser *no anticipante*, es decir, no debe depender de información futura. El modelo $MA(1)$ cumple esta condición, ya que y_t depende únicamente de a_t y a_{t-1}, ambos observables en el tiempo t o anteriores.

Además, para que el modelo sea *invertible*, es decir, que pueda representarse como una combinación infinita de valores pasados de y_t con coeficientes convergentes, se requiere que: $|\theta| < 1$.

Esta condición asegura que la representación equivalente en forma autorregresiva infinita converja. Por lo tanto, aunque el modelo $MA(1)$ es estacionario para todo θ, solo es invertible cuando $|\theta| < 1$, [6].

Modelo con media distinta de cero. El modelo $MA(1)$ puede extenderse para incluir una media diferente de cero:

$$y_t = \delta + a_t - \theta a_{t-1}. \tag{3.4}$$

La media del proceso es entonces:

$$\mathbb{E}(y_t) = \mathbb{E}[\delta + a_t - \theta a_{t-1}] = \delta,$$

y tanto la función de autocovarianza como la de autocorrelación permanecen iguales a las del modelo centrado.

3.2.2. MA(q)

La extensión natural del modelo $MA(1)$ es el proceso de medias móviles de orden q, conocido como $MA(q)$. En este caso, el valor presente de la serie y_t depende no solo de la innovación contemporánea a_t, sino también de las q innovaciones anteriores. El modelo se define como:

$$y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}, \tag{3.5}$$

donde $\{\theta_1, \theta_2, \ldots, \theta_q\}$ son los parámetros de medias móviles y $\{a_t\}$ es un proceso de ruido blanco con media cero y varianza σ^2.

Utilizando el operador de rezago L, el modelo se puede escribir de forma compacta como:

$$y_t = \left(1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_q L^q\right) a_t = \theta_q(L) a_t,$$

donde $\theta_q(L)$ es el *polinomio de medias móviles* de orden q.

El modelo $MA(q)$ generaliza al $MA(1)$, por lo que comparten muchas propiedades. Sin embargo, al incluir más retardos de las perturbaciones pasadas, la estructura dinámica se vuelve más compleja y el proceso adquiere una memoria más prolongada. En particular, una innovación ocurrida en el tiempo t afecta a $y_t, y_{t+1}, \ldots, y_{t+q}$, pero no influye en observaciones más alejadas en el tiempo.

De acuerdo con Box, Jenkins y Reinsel [4], un proceso $MA(q)$ es siempre *estacionario*, ya que se construye como una combinación lineal finita de perturbaciones independientes e idénticamente distribuidas. Esto implica que sus momentos (esperanza, varianza y autocovarianza) no dependen del tiempo t.

Esperanza:

$$\mathbb{E}[y_t] = \mathbb{E}[a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}] = 0. \tag{3.6}$$

Autocovarianza:
Multiplicando ambos lados del modelo por y_{t-k} y tomando esperanza, se obtiene la función de autocovarianza γ_k. Para $k = 0$, la varianza es:

$$\gamma_0 = \mathrm{Var}(y_t) = \left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right)\sigma^2.$$

Para $k = 1, 2, \ldots, q$, las autocovarianzas se calculan como:

$$\gamma_k = \sigma^2 \sum_{j=0}^{q-k} \theta_j \theta_{j+k},$$

donde se define $\theta_0 = 1$. Para $k > q$, las autocovarianzas se anulan:

$$\gamma_k = 0 \quad \text{para } k > q.$$

Autocorrelación:
La función de autocorrelación se define como $\rho_k = \gamma_k/\gamma_0$, y por lo tanto:

$$\rho_k = \frac{\sum_{j=0}^{q-k} \theta_j \theta_{j+k}}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2}, \quad \text{para } 1 \leq k \leq q,$$

y

$$\rho_k = 0, \quad \text{para } k > q.$$

Este comportamiento truncado de la función de autocorrelación es característico de los procesos $MA(q)$.

Modelo con media distinta de cero:

El modelo puede extenderse fácilmente para representar procesos con media no nula mediante la inclusión de un término constante δ:

$$y_t = \delta + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}. \tag{3.7}$$

En este caso, la esperanza del proceso es:

$$\mathbb{E}(y_t) = \delta,$$

mientras que la varianza, autocovarianzas y autocorrelaciones permanecen sin cambio respecto al modelo centrado.

La prueba del siguiente resultado puede ser consultada en [4].

Teorema 2. Un proceso de medias móviles finito $MA(q)$ es *invertible* si y solo si todas las raíces del polinomio de medias móviles $\theta_q(L)$ tienen módulo estrictamente mayor que uno, es decir, se encuentran fuera del círculo unidad en el plano complejo.

3.3. Procesos Autorregresivos de Medias Móviles: ARMA(p, q)

Un proceso autorregresivo de medias móviles, denotado $ARMA(p, q)$, combina las estructuras de los modelos $AR(p)$ y $MA(q)$. En este caso, el valor de y_t depende de sus p valores pasados, de la innovación contemporánea a_t, y de los q retardos de dicha innovación. El modelo se expresa como:

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}. \tag{3.8}$$

Utilizando el operador de rezago L, el modelo se puede escribir de manera compacta como:

$$\left(1 - \phi_1 L - \ldots - \phi_p L^p\right) y_t = \left(1 - \theta_1 L - \ldots - \theta_q L^q\right) a_t,$$
$$\phi_p(L) y_t = \theta_q(L) a_t,$$

donde $\phi_p(L)$ es el polinomio autorregresivo y $\theta_q(L)$ el polinomio de medias móviles [10].

¿Es estacionario el modelo $ARMA(p, q)$ para cualquier valor de sus parámetros?

Teorema 3. Un proceso $ARMA(p, q)$ es *estacionario* si y solo si todas las raíces del polinomio autorregresivo $\phi_p(L)$ tienen módulo estrictamente mayor que uno, es decir, se encuentran fuera del círculo unidad en el plano complejo.

Para que un proceso $ARMA(p, q)$ sea estadísticamente adecuado, además de la estacionariedad, debe cumplir dos propiedades fundamentales: ser *no anticipante* e *invertible*, las cuales garantizan que el modelo sea estable y computacionalmente tratable.

a) **No anticipante:** Un proceso es no anticipante si el valor presente de la serie y_t depende únicamente de valores pasados y actuales, y no de valores futuros. Esta propiedad está directamente relacionada con la *estabilidad* del modelo.
Para verificar esta condición, se analiza el polinomio autorregresivo $\phi_p(L)$: todas sus raíces deben estar fuera del círculo unidad. Esta condición garantiza que el proceso no explote y que su comportamiento sea estable en el tiempo.

b) **Invertible:** Un proceso es invertible si puede representarse como un modelo autorregresivo infinito, es decir, si los valores presentes pueden expresarse como una combinación convergente de errores pasados. Esto permite identificar las perturbaciones a_t a partir de los valores observados de y_t, y es esencial para una estimación eficiente de los parámetros.
La invertibilidad se verifica analizando el polinomio de medias móviles $\theta_q(L)$: todas sus raíces deben tener módulo mayor que uno. Esta condición evita la dependencia de errores futuros y garantiza que el proceso sea recuperable desde sus innovaciones pasadas.

Teorema 4. Un proceso autorregresivo de medias móviles finito $ARMA(p, q)$ es invertible sí y solo sí el módulo de las raíces del polinomio medias móviles $\phi_p(L)$ está fuera del círculo unidad.

Las condiciones de invertibilidad del modelo $ARMA(p, q)$ están determinadas exclusivamente por la parte de medias móviles, ya que la parte autorregresiva ya está expresada directamente en forma invertida. Si bien la parte AR regula la estacionariedad, la parte MA define la invertibilidad del proceso [4, 10].

El modelo $ARMA(p, q)$ combina las estructuras de los modelos $AR(p)$ y $MA(q)$, compartiendo sus principales propiedades. Si es estacionario, posee media constante (cero en su forma centrada), varianza finita e independiente del tiempo, y una función de autocovarianza de soporte infinito. A diferencia del modelo $MA(q)$, cuya función de autocorrelación se trunca en el desfase q, en el modelo $ARMA(p, q)$ la autocorrelación es también infinita, decreciendo gradualmente hacia cero sin truncamiento [10].

Consideremos ahora el modelo específico $ARMA(1, 1)$, definido por:

$$y_t = \phi_1 y_{t-1} + a_t - \theta_1 a_{t-1}, \qquad t = 1, 2, \ldots \tag{3.9}$$

La memoria extendida del proceso se debe a la estructura autorregresiva. Aunque la perturbación a_t afecta directamente a y_t y a través del término $-\theta_1 a_{t-1}$ también a y_{t+1}, su influencia se transmite indirectamente al futuro mediante la secuencia de valores de y.

Estacionariedad: Para verificar la estacionariedad del proceso, se analizan las raíces del polinomio autorregresivo:

$$1 - \phi L = 0 \quad \Rightarrow \quad L = \frac{1}{\phi} \quad \Rightarrow \quad |\phi| < 1.$$

Invertibilidad: De forma análoga, la condición de invertibilidad se obtiene a partir del polinomio de medias móviles:

$$1 - \theta L = 0 \quad \Rightarrow \quad |\theta| < 1.$$

Las características de un proceso $ARMA(1, 1)$ estacionario son:

a) **Media:**

$$\begin{aligned} \mathbb{E}(y_t) &= \phi_1 \mathbb{E}(y_{t-1}) + \mathbb{E}(a_t) - \theta_1 \mathbb{E}(a_{t-1}) \\ &= \phi_1 \mu + 0 - 0 = \phi_1 \mu \quad \Rightarrow \quad (1 - \phi_1)\mu = 0 \quad \Rightarrow \quad \mu = 0. \end{aligned}$$

b) **Función de autocovarianza:**
La varianza del proceso (autocovarianza de orden 0) es:

$$\gamma_0 = \frac{(1 + \theta_1^2 - 2\phi_1\theta_1)\sigma^2}{1 - \phi_1^2}.$$

La autocovarianza de primer orden es:

$$\gamma_1 = \phi_1 \gamma_0 - \theta_1 \sigma^2.$$

Para órdenes mayores, se obtiene la relación recursiva:

$$\gamma_k = \phi_1 \gamma_{k-1}, \quad \text{para } k > 1.$$

Esto muestra que la memoria del proceso más allá de $k = 1$ está determinada únicamente por la estructura autorregresiva.

c) **Función de autocorrelación:**
Dividiendo γ_k entre γ_0, se obtiene la función de autocorrelación:

$$\rho_k = \begin{cases} 1 & k = 0, \\ \dfrac{\gamma_1}{\gamma_0} = \phi_1 - \dfrac{\theta_1 \sigma^2}{\gamma_0} & k = 1, \\ \phi_1 \rho_{k-1} & k > 1. \end{cases}$$

La función de autocorrelación es infinita: el primer coeficiente depende tanto del componente autorregresivo como del de medias móviles, mientras que los coeficientes siguientes decaen exponencialmente con razón ϕ_1, siguiendo la dinámica impuesta por el componente $AR(1)$. Esta estructura permite una mayor flexibilidad en la representación de series temporales con autocorrelaciones iniciales no monótonas [9].

Los resultados obtenidos para el modelo $ARMA(1,1)$ pueden extenderse al modelo general $ARMA(p, q)$. En este caso, la función de autocorrelación (FAC) también tiene soporte infinito, es decir, sus valores no se anulan para ningún desfase k, aunque decrecen rápidamente hacia cero.

Los primeros q coeficientes de autocorrelación, $\rho_1, \ldots, \rho_q$, están influenciados tanto por los parámetros autorregresivos $\phi_1, \ldots, \phi_p$ como por los parámetros de medias móviles $\theta_1, \ldots, \theta_q$. A partir del desfase $q + 1$, los coeficientes de la FAC ya no dependen explícitamente de los términos de medias móviles y su comportamiento está determinado por la parte autorregresiva del modelo.

En particular, si todas las raíces del polinomio autorregresivo $\phi_p(L)$ son reales, la FAC decrece exponencialmente hacia cero. Si el polinomio tiene raíces complejas conjugadas, la FAC presenta un comportamiento oscilatorio, con amortiguamiento sinusoidal (combinación de senos y cosenos) [7].

El modelo $ARMA(p, q)$ puede generalizarse para incluir una media distinta de cero mediante la adición de una constante δ. La forma general es:

$$y_t = \delta + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}. \tag{3.10}$$

Este modelo mantiene las mismas propiedades dinámicas que el modelo centrado, pero su esperanza cambia. Calculando la media del proceso:

$$\begin{aligned} \mathbb{E}(y_t) &= \mathbb{E}\left[\delta + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}\right] \\ &= \delta + \phi_1 \mu + \cdots + \phi_p \mu + 0 \\ &= \delta + \mu(\phi_1 + \cdots + \phi_p), \end{aligned}$$

donde $\mu = \mathbb{E}(y_t)$. Despejando:

$$(1 - \phi_1 - \cdots - \phi_p)\mu = \delta \quad \Rightarrow \quad \mathbb{E}(y_t) = \frac{\delta}{1 - \phi_1 - \cdots - \phi_p}.$$

Esta expresión es válida siempre que el proceso sea estacionario, es decir, cuando todas las raíces del polinomio autorregresivo $\phi_p(L)$ se encuentren fuera del círculo unidad.

En este capítulo se han presentado los principales modelos lineales para series temporales estacionarias: los procesos autorregresivos $AR(p)$, de medias móviles $MA(q)$, y su combinación $ARMA(p, q)$. Se expusieron sus propiedades fundamentales, condiciones de estacionariedad e invertibilidad, así como el comportamiento de sus funciones de autocovarianza y autocorrelación. Estos modelos constituyen la base teórica para el análisis dinámico de series temporales y sirven como punto de partida para modelos más generales que permiten capturar tendencias, estacionalidades y transformaciones estructurales. En el siguiente capítulo se abordará la metodología para la identificación, estimación y validación de estos modelos.

Capítulo 4

Modelo ARIMA

El modelo $ARMA$ combina dos componentes fundamentales para el análisis de series temporales estacionarias: una parte autorregresiva $AR(p)$ y una parte de medias móviles $MA(q)$. Sin embargo, muchas series temporales reales presentan tendencias o estructuras no estacionarias, lo que impide aplicar directamente este tipo de modelos. En tales casos, es necesario transformar la serie mediante diferenciación para eliminar la no estacionariedad y permitir el uso de modelos basados en procesos estacionarios.

Esta idea da origen al modelo $ARIMA(p, d, q)$, donde la letra "I" representa la parte integrada del modelo y corresponde al número de veces que debe diferenciarse la serie para que sea estacionaria. El modelo $ARIMA$ (Autorregresivo Integrado de Medias Móviles) fue propuesto por Box y Jenkins en 1976 como una herramienta para la modelación y predicción de series temporales no estacionarias [3].

La forma general del modelo, una vez aplicada la diferenciación de orden d, se expresa como:

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}, \tag{4.1}$$

donde $\{a_t\}$ es un proceso de ruido blanco con media cero y varianza constante. En la notación $ARIMA(p, d, q)$, el parámetro p representa el orden del componente autorregresivo, d el orden de integración (número de diferencias aplicadas), y q el orden del componente de medias móviles.

La construcción de un modelo $ARIMA$ adecuado sigue una metodología sistemática, comúnmente conocida como la metodología de Box-Jenkins [3], la cual consta de los siguientes pasos:

a) **Identificación:** A partir del análisis exploratorio de los datos y de la estructura de autocorrelación, se propone una subclase de modelos $ARIMA(p, d, q)$ candidatos. El objetivo es determinar los valores apropiados de p, d y q, así como decidir si se debe incluir una constante δ en el modelo.

b) **Estimación:** Se realiza una inferencia estadística eficiente de los parámetros del modelo propuesto, condicionada a la especificación estructural elegida. Esta etapa busca obtener estimaciones de los coeficientes $\phi_1, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q$, la varianza del término de error σ^2, y, en su caso, la constante δ.

c) **Validación:** Se aplican pruebas de diagnóstico para evaluar la adecuación del modelo a los datos. Si se detectan discrepancias sistemáticas entre el modelo y la serie observada, se reconsidera la especificación o se ajustan los parámetros hasta alcanzar un modelo estadísticamente válido.

d) **Predicción:** Una vez validado el modelo, se utiliza para generar predicciones probabilísticas de los valores futuros de la serie. Esta etapa también permite evaluar el desempeño predictivo del modelo, lo cual es un aspecto de suma importancia en aplicaciones prácticas.

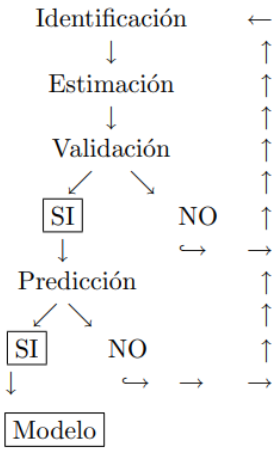


Figura 4.1: Pasos Modelo ARIMA

La **Figura 4.1** ilustra el esquema general del procedimiento Box-Jenkins para la construcción de modelos ARIMA. El proceso es iterativo: inicia con la identificación del modelo adecuado mediante el análisis de la serie y sus correlogramas; posteriormente, se estiman los parámetros del modelo propuesto. A continuación, se realiza una fase de validación, en la que se aplican pruebas estadísticas para evaluar la adecuación del modelo. Si los resultados no son satisfactorios, se vuelve a la etapa de identificación para ajustar la especificación. Una vez validado, se procede a la etapa de predicción. Si el modelo muestra un buen desempeño predictivo, se acepta como modelo final. En caso contrario, se reevalúa la estructura planteada. Este ciclo puede repetirse varias veces hasta encontrar un modelo que cumpla con los criterios de ajuste y predicción deseados.

4.1. Identificación

El propósito de esta fase es seleccionar un modelo $ARIMA(p, d, q)$ que represente adecuadamente la estructura dinámica de la serie temporal bajo análisis. Esta etapa busca determinar los valores apropiados de los órdenes p, d y q, garantizando que el modelo resultante capture las características fundamentales del comportamiento de los datos, como tendencia, estacionalidad, persistencia y autocorrelación.

La identificación del modelo se divide en dos subfases principales:

a) **Análisis de estacionariedad.**
El primer paso consiste en diagnosticar si la serie original es estacionaria y, en caso contrario, aplicar las transformaciones necesarias para alcanzar la estacionariedad. Este análisis comprende dos aspectos:

1) **Estacionariedad en media:** se determina el número de diferencias necesarias (d) para eliminar la tendencia o estructuras no estacionarias en la media. Esto puede hacerse de forma visual (a través de gráficos) o mediante pruebas estadísticas como la prueba de Dickey-Fuller aumentada (ADF) [10].

2) **Estacionariedad en varianza:** si la varianza de la serie cambia con el tiempo (heterocedasticidad), pueden aplicarse transformaciones estabilizadoras como logaritmos, raíz cuadrada o Box-Cox [4]. Este paso es de suma importancia cuando se observan cambios abruptos o crecimiento exponencial en la serie.

b) **Determinación de los órdenes p y q.**
Una vez que se ha transformado la serie en una serie estacionaria en media y varianza (si es necesario), se procede a identificar los valores de p (autorregresivo) y q (media móvil). El objetivo es encontrar un modelo $ARMA(p, q)$ que represente adecuadamente la dinámica de la serie diferenciada.

A continuación se describe con mayor precisión cada uno de los pasos comentados.

4.1.1. Estacionariedad en Media

En esta sección, se debe determinar si la serie es estacionaria en media, es decir, si fluctúa alrededor de un nivel constante. Para tomar esta decisión, nos basaremos en las características que distinguen a las series estacionarias de las no estacionarias.

Una **serie estacionaria** en media es aquella en la que se puede suponer que existe una única media constante a lo largo de toda la serie temporal, lo que implica que las fluctuaciones se mantienen alrededor de un valor promedio fijo. En un proceso estacionario en media, la función de autocorrelación teórica disminuye de manera rápida y exponencial.

Una **serie no estacionaria** en su media muestra una tendencia o está compuesta por segmentos con medias distintas. En términos generales, un proceso con una raíz unitaria presenta una función de autocorrelación muestral que decae de manera muy lenta, sin que sea necesario que se acerque constantemente a la unidad.

Si la serie no es estacionaria en media, se puede alcanzar la estacionariedad aplicando diferencias. Por lo tanto, si la serie no es estacionaria en media, se realizarán diferencias sucesivas de orden 1 hasta obtener una serie estacionaria:

$$z_t = (1 - L)^d y_t. \tag{4.2}$$

- $(1 - L)$ representa la **diferenciación en primer orden**, que se calcula como:

$$(1 - L)y_t = y_t - y_{t-1}. \tag{4.3}$$

  Esto nos da la primera diferencia de la serie, que ayuda a eliminar tendencias lineales y hacer que la serie sea estacionaria.

- $(1 - L)^d$ denota una **diferenciación de orden** d. Para $d = 2$, por ejemplo:

$$(1 - L)^2 y_t = (1 - L)(1 - L)y_t = y_t - 2y_{t-1} + y_{t-2}, \tag{4.4}$$

  que corresponde a la **segunda diferencia**, útil cuando la serie tiene una tendencia cuadrática.

Sin embargo, el problema radica en determinar la cantidad exacta de diferencias necesarias para que la serie se vuelva estacionaria, ya que pueden surgir los siguientes problemas:

a) **Sobre-diferenciación**: Si se aplican demasiadas diferencias, la serie puede volverse demasiado "plana" o perder características importantes de la estructura temporal, como las tendencias subyacentes o la estacionalidad. Esto puede dificultar la interpretación del modelo y generar una predicción menos precisa.

b) **Sub-diferenciación**: Si no se aplican suficientes diferencias, la serie puede seguir mostrando características no estacionarias, como tendencias o ciclos, lo que llevará a un modelo que no capte correctamente la dinámica temporal y, por lo tanto, a una mala predicción.

c) **Raíces unitarias múltiples**: En algunos casos, puede haber más de una raíz unitaria, lo que complica la identificación de cuántas diferencias son realmente necesarias para alcanzar la estacionariedad. Este tipo de problemas requiere técnicas más complejas para detectarlas.

d) **Efecto en la autocorrelación**: Al aplicar diferencias, se pueden alterar las relaciones de autocorrelación, lo que puede hacer que los patrones temporales originales sean difíciles de identificar o modelar de manera adecuada.

e) **Dificultad en la interpretación**: Después de aplicar una serie de diferencias, la interpretación de los resultados puede volverse más complicada, ya que los valores de la serie original ya no son directamente observables.

Por ejemplo, los valores de d más habituales son $d = 0, 1, 2$, y para decidir cuál es el más apropiado para la serie bajo estudio utilizaremos los siguientes instrumentos:

a) Gráfico de la serie original y las transformaciones correspondientes, para observar si se cumple o no la condición de estacionariedad de oscilar en torno a un nivel constante.

b) Correlograma estimado de la serie original y de las transformaciones correspondientes, para comprobar si decrece rápidamente hacia cero o no.

c) Contrastes de raíces unitarias.

Las pruebas de raíces unitarias proporcionan herramientas estadísticas que permiten inferir, a partir del conjunto de información disponible, la existencia o no de una raíz unitaria en una serie, es decir, la no estacionariedad de la serie. Si se rechaza la hipótesis nula de existencia de raíz unitaria, no se realizarán más diferencias en la serie. En caso contrario, si no se rechaza la hipótesis nula, se tomará una diferencia adicional de orden 1.

Los contrastes de raíces unitarias más utilizados en la práctica del análisis de series temporales económicas y que, a la vez, fueron los pioneros en este campo, es el de **Dickey-Fuller Aumentado**, que es una generalización de la prueba de Dickey-Fuller [10, 17].

4.1.2. Estacionariedad en Varianza

Una serie se considera estacionaria en varianza cuando tiene una varianza constante a lo largo de toda la serie temporal. Es decir, la dispersión de los datos alrededor de la media debe mantenerse uniforme en el tiempo. Si la serie no cumple con esta condición, se pueden aplicar transformaciones estabilizadoras de varianza, como las transformaciones **Box-Cox**, para lograrla.

$$y_t^{(\lambda)} = \begin{cases} \frac{y_t^{\lambda}-1}{\lambda} & \lambda \neq 0, \\ \ln(y_t) & \lambda = 0. \end{cases}$$

Las transformaciones **Box-Cox** incluyen una amplia variedad de funciones, como la raíz cuadrada y la inversa, entre otras. Debido a que las series económicas suelen ser positivas y no presentan valores cero, la transformación logarítmica es la más utilizada en el ámbito

económico. Para analizar la estacionariedad en varianza de una serie, se emplean herramientas como los gráficos de la serie original y sus correspondientes transformaciones.

4.1.3. Prueba de Dickey-Fuller Aumentada

Prueba de Dickey-Fuller

La prueba de *Dickey-Fuller* es un contraste estadístico diseñado para detectar la presencia de raíces unitarias en un proceso autorregresivo de primer orden, es decir, en un modelo $AR(1)$, en este, se parte de la suposición de que se tienen t observaciones de una serie de tiempo $\{y_t\}$, restando y_{t-1} en ambos lados de la ecuación (3.1) se tiene:

$$y_t - y_{t-1} = c + \phi y_{t-1} - y_{t-1} + a_t,$$

$$\nabla y_t = c + \beta y_{t-1} + a_t, \tag{4.5}$$

donde $\Delta y_t = y_t - y_{t-1}$ representa la primera diferencia de la serie y el parámetro $\beta = \phi - 1$. Esta reparametrización es conocida como el *modelo de corrección de errores* (4.5). El objetivo de la prueba es determinar si la serie $\{y_t\}$ es estacionaria o posee una raíz unitaria. Para ello se plantea el siguiente conjunto de hipótesis:

$$\mathbf{H_0} : \beta = 0,$$

vs

$$\mathbf{H_a} : \beta < 0.$$

Es decir, que si $\beta = 0$ entonces la serie tendrá una raíz unitaria que implica que la serie no es estacionaria, por el contrario, si $\beta < 0$ entonces la serie no tiene raíces unitarias y por lo tanto, la serie es estacionaria.

El estadístico de la prueba se construye como:

$$t_\nu = \frac{\hat{\beta}}{\hat{s}_\beta}, \tag{4.6}$$

donde $\hat{\beta}$ es estimador de mínimos cuadrados de β y $\hat{s}_\beta$ es el error estándar asociado a dicho estimador.

Este estadístico, conocido como el *estadístico de Dickey-Fuller*, no se ajusta a ninguna distribución conocida. Por lo tanto, Dickey y Fuller calcularon sus percentiles bajo la hipótesis nula, $\mathbf{H_0}$, proporcionando tablas con los niveles críticos apropiados para el estadístico, según el tamaño de la muestra y el nivel de significancia α.

Se rechaza $\mathbf{H_0}$ (y se concluye que la serie es estacionaria) cuando el *p-value* es menor que un umbral de significancia, típicamente 0.05.

La prueba de Dickey-Fuller puede ser extendida para considerar dinámicas más complejas o errores autocorrelacionados, lo cual da lugar a la llamada **prueba de Dickey-Fuller aumentada** (Augmented Dickey-Fuller, ADF), que se describirá a continuación.

Prueba de Dickey-Fuller Aumentada

Los contrastes anteriores de Dickey-Fuller se han desarrollado bajo la suposición restrictiva de que la serie sigue un proceso $AR(1)$. Sin embargo, la serie puede seguir procesos más generales como $ARMA(p,q)$. Es bien sabido que cualquier proceso $ARMA(p,q)$ puede aproximarse con la precisión necesaria utilizando un $AR(p)$:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + a_t.$$

Este modelo se puede reparametrizar como:

$$\nabla y_t = \beta y_{t-1} + \alpha_1 \nabla y_{t-1} + \ldots + \alpha_{p-1} \nabla y_{t-p+1} + a_t, \tag{4.7}$$

donde:

$$\beta = \sum_{i=1}^{p} \phi_i - 1, \quad \text{y} \quad \alpha_i = -\sum_{j=i+1}^{p} \phi_j \quad \text{para } i = 1, \ldots, p-1.$$

La hipótesis nula que se plantea en este contexto es la existencia de una raíz unitaria, lo cual implica no estacionariedad en la serie. Esto ocurre si la suma de los coeficientes autorregresivos es igual a uno. El contraste estadístico se define como:

$$\begin{cases} \mathbf{H_0} : \beta = 0 & \text{(la serie tiene una raíz unitaria)} \\ \mathbf{H_a} : \beta < 0 & \text{(la serie es estacionaria)} \end{cases}$$

El estadístico de prueba es el mismo utilizado en la prueba de Dickey-Fuller simple:

$$t_\nu = \frac{\hat{\beta}}{\hat{s}_\beta}, \tag{4.8}$$

y su distribución bajo la hipótesis nula continúa siendo no estándar. Por tanto, se utiliza la misma tabla de valores críticos de Dickey-Fuller, ajustados para diferentes tamaños muestrales y niveles de significancia α. La regla de decisión sigue siendo:

$$\text{Rechazar } \mathbf{H_0} \text{ si } t_\nu < DF_\alpha.$$

Esta extensión de la prueba se conoce como la **prueba de Dickey-Fuller aumentada (ADF)** y permite contrastar la presencia de raíces unitarias en series con estructura autorregresiva más general. Además de comparar el valor del estadístico de prueba t_ν con los valores críticos tabulados por Dickey y Fuller, es común reportar el *p-valor* asociado al test. Este valor representa la probabilidad, bajo la hipótesis nula $\mathbf{H_0}$, de obtener un valor del estadístico tan extremo o más extremo que el observado. La regla general de decisión es:

- Si p-valor $< \alpha$, se rechaza $\mathbf{H_0}$: la serie no tiene raíz unitaria, por lo tanto, es **estacionaria**.
- Si p-valor $\geq \alpha$, no se rechaza $\mathbf{H_0}$: no hay evidencia suficiente para descartar la presencia de una raíz unitaria, es decir, la serie puede ser **no estacionaria**.

El valor típico para el nivel de significancia α es 0.05 o 0.01, aunque puede ajustarse según el contexto de análisis.

4.1.4. Determinación del Orden de p y q

Las características dinámicas de un proceso estacionario se manifiestan a través de la función de autocorrelación (FAC), que es la herramienta clave para determinar los órdenes p y q del modelo $ARMA$ que mejor describe las propiedades de la serie estacionaria z_t (4.2).

Los coeficientes de autocorrelación muestral de z_t son:

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{T^*}(z_t - \bar{z})(z_{t-k} - \bar{z})}{\sum_{t=1}^{T^*}(z_t - \bar{z})^2} \qquad k = 1, 2, \ldots, T^*/3, \tag{4.9}$$

donde $T^* = T - d$ es la longitud de la serie estacionaria z_t.

Para identificar los órdenes p y q, comparamos las funciones de autocorrelación muestrales con las FAC teóricas de los modelos $ARMA$ cuyas características son:

	FAC
$MA(q)$	Se anula para $j > q$.
$AR(q)$	Decrecimiento rápido. No se anula.
$ARMA(p,q)$	Decrecimiento rápido. No se anula.

Tabla 4.1: Comparación de funciones de autocorrelación muestrales con las **FAC**.

Para ayudarnos en la identificación de modelos $ARMA(p,q)$ recurrimos en este caso a la función de autocorrelación parcial.

El coeficiente de autocorrelación parcial de orden k, denotado por p_k, mide el grado de asociación lineal existente entre las variables y_t e y_{t-k} una vez ajustado el efecto lineal de todas las variables intermedias, es decir:

$$p_k = \rho_{y_t y_{t-k}.y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}}.$$

Por lo tanto, el coeficiente de autocorrelación parcial p_k es el coeficiente de la siguiente regresión lineal:

$$y_t = \alpha + p_1 y_{t-1} + p_2 y_{t-2} + \ldots + p_k y_{t-k} + e_t.$$

Las propiedades de la función de autocorrelación parcial (FACP), p_k, $k = 0, 1, 2, 3, \ldots$, son:

- $p_0 = 1$ y $p_1 = \rho_1$.
- Los coeficientes p_k no dependen de unidades y son menores que la unidad en valor absoluto.
- La FACP es una función simétrica.
- La FACP de un proceso estocástico estacionario decrece rápidamente hacia cero cuando $k \to \infty$.

La estructura de la función de autocorrelación parcial para un modelo estacionario $ARMA(p,q)$ es como sigue:

- Modelo $AR(p)$:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \ldots + \phi_p y_{t-p} + a_t.$$

  La estructura de la FACP es la siguiente:

$$p_k = \begin{cases} p_k \neq 0, & k = 1, 2, \ldots, p, \\ p_k = 0, & k = p+1, p+2, \ldots \end{cases}$$

  En un modelo $AR(p)$, y_t depende linealmente de sus p valores pasados. Las variables separadas por hasta p periodos tienen una relación directa, pero para periodos mayores, el coeficiente de autocorrelación parcial es cero porque no hay relación directa, solo indirecta a través de las intermedias.

- Modelo $MA(q)$:

$$y_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \theta_3 a_{t-3} + \ldots + \theta_q a_{t-q}.$$

  En un modelo de medias móviles finito, la FACP es infinita y disminuye rápidamente: exponencialmente si las raíces son reales o como una onda amortiguada si son complejas, coherente con su representación AR infinita.

- Modelo $ARMA(p,q)$:

$$y_t = a_t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}.$$

  En un modelo $ARMA(p,q)$, la FACP es infinita. Los primeros p coeficientes dependen de los parámetros AR y MA, mientras que desde el retardo $p+1$, solo de la parte MA, disminuyendo rápidamente: exponencialmente si las raíces son reales o como una onda amortiguada si son complejas.

Comparando la estructura de las funciones de autocorrelación simple y parcial estimadas con las características básicas de las funciones de autocorrelación teóricas tenemos la siguiente tabla:

	FAC	**FACP**
$MA(q)$	Se anula para $j > q$.	Decrecimiento rápido. No se anula.
$AR(q)$	Decrecimiento rápido. No se anula.	Se anula para $j > q$.
$ARMA(p,q)$	Decrecimiento rápido. No se anula.	Decrecimiento rápido. No se anula.

Tabla 4.2: Comparación entre **FAC** y **FACP**.

Una de las características de los procesos $AR(p)$ es que su FAC decrece geométricamente, mientras que en su FACP las primeras p observaciones son diferentes a cero, y a partir de la observación $p+1$ el valor de esta es cero. En el caso de un proceso $MA(q)$, el comportamiento de estas funciones es contrario: la función de autocorrelación simple decae a cero a partir de la observación $q+1$, mientras que la representación de la FACP presenta un decaimiento geométrico [4, 22].

Es por este motivo que la identificación de un proceso $AR(p)$ se logra utilizando la FACP, mientras que para identificar un proceso $MA(q)$ se requiere de la FAC.

Recordando la ecuación (4.8) tenemos que, para determinar cuándo un coeficiente $\hat{\rho}_k$ es diferente de cero, es necesario conocer su error estándar. Utilizando $\frac{1}{\sqrt{T^*}}$ como el error, se aproxima al error de un coeficiente de correlación entre variables independientes. Si todos los coeficientes de autocovarianza fueran cero, las desviaciones típicas serían aproximadamente $\frac{1}{\sqrt{T^*}}$. De este modo, se pueden construir bandas de confianza de $\pm\frac{2}{\sqrt{T^*}}$ y considerar significativos aquellos coeficientes que se encuentren fuera de dichas bandas.

La identificación de un modelo $ARMA(p,q)$ con funciones de autocorrelación no es sencilla. El objetivo inicial es reducir posibles modelos $ARIMA$, ajustando y validando después para refinar la identificación.

Al interpretar las funciones de autocorrelación, se busca identificar modelos simples que expliquen la serie. Hay que considerar posibles coeficientes significativos por azar, la correlación entre coeficientes muestrales y las ondulaciones del correlograma. Una muestra mayor facilita la identificación.

4.2. Estimación

Una vez identificados los procesos estocásticos que podrían haber generado la serie temporal y_t, el siguiente paso es estimar los parámetros desconocidos de esos modelos:

$$\beta = (\delta, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)' \qquad \text{y} \qquad \sigma_a^2.$$

Estos parámetros se pueden estimar de forma consistente por *Mínimos Cuadrados* o *Máxima Verosimilitud*. Ambos métodos de estimación se basan en el cálculo de las innovaciones, a_t, a partir de los valores de la serie estacionaria [4, 6].

El método de *Mínimos Cuadrados* minimiza la suma de cuadrados:

$$Min \sum_t a_t^2. \tag{4.10}$$

Para resolver el problema de estimación, las ecuaciones (4.9) y (4.10) se expresa en función del conjunto de información y de los parámetros desconocidos del modelo.

Para un modelo $ARMA(p,q)$, la innovación se puede escribir como:

$$a_t = Z_t - \delta - \sum_{i=1}^{p} \phi_i Z_{t-i} - \sum_{i=1}^{q} \theta_i a_{t-i}. \tag{4.11}$$

Por lo tanto, para calcular las innovaciones a partir de un conjunto de información y de un vector de parámetros desconocidos, se necesitan un conjunto de valores iniciales $Z_0, Z_1, \ldots, Z_{p-1}$ y $a_0, a_1, \ldots, a_{q-1}$.

Para aproximar las innovaciones, se imponen condiciones iniciales, como asumir que las primeras p observaciones son los valores iniciales y las innovaciones previas son cero. Esto genera estimadores de *Mínimos Cuadrados Condicionados* y de *Máxima Verosimilitud Condicionados*, que son equivalentes.

Los estimadores de *Máxima Verosimilitud* no condicionados se obtienen maximizando la verosimilitud exacta, que combina la verosimilitud condicionada y la densidad no condicionada de los valores iniciales.

4.3. Validación

Al construir un modelo $ARIMA$, el objetivo es encontrar el modelo que mejor represente el comportamiento de la serie estudiada. El modelo ideal debe cumplir con los siguientes criterios:

a) Los coeficientes del modelo son estadísticamente significativos y no están correlacionados entre sí.

b) Los residuos del modelo estimado se aproximan al comportamiento de un ruido blanco.

c) El grado de ajuste es elevado en comparación al de otros modelos alternativos.

d) El modelo es estacionario e invertible.

Si el modelo no cumple con los puntos anteriores entonces se debe modificar y se coteja con las características anteriores.

En primer lugar, es necesario realizar pruebas de significancia de los coeficientes AR y MA. Es decir, se debe verificar si el modelo incluye algún coeficiente o estructura que no sea relevante.

El juego de hipótesis para cada caso es:

$$\begin{aligned}
&\mathbf{H_0}: c = 0 \quad \text{vs} \quad \mathbf{H_a}: c \neq 0, \\
&\mathbf{H_0}: \phi_i = 0 \quad \text{vs} \quad \mathbf{H_a}: \phi_i \neq 0, \quad i = 1, \ldots, p, \\
&\mathbf{H_0}: \theta_j = 0 \quad \text{vs} \quad \mathbf{H_a}: \theta_j \neq 0, \quad j = 1, \ldots, q.
\end{aligned}$$

Si $\beta = (c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ generalmente la distribución asintótica de los estimadores β sigue una distribución normal con media $\hat{\beta}_i$ y varianza $V(\hat{\beta}_i)$, por lo que para contrastar $\mathbf{H_0}$ de no significatividad individual de los parámetros se utiliza el estadístico:

$$t = \frac{\hat{\beta}_i}{\sqrt{V(\hat{\beta}_i)}} \sim N(0,1).$$

Se rechaza $\mathbf{H_0}$ a un nivel de significancia $\alpha = 0.05$ si

$$\left| \frac{\hat{\beta}_i}{\sqrt{V(\hat{\beta}_i)}} \right| > N_{\alpha/2}(0,1) \simeq 2.$$

- Con $N_{\alpha/2}(0,1)$ el valor crítico de una distribución normal estándar con significancia $\alpha/2$.

Una vez confirmada la significancia de todos los parámetros, es necesario realizar un contraste sobre los residuos estimados para determinar si están correlacionados. Esto requiere calcular la FAC mediante:

$$\hat{r}_k = \frac{\sum_{t=1}^{T-k} (\hat{a}_t - \bar{a})(\hat{a}_{t+k} - \bar{a})}{\sum_{t=1}^{T-k} (\hat{a}_t - \bar{a})^2}.$$

donde $\bar{a}$ es la media de los T residuos; si los residuos no están correlacionados, los coeficientes $\hat{r}_k$ (para k grande) serán aproximadamente variables aleatorias con media cero y varianza asintótica $\frac{1}{T}$, y distribución normal.

El método más utilizado para verificar la falta de correlación en los residuos es trazar dos líneas paralelas, separadas por una distancia de $\frac{2}{\sqrt{T}}$ desde el origen, en las funciones de autocorrelación simple o parcial de los residuos. Es necesario asegurarse de que todos los coeficientes $\hat{r}_k$ estén dentro de estos límites de confianza. Esta representación se conoce como **correlograma de errores**.

Otro método comúnmente utilizado para verificar la ausencia de autocorrelación en los residuos es la prueba de **Box-Ljung**.

4.3.1. Prueba de Box-Ljung

La prueba de *Box-Ljung* fue creado por George E. P. Box y Greta M. Ljung en el año 1978. El objetivo de la prueba es comprobar si los residuos del modelo (es decir, la parte del modelo que no se puede explicar con sus componentes) están correlacionados entre sí.

Las Hipótesis de esta prueba son:

- $\boldsymbol{H_0}$ **(nula):** No hay autocorrelación significativa en los residuos (es decir, el modelo explica bien la estructura temporal).
- $\boldsymbol{H_1}$ **(alternativa):** Sí hay autocorrelación (el modelo podría estar mal especificado).

El estadístico de la prueba es:

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{r}_k^2}{n-k}$$

donde:

- n es el número de observaciones,
- h es el número de rezagos que se están probando,
- $\hat{r}_k$ es la autocorrelación muestral en el rezago k.

Este estadístico Q se compara con una distribución **chi-cuadrado** con $h - p - q$ grados de libertad.

Después de ajustar un modelo $ARIMA$, se usa la prueba de Box-Ljung para verificar que los residuos son ruido blanco. Si el valor p del test es mayor que 0.05, no se rechaza $\boldsymbol{H_0}$, lo que es bueno (el modelo captura bien la dependencia temporal) [5].

Una vez confirmada la falta de correlación de los residuos, se procede a realizar una prueba de normalidad sobre ellos, como la prueba de **Jarque-Bera**.

4.3.2. Prueba de Jarque-Bera

La prueba de Jarque-Bera se emplea para decidir si una serie sigue una distribución normal, el estadístico de prueba es:

$$JB = \frac{N}{6}\left[S^2 + \frac{(K-3)^2}{4}\right].$$

S representa el sesgo de la serie, K la kurtosis, N es el número de observaciones, JB es el estadístico de prueba, se distribuye de acuerdo con una distribución χ^2 con dos grados de libertad.

$\mathbf{H_0}$: La serie no sigue una distribución normal.
$\mathbf{H_a}$: La serie sigue una distribución normal.

Se rechaza $\mathbf{H_0}$ si $JB > \chi^2_{(2)} = 5.99$ con $\alpha = 0.05$.

4.4. Predicción

La predicción constituye el objetivo principal del modelo $ARIMA$. A través de la función de predicción, es posible elaborar estimaciones sobre el comportamiento futuro de la serie temporal. Esta función se fundamenta en la estructura del modelo seleccionado, que se considera el más adecuado para describir el comportamiento de los datos durante el periodo de estudio [4, 10].

Una vez obtenidos los valores predichos, es esencial evaluar la magnitud del error cometido, conocido como error de predicción. Este se define como la diferencia entre los valores observados de la serie y las predicciones realizadas.

4.4.1. Predicción con Modelos MA(q)

Comencemos por un modelo de medias móviles sencillo, por ejemplo, el $MA(2)$ de media cero:

$$y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}, \qquad t = 1, 2, \ldots$$

La función de predicción es:

$$\begin{aligned}
y_{T+1} =\ & a_{T+1} - \theta_1 a_T - \theta_2 a_{T-1}, \\
& y_T(1) = E_T[y_{T+1}] = E_T[a_{T+1} - \theta_1 a_T - \theta_2 a_{T-1}] = -\theta_1 a_T - \theta_2 a_{T-1}, \\
y_{T+2} =\ & a_{T+2} - \theta_1 a_{T+1} - \theta_2 a_T, \\
& y_T(2) = E_T[y_{T+2}] = E_T[a_{T+2} - \theta_1 a_{T+1} - \theta_2 a_T] = -\theta_2 a_{T-1}, \\
y_{T+3} =& a_{T+3} - \theta_1 a_{T+2} - \theta_2 a_{T+1}, \\
& y_T(3) = E_T[y_{T+3}] = E_T[a_{T+3} - \theta_1 a_{T+2} - \theta_2 a_{T+1}] = 0.
\end{aligned}$$

Donde: $\boldsymbol{y_T}$ representa la predicción en el tiempo T, $\boldsymbol{a_T}$ es el término de error (innovación) en el tiempo T y $\boldsymbol{\theta_1}$ y $\boldsymbol{\theta_2}$, son los parámetros del modelo que determinan la relación entre el error actual y los errores pasados en el proceso autorregresivo.

¿Por qué cambiamos de un tiempo t a un tiempo T? La respuesta es sencilla: en el modelo original, t representa cualquier instante en la serie temporal, mientras que en la predicción, T es un punto específico a partir del cual queremos estimar valores futuros. Utilizamos T porque estamos fijando un momento en el tiempo desde el cual realizamos las predicciones.

Este es un modelo $AR(2)$, en el cual las predicciones se basan en los errores pasados y los parámetros del modelo. A medida que se avanza en el tiempo, la influencia de los errores pasados disminuye gradualmente, hasta que las predicciones terminan por converger a cero.

De modo que tenemos que:

$$y_T(\ell) = E_T[y_{T+\ell}] = 0 \ \ (= E(y_t)), \quad \forall \ell > 2.$$

Por lo tanto, la función de predicción de un $MA(2)$, depende del conjunto de información, $\boldsymbol{I_T}$, que representa la información conocida en el momento T, para $\ell = 1, 2$. A partir de $\ell > 2$, la predicción óptima viene dada por la media del proceso.

Estos resultados se pueden generalizar fácilmente para el modelo $MA(q)$:

$$y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}, \qquad t = 1, 2, \ldots$$

Y entonces la función de predicción es:

$$y_T(\ell) = \begin{cases} -\theta_1 a_T - \theta_2 a_{T-1} - \ldots - \theta_q a_{T+1-q}, & \ell = 1, \\ -\theta_2 a_T - \theta_3 a_{T-2} - \ldots - \theta_q a_{T+2-q}, & \ell = 2, \\ \cdots & \cdots \\ -\theta_q a_T, & \ell = q, \\ 0, & \forall \ell = q+1, q+2, \ldots \end{cases}$$

Cabe señalar que la ecuación del error de predicción, denotado por e_T se define como:

$$e_T = y_T - \hat{y}_T.$$

Donde:

- y_T es el valor real observado en el tiempo T,
- $\hat{y}_T$ es el valor predicho por el modelo $ARIMA$ en el mismo tiempo.

Ahora, como el modelo $MA(2)$ está escrito directamente en forma de medias móviles, se obtiene la varianza del error de predicción con $\psi_i = \theta_i$, $i = 1, 2, \ldots, q$ y $\psi_i = 0$, $\forall i > q$, donde $\boldsymbol{\psi_i}$ describe cómo los errores pasados contribuyen a la varianza de la predicción.

$$V(e_T(\ell)) = \begin{cases} \sigma^2, & \ell = 1, \\ (1 + \theta_1^2)\sigma^2, & \ell = 2, \\ \cdots & \cdots \\ (1 + \theta_1^2 + \theta_2^2 + \ldots + \theta_{q-1}^2)\sigma^2, & \ell = q, \\ (1 + \theta_1^2 + \theta_2^2 + \ldots + \theta_q^2)\sigma^2 \ \ (= V(y_t)), & \ell = q+1, q+2, \ldots \end{cases}$$

Aunque la varianza del error de predicción ($V(e_T(\ell))$) es una función creciente de ℓ, el horizonte de predicción tiene una cota máxima que viene dada por la varianza no condicionada del proceso y que se alcanza para $\ell = q$.

Se puede concluir que, en un modelo $MA(q)$, las predicciones para $\ell \leq q$ usan los errores pasados, reduciendo la varianza del error en comparación con la media no condicionada. Para $\ell > q$, las predicciones se basan en la media no condicionada, y la información pasada deja de ser relevante.

La predicción por intervalo viene dado por:

$$\begin{aligned} \ell = 1, &\quad \left[-\theta_1 a_T - \theta_2 a_{T-1} - \ldots - \theta_q a_{T+1-q} \pm N_{\alpha/2}\sqrt{\sigma^2}\right], \\ \ell = 2, &\quad \left[-\theta_2 a_T - \theta_3 a_{T-2} - \ldots - \theta_q a_{T+2-q} \pm N_{\alpha/2}\sqrt{\sigma^2(1+\theta_1^2)}\right], \\ \cdots &\quad \cdots \\ \ell = q, &\quad \left[-\theta_q a_T \pm N_{\alpha/2}\sqrt{\sigma^2(1+\theta_1^2+\ldots+\theta_{q-1}^2)}\right], \\ \ell > q, &\quad \left[0 \pm N_{\alpha/2}\sqrt{\sigma^2(1+\theta_1^2+\ldots+\theta_{q-1}^2+\theta_q^2)}\right]. \end{aligned}$$

La amplitud de los intervalos de predicción va creciendo con ℓ, con el límite impuesto por:

$$\pm N_{\alpha/2}\sqrt{\sigma^2(1+\theta_1^2+\ldots+\theta_{q-1}^2+\theta_q^2)} \;=\; \pm N_{\alpha/2}\sqrt{V(y_t)}.$$

4.4.2. Predicción con Modelos AR(p)

Consideremos el modelo autorregresivo más sencillo, el $AR(1)$.

$$y_t = \phi y_{t-1} + a_t, \qquad t = 1, 2, \ldots$$

La función de predicción es:

$$\begin{aligned}
y_{T+1} &= \phi y_T + a_{T+1},\\
&\quad y_T(1) = E_T[y_{T+1}] = E_T[\phi y_T + a_{T+1}] = \phi y_T,\\
y_{T+2} &= \phi y_{T+1} + a_{T+2},\\
&\quad y_T(2) = E_T[y_{T+2}] = E_T[\phi y_{T+1} + a_{T+2}] = \phi E_T[y_{T+1}] = \phi y_T(1),\\
y_{T+3} &= \phi y_{T+2} + a_{T+3},\\
&\quad y_T(3) = E_T[y_{T+3}] = E_T[\phi y_{T+2} + a_{T+3}] = \phi E_T[y_{T+2}] = \phi y_T(2).
\end{aligned}$$

De forma que la función de predicción es:

$$y_T(\ell) = \phi y_T(\ell-1), \quad \ell = 2, 3, \ldots \tag{4.12}$$

dado que:

$$E_T(y_{T+j}) = \begin{cases} y_{T+j} & j \le 0,\\ E(y_T(j)) & j > 0. \end{cases}$$

$E_T(y_{T+j})$ representa la **esperanza condicional**, es decir, la mejor predicción de y_{T+j} con la información disponible en T.

Donde:

- Para $j \le 0$, simplemente tomamos el valor real de la serie porque ya ocurrió.
- Para $j > 0$, usamos la mejor predicción basada en la información disponible en T.

La función de predicción (4.12) utiliza una regla de cadena para generar predicciones en un proceso autorregresivo, donde cada predicción se basa en las anteriores hasta un intervalo temporal indefinido.

La trayectoria de esta función de predicción está determinada por la estructura de la parte autorregresiva:

$$\begin{aligned}
y_T(1) &= \phi y_T,\\
y_T(2) &= \phi y_T(1) = \phi\phi y_T = \phi^2 y_T,\\
y_T(3) &= \phi y_T(2) = \phi\phi^2 y_T = \phi^3 y_T,\\
y_T(4) &= \phi y_T(2) = \phi\phi^3 y_T = \phi^4 y_T,\\
y_T(\ell) &= \phi^\ell y_T, \quad \ell = 1, 2, 3, \ldots
\end{aligned}$$

Dado que el proceso autorregresivo es estacionario, se cumple que $|\phi| < 1$. Por lo tanto, al avanzar en el tiempo, la función de predicción converge hacia la media incondicional del proceso:

$$\lim_{\ell\to\infty} y_T(\ell) = 0 \;(= E(y_T)).$$

Para construir los intervalos de predicción, se ha de obtener la varianza del error de predicción. Para ello es preciso partir del modelo escrito en forma medias móviles. En el caso del $AR(1)$:

$$\begin{aligned}(1-\phi L)y_t = a_t \implies y_t &= \frac{1}{1-\phi L}a_t,\\ \implies y_t &= (1+\phi L+\phi^2L^2+\phi^3L^3+\ldots)a_t,\\ \implies y_t &= a_t+\phi a_{t-1}+\phi^2 a_{t-2}+\phi^3 a_{t-3}+\ldots\end{aligned}$$

Por lo que la varianza del error de predicción se obtiene con $\psi_1=\phi^i$, $\forall_i$:

$$V(e_T(\ell)) = \begin{cases}\sigma^2, & \ell=1\\ (1+\phi^2)\sigma^2, & \ell=2,\\ (1+\phi^2+(\phi^2)^2)\sigma^2, & \ell=3\\ (1+\phi^2+(\phi^2)^2+(\phi^3)^2)\sigma^2, & \ell=4,\\ (1+\phi^2+(\phi^2)^2+\ldots+(\phi^{\ell-1})^2)\sigma^2, & \ell>4.\end{cases}$$

La varianza del error de predicción aumenta de manera monótona a medida que avanzamos en el futuro. Sin embargo, dado que el proceso es estacionario, esta varianza no crece sin límite, sino que está acotada por la varianza incondicional del proceso:

$$\lim_{\ell\to\infty} V(e_T(\ell)) = \lim_{\ell\to\infty}\sigma^2\left[1+\phi^2+(\phi^2)^2+\ldots+(\phi^{\ell-1})^2\right] = \frac{\sigma^2}{1-\phi^2} = V(y_t).$$

La predicción por intervalo es:

$$\begin{aligned}&\ell=1, && \left[\phi y_T \pm N_{\alpha/2}\sqrt{\sigma^2}\right],\\ &\ell=2, && \left[\phi y_T(1) \pm N_{\alpha/2}\sqrt{\sigma^2(1+\phi^2)}\right],\\ &\ell=3, && \left[\phi y_T(2) \pm N_{\alpha/2}\sqrt{\sigma^2(1+\phi^2+(\phi^2)^2)}\right],\\ &\ldots && \ldots\\ &\ell>3, && \left[\phi y_T(\ell-1) \pm N_{\alpha/2}\sqrt{\sigma^2(1+\phi^2+(\phi^2)^2+\ldots+(\phi^{\ell-1})^2)}\right].\end{aligned}$$

La amplitud de los intervalos de predicción va creciendo con ℓ, con el límite impuesto por:

$$\pm N_{\alpha/2}\sqrt{\frac{\sigma^2}{1-\phi^2}} \;=\; \pm N_{\alpha/2}\sqrt{V(y_t)}.$$

Los resultados obtenidos para el modelo $AR(1)$ (3.1) se pueden extender para el modelo $AR(p)$ [3]. En general, las funciones de predicción de procesos autorregresivos puros se obtendrán a partir de reglas de cadena:

$$y_T(\ell) = \phi_1 y_T(\ell-1)+\phi_2 y_T(\ell-2)+\phi_3 y_T(\ell-3)+\ldots+\phi_p y_T(\ell-p), \quad \ell=1,2,3,\ldots$$

Entonces, en un proceso $AR(1)$, la predicción se basa en la última observación y_T, y a partir de ahí se generan las predicciones futuras. En un proceso $AR(p)$, se usan las últimas p observaciones para predecir los primeros p periodos, y las predicciones posteriores se derivan de estas.

4.4.3. Predicción con Modelos ARMA(p, q)

Consideremos un modelo $ARMA(p, q)$ sencillo, el $ARMA(1, 2)$:

$$y_t = \delta + \phi y_{t-1} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}, \qquad t = 1, 2, \ldots \tag{4.13}$$

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

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

Las predicciones por punto son:

$$\begin{aligned}
y_{T+1} &= \delta + \phi y_T + a_{T+1} - \theta_1 a_T - \theta_2 a_{T-1} \\
&\quad y_T(1) = E_T[y_{T+1}] = E_T[\delta + \phi y_T + a_{T+1} - \theta_1 a_T - \theta_2 a_{T-1}] = \\
&\quad = \delta + \phi y_T - \theta_1 a_T - \theta_2 a_{T-1}, \\
y_{T+2} &= \delta + \phi y_{T+1} + a_{T+2} - \theta_1 a_{T+1} - \theta_2 a_T \\
&\quad y_T(2) = E_T[y_{T+2}] = E_T[\delta + \phi y_{T+1} + a_{T+2} - \theta_1 a_{T+1} - \theta_2 a_T] = \\
&\quad = \delta + \phi y_T(1) - \theta_2 a_T, \\
y_{T+3} &= \delta + \phi y_{T+2} + a_{T+3} - \theta_1 a_{T+2} - \theta_2 a_{T+1} \\
&\quad y_T(3) = E_T[y_{T+3}] = E_T[\delta + \phi y_{T+2} + a_{T+3} - \theta_1 a_{T+2} - \theta_2 a_{T+1}] = \\
&\quad = \delta + \phi y_T(2),
\end{aligned}$$

$$\Longrightarrow y_T(\ell) = E_T[y_{T+\ell}] = \delta + \phi y_T(\ell - 1) \quad \forall \ell > 2.$$

La función de predicción se basa en la última observación y_T y los errores de predicción previos para los primeros dos periodos. Para $\ell > 2$, la parte de medias móviles desaparece, y las predicciones se obtienen de las anteriores a través de la parte autorregresiva, acercándose a la media del proceso con el tiempo:

$$\begin{aligned}
y_T(3) &= \delta + \phi y_T(2), \\
y_T(4) &= \delta + \phi y_T(3) = \delta + \phi(\delta + \phi y_T(2)) = \delta(1+\phi) + \phi^2 y_T(2), \\
y_T(5) &= \delta + \phi y_T(4) = \delta + \phi(\delta(1+\phi) + \phi^2 y_T(2)) = \delta(1 + \phi + \phi^2) + \phi^3 y_T(2), \\
&\cdots \qquad \cdots \\
y_T(\ell) &= \delta + \phi y_T(\ell - 1) = \delta(1 + \phi + \phi^2 + \ldots + \phi^{\ell-3}) + \phi^{\ell-2} y_T(2),
\end{aligned}$$

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

$$\lim_{\ell \to \infty} y_T(\ell) = \lim_{\ell \to \infty} \delta \sum_{i=0}^{\ell-3} \phi^i = \frac{\delta}{1-\phi} \quad (= E(y_t)).$$

Ahora, para calcular la varianza del error de predicción y, en consecuencia, construir los intervalos de predicción, es necesario derivar la representación de medias móviles infinitas ($MA(\infty)$).

Comenzamos con el modelo $ARMA(1, 2)$:

$$(1 - \phi L) y_t = (1 - \theta_1 L - \theta_2 L^2) a_t.$$

Dividiendo por $(1 - \phi L)$ en ambos lados, obtenemos la representación $MA(\infty)$:

$$y_t = \frac{(1 - \theta_1 L - \theta_2 L^2)}{(1 - \phi L)} a_t.$$

Usando la expansión en serie geométrica:

$$\frac{1}{1-\phi L} = 1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \dots$$

Multiplicando esta expansión por $(1 - \theta_1 L - \theta_2 L^2)$, obtenemos:

$$1 - \theta_1 L - \theta_2 L^2 = (1 - \phi L)(1 + \psi_1 L + \psi_2 L^2 + \psi_3 L^3 + \dots)$$

Igualamos coeficientes para obtener ψ_k

Para L,

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

Para L^2,

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

Para L^3,

$$0L^3 = (\psi_3 - \phi\psi_2)L^3 \implies 0 = \psi_3 - \phi\psi_2 \implies \psi_3 = \phi\psi_2$$

Entonces los coeficientes de la forma de medias móviles infinita son:

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

con estos coeficientes se pueden construir los 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 $\left[\pm N_{\alpha/2} \times \sqrt{V(y_t)}\right]$.

4.4.4. Predicción con Modelos no Estacionarios

La predicción utilizando modelos no estacionarios $ARIMA(p,d,q)$ se realiza de manera similar a los modelos estacionarios $ARMA(p,q)$. El predictor por punto óptimo de $y_{T+\ell}$ viene dado por la esperanza condicionada al conjunto de información $y_T(\ell) = E_T[y_{T+\ell}]$. Para calcular esta esperanza condicionada, es suficiente expresar el modelo en forma de ecuación en diferencias y determinar las esperanzas condicionadas, teniendo en cuenta que:

$$E_T[y_{T+j}] = \begin{cases} y_{T+j} & j \le 0, \\ y_T(j) & j > 0. \end{cases} \qquad E_T[a_{T+j}] = \begin{cases} a_{T+j} & j \le 0, \\ 0 & j > 0. \end{cases}$$

Para construir los intervalos de predicción,

$$\left[y_T(\ell) \pm N_{\alpha/2}\sqrt{V(e_T(\ell))}\right], \qquad \text{donde: } V(e_T(\ell)) = \sigma \sum_{j=0}^{\ell-1} \psi_j^2,$$

el modelo ha de estar escrito en forma $MA(\infty)$ ya que son los coeficientes del modelo $ARIMA$ escrito en forma de medias móviles [22].

En este capítulo se ha completado la construcción formal de modelos $ARIMA$, abordando los criterios de validación que permiten evaluar su idoneidad estadística y su capacidad para capturar adecuadamente la dinámica subyacente de una serie temporal. El siguiente paso consiste en aplicar esta metodología a un conjunto de datos reales, lo cual permitirá ilustrar de manera práctica todo el proceso de identificación, estimación, validación y predicción desarrollado hasta ahora.

Capítulo 5

Aplicación del Modelo ARIMA

En este capítulo se lleva a cabo la aplicación práctica de los modelos ARIMA al análisis de una serie de tiempo relacionada con los casos de esquizofrenia, con el objetivo de modelar su comportamiento histórico y realizar predicciones. Se comienza con un estudio descriptivo de la base de datos, destacando sus principales características estadísticas y visuales. A continuación, se implementa el proceso metodológico completo de modelado: identificación de la estructura temporal mediante análisis de autocorrelación, estimación de parámetros, validación del modelo a través de pruebas estadísticas y evaluación de residuos, y finalmente la generación de pronósticos.

5.1. Estudio Descriptivo de los Datos

5.1.1. Base de Datos

La base de datos utilizada en este estudio fue obtenida del sitio web del Instituto de Métricas y Evaluación de la Salud (IHME, por sus siglas en inglés) [24]. Este sitio recopila datos globales sobre diversas enfermedades, tanto físicas como mentales, y abarca información desde 1990 hasta 2021. El contenido de esta página permite al usuario personalizar la descarga de información mediante la selección de diversas variables y criterios. Para el análisis específico de la esquizofrenia, se seleccionaron las siguientes variables:

- **Sexo:** Masculino y Femenino.
- **Edad:** Se eligieron los siguientes rangos: 0-14 años, 15-19 años, 20-24 años, 25-29 años, 30-34 años, 35-39 años, 40-44 años, 45-49 años, 50-54 años, y 55+ años.
  La página ofrece la opción de seleccionar distintos rangos de edades. Se eligieron estos rangos para obtener una visión más completa de cómo se manifiesta esta enfermedad en las diferentes edades.
- **Años:** 1990-2021, con datos manejados anualmente.
- **Valor:** Para la variable valor se consideraron tres aspectos importantes, la primera de ellas es la correspondiente a *Medida*, es decir la prevalencia, esta, según la definición proporcionada en la misma página, se refiere a la prevalencia puntual, entendida como el número de personas que presentan la enfermedad en un momento específico del tiempo. El segundo aspecto es la *Métrica*, que representa el valor numérico asociado a dicha prevalencia, sin transformaciones ni ajustes adicionales. Mientras que la medida expresa un concepto epidemiológico, la métrica proporciona su cuantificación en un contexto

particular. Finalmente, se consideró la *Ubicación*, es decir, la población de la que se obtienen estos datos; en este caso, corresponden a la población del país de México.

Una vez seleccionadas todas las características, se procede a descargar la base de datos. La descarga se realiza a través de un enlace enviado al correo electrónico registrado previamente, ya que el acceso a esta página requiere una cuenta de usuario. Durante el registro, solo se solicita un correo electrónico y una breve descripción del uso previsto para los datos. Este proceso es necesario para garantizar la autenticidad y seguridad de la información proporcionada. El correo enviado contiene un enlace que dirige a otra página desde la cual se pueden descargar los datos. Estos se descargan en una carpeta comprimida que incluye un archivo Excel con la base de datos requerida.

La base de datos está compuesta por 640 registros, los cuales se desglosan por sexo, edad y año. Esta estructura es importante, ya que esto permite realizar un análisis detallado tanto por año como por sexo y edad, lo que facilita una mejor visualización y comprensión de cómo se comportan los datos desde diferentes perspectivas. Dada la cantidad de variables involucradas, la **Figura 5.1** presenta únicamente una vista parcial de la base de datos, con fines ilustrativos.

Sexo	Edad	Año	Valor	Superior	Inferior
Femenino	0-14 años	1990	409	672	227
Masculino	0-14 años	1990	525	847	299
Femenino	15-19 años	1990	2999	4241	1932
Masculino	15-19 años	1990	3707	5255	2420
Femenino	20-24 años	1990	8464	12177	5511
Masculino	20-24 años	1990	10004	14594	6567
Femenino	25-29 años	1990	12099	16038	8692
Masculino	25-29 años	1990	13863	18567	10030
Femenino	30-34 años	1990	12545	15801	9732
Masculino	30-34 años	1990	14352	18292	11110
Femenino	35-39 años	1990	11106	13706	8663
Masculino	35-39 años	1990	12847	15887	10021
Femenino	40-44 años	1990	8790	10582	7125
Masculino	40-44 años	1990	10295	12456	8326
Femenino	45-49 años	1990	6740	8041	5454
Masculino	45-49 años	1990	7889	9372	6376
Femenino	50-54 años	1990	5079	6048	4154
Masculino	50-54 años	1990	5854	6931	4828
Femenino	55+ años	1990	9724	11510	8268
Masculino	55+ años	1990	10618	12459	8978
Femenino	0-14 años	1991	406	665	226
Masculino	0-14 años	1991	524	842	299
Femenino	15-19 años	1991	3043	4297	1959
Masculino	15-19 años	1991	3798	5375	2476
Femenino	20-24 años	1991	8612	12408	5595
Masculino	20-24 años	1991	10281	14894	6738
Femenino	25-29 años	1991	12386	16405	8912
Masculino	25-29 años	1991	14279	19080	10350
Femenino	30-34 años	1991	12999	16382	10085
Masculino	30-34 años	1991	14856	18943	11517
Femenino	35-39 años	1991	11631	14358	9084
Masculino	35-39 años	1991	13421	16614	10491
Femenino	40-44 años	1991	9026	10887	7306
Masculino	40-44 años	1991	10618	12808	8604
Femenino	45-49 años	1991	7013	8361	5675
Masculino	45-49 años	1991	8193	9748	6625
Femenino	50-54 años	1991	5233	6226	4291
Masculino	50-54 años	1991	6042	7153	4985
Femenino	55+ años	1991	10032	11858	8537
Masculino	55+ años	1991	10969	12877	9286
Femenino	0-14 años	1992	404	661	225
Masculino	0-14 años	1992	525	842	300
Femenino	15-19 años	1992	3074	4345	1978

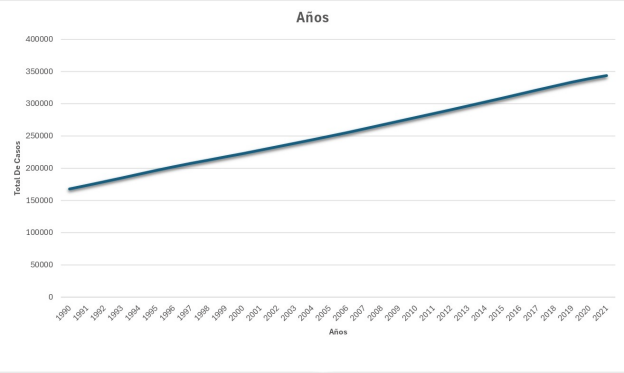
Figura 5.1: Base de Datos

5.1.2. Análisis de la Base de Datos

La plataforma desde la cual se descargó la base de datos no ofrecía la opción de visualizar los datos gráficamente. Sin embargo, se consideró esencial incluir este tipo de análisis, ya que las gráficas permiten presentar la información de manera más clara y comprensible. Para

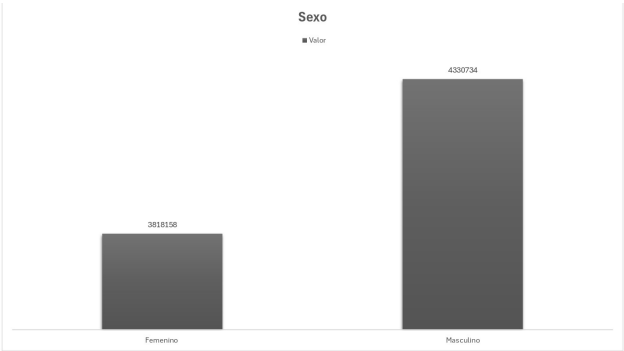
muchos usuarios, los patrones y tendencias resultan más evidentes cuando la información se presenta de forma visual.

Una de las gráficas más importantes para el análisis de los datos es aquella que muestra cómo varían los casos de esquizofrenia a lo largo de los años, dado que los datos se registran de manera anual; esta gráfica, etiquetada como "Años", representa el total de casos por año y se muestra en la **Gráfica 5.2**. Esta gráfica ilustra la tendencia de nuestros valores desde 1990 hasta 2021. Se observa un aumento lineal en los datos a lo largo de los años, lo que sugiere que esta tendencia podría continuar en el futuro.



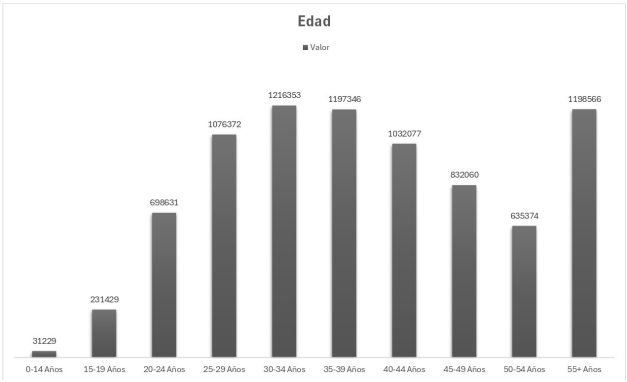
Gráfica 5.2: Casos por Año

La gráfica que se presenta en la **Gráfica 5.3**, muestra los casos de esquizofrenia según el sexo. Podemos apreciar con claridad que hay significativamente más individuos masculinos quienes padecen más casos de esquizofrenia. Esta gráfica es importante, ya que nos permite identificar si existe una mayor prevalencia de casos en hombres o en mujeres. Sin embargo, es importante señalar que esta conclusión se basa únicamente en los datos recopilados en esta base de datos, y no podemos generalizar que esta tendencia se mantenga en todos los casos de esquizofrenia.



Gráfica 5.3: Casos por Sexo

Finalmente, la **Gráfica 5.4** nos muestra el comportamiento de los casos de esquizofrenia en relación con las edades, las cuales fueron seleccionadas y filtradas en el momento en el que descargamos la base de datos. Las edades fueron elegidas de tal modo en que pudiéramos observar mejor en que rangos de edades se presentan más casos de Esquizofrenia. Este análisis revela que el rango de edad con mayor cantidad de casos abarca desde los 25 hasta los 39 años. También podemos observar una gran cantidad de casos en la categoría 55+ años.



Gráfica 5.4: Casos por Edad

En resumen, los datos muestran una tendencia lineal creciente en el número de casos de esquizofrenia a lo largo del tiempo. Asimismo, se observa que la mayor concentración de casos se presenta en personas de entre 25 y 39 años, así como en el grupo de 55 años o más. Por otro lado, la cantidad de casos registrados en hombres supera consistentemente a la de mujeres.

Como un comentario adicional, podemos notar cómo, si bien nuestra base de datos puede proporcionarnos toda la información necesaria, el uso de métodos visuales, en este caso las gráficas, nos permite apreciar de manera más efectiva cómo se comportan nuestros datos y cómo se relacionan entre sí.

5.2. Análisis por Series de Tiempo

Como se mencionó anteriormente, el análisis de series de tiempo, específicamente mediante el modelo $ARIMA$, se realizará con el apoyo de RStudio. El objetivo es pronosticar los datos obtenidos de la base de datos sobre esquizofrenia para determinar si, según las tendencias observadas, el número de personas con esquizofrenia aumentará, se mantendrá estable o, en un escenario ideal, disminuirá con el paso de los años. RStudio facilita considerablemente el análisis de nuestros datos gracias a su amplia gama de paquetes especializados. Estos paquetes no solo aceleran el proceso, sino que también permiten un manejo más eficiente y preciso de las técnicas estadísticas involucradas. Además, RStudio ofrece un entorno integrado que favorece la visualización, la manipulación de datos y la interpretación de resultados, lo que contribuye a un análisis más detallado.

Es importante recordar que para implementar el modelo $ARIMA$ de manera efectiva, es necesario seguir una serie de pasos que garantizan la obtención de un pronóstico preciso. Estos pasos incluyen la **identificación, estimación, validación y predicción**. Cada uno de estos pasos contribuye a desarrollar un modelo que optimice la precisión del pronóstico. En el **Capítulo 4**, titulado *Modelo ARIMA*, se ofrece una explicación detallada de cada uno de estos pasos, brindando todo lo necesario para poder llevar a cabo un análisis riguroso y obtener resultados confiables.

De manera más detallada, cada uno de estos pasos se compone de los siguientes elementos:

a) **Identificación.**

1) Realizar la gráfica de la serie.
2) Determinar estacionariedad por medio de:
   i. Test de Dickey-Fuller.
   ii. Correlograma de errores.
3) En caso de no ser estacionaria diferenciar.
4) Realizar la gráfica de la serie diferenciada y determinar de nueva cuenta estacionariedad. En caso de ser estacionaria realizar el siguiente paso.

b) **Estimación.**

1) Estimar los parámetros del modelo.

c) **Validación.**

1) Realizar pruebas de normalidad con el test de Jarque-Bera e histograma.
2) Determinar si los parámetros estimados son significativos por medio del correlograma de errores del modelo y el test de Box-Ljung.
3) Calcular raíces unitarias del modelo y graficarlas.

d) **Predicción.**

1) Realizar una predicción.
2) Calcular los errores de predicción.

5.2.1. Estudio de los Casos por Años

Todos los datos presentados a continuación fueron obtenidos del Instituto de Métricas y Evaluación de la Salud (IHME). Estos datos se actualizan anualmente y, dado que solo se obstuvieron observaciones hasta 2021. El análisis se realizará por años, sin considerar el género ni la edad.

En total, trabajaremos con 29 datos. Esto se debe a que, según se explica en la página de la cual se descargó la base de datos, se recopilaron los casos de esquizofrenia registrados en cada año en cuestión. Por ejemplo, si en el siguiente año se mantienen los mismos casos del año anterior más uno adicional, ese sería el total para ese año. Este procedimiento podría explicar por qué en la **Gráfica 5.2** los casos de esquizofrenia muestran un comportamiento creciente.

Ahora, explicando con mayor detalle el por qué obtuvimos únicamente 29 datos, lo que hicimos fue lo siguiente: en primer lugar, sumamos todos los casos de esquizofrenia registrados por año, es decir, agrupamos y totalizamos los casos correspondientes a 1990, 1991, y así sucesivamente. A partir de estos totales anuales, calculamos la diferencia entre años consecutivos; por ejemplo, al total de casos del año 1991 le restamos los del año 1990. De este modo, obtuvimos únicamente los nuevos casos registrados específicamente en 1991. Este procedimiento explica por qué en el año 1990 no se obtuvo un valor (marcado como NA), ya que no existía un año previo con el cual comparar.

Finalmente, debido a que los casos reportados en 2020 y 2021 generaban comportamientos atípicos en la serie temporal, afectando negativamente el análisis y la modelación con $ARIMA$, se optó por excluir dichos años. Como resultado de todo este proceso, se obtuvo la tabla que se muestra en la **Figura 5.5**.

Año	Valor
1990	NA
1991	5453
1992	5712
1993	5871
1994	5912
1995	5840
1996	5557
1997	5293
1998	5093
1999	5025
2000	5138
2001	5252
2002	5310
2003	5332
2004	5388
2005	5499
2006	5641
2007	5828
2008	5900
2009	5898
2010	5915
2011	5924
2012	6006
2013	6035
2014	6074
2015	6162
2016	6217
2017	6236
2018	6177
2019	6019

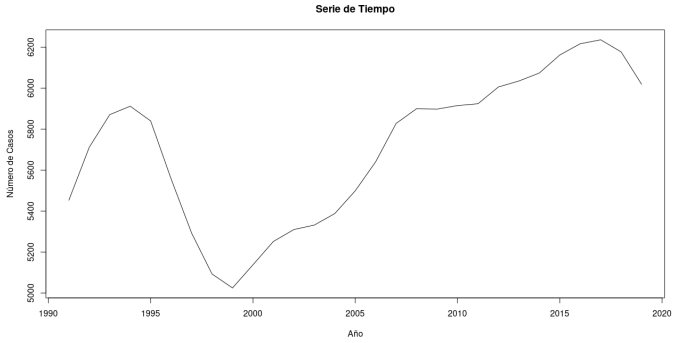
Figura 5.5: Tabla de Casos por Años

Ahora, utilizando el software RStudio, comenzamos nuestro análisis de series de tiempo.

Identificación

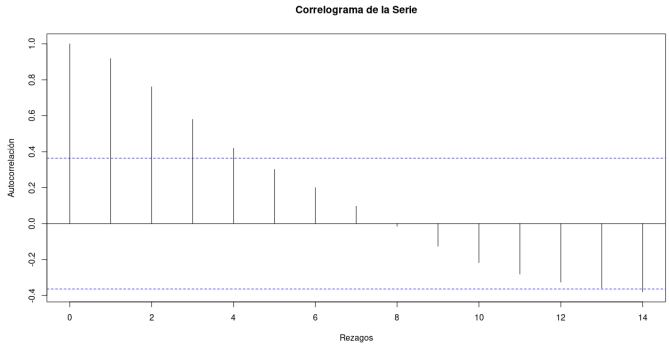
Recordemos que el primer paso en la construcción de nuestro modelo $ARIMA$ es la identificación. En esta etapa, lo primero que debemos hacer es graficar la serie de tiempo, lo cual nos permite observar el comportamiento de los casos de esquizofrenia a lo largo del periodo analizado. Esta gráfica se muestra en la **Gráfica 5.6**, donde se observa una ligera disminución en los casos de esquizofrenia alrededor del año 1999. Posteriormente, se presenta un incremento, aproximadamente, en los años 2017–2018. A partir de ese punto, los datos sugieren un nuevo descenso en los casos en los años posteriores.

Más adelante, en el análisis de nuestras predicciones, evaluaremos si esta tendencia a la baja continúa en los años siguientes o si, por el contrario, se revierte.



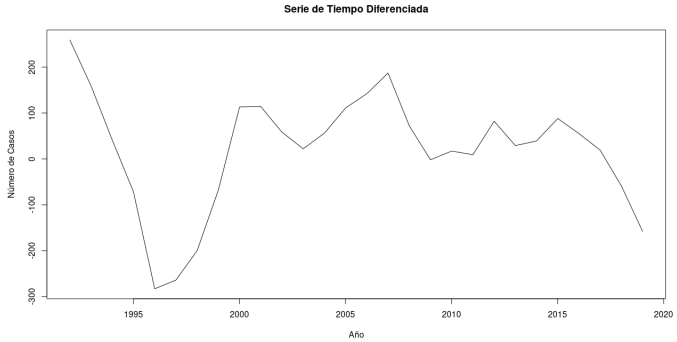
Gráfica 5.6: Serie de Tiempo de los Casos por Años

También se evaluó la estacionariedad de la serie utilizando el correlograma de errores (**Gráfica 5.7**), en el cual se observa que varios coeficientes de autocorrelación se encuentran fuera de las bandas de confianza. Esto sugiere, al menos gráficamente, que la serie de tiempo NO es estacionaria. Adicionalmente, se aplicó la prueba de Dickey-Fuller aumentada, la cual arrojó un valor p de 0.33. Este resultado no permite rechazar la hipótesis nula de no estacionariedad. Por consiguiente, fue necesario aplicar la primera diferenciación a la serie.



Gráfica 5.7: Correlograma de la Serie de los Casos por Años

En la **Gráfica 5.8** se muestra el comportamiento de nuestra serie de tiempo tras aplicar la primera diferenciación.

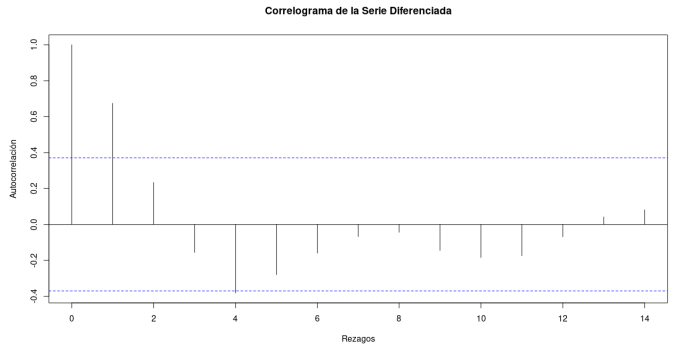


Gráfica 5.8: Gráfica de la Serie Diferenciada de los Casos por Años

Posteriormente, se generó nuevamente el correlograma de errores de la serie de tiempo, esta vez después de aplicar la diferenciación. Dicho correlograma, mostrado en la **Gráfica 5.9**, revela que, a diferencia del anterior, ahora se observan menos valores que sobrepasan las bandas de confianza, lo cual indica una posible mejora en la estacionariedad de la serie.

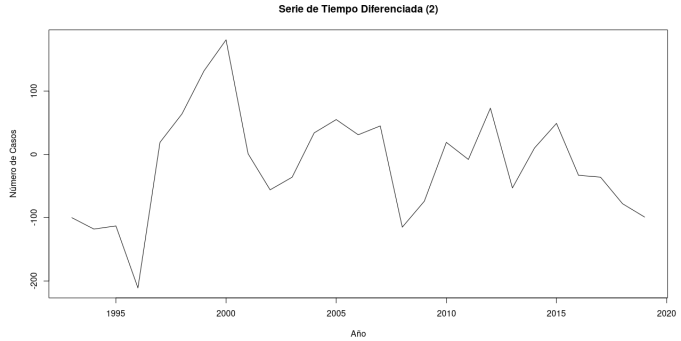
Sin embargo, la prueba de Dickey-Fuller aplicada a la serie diferenciada arrojó un p-valor de 0.4565, lo cual indica que la serie aún no es estacionaria. Por lo tanto, es necesario aplicar una segunda diferenciación para intentar alcanzar la estacionariedad.

Como nota importante, debemos tener presente que lo más recomendable es aplicar como máximo dos diferenciaciones a una serie de tiempo para lograr su estacionariedad. Aplicar más de dos diferenciaciones puede eliminar parte de la estructura original de la serie, dificultando la identificación adecuada de los componentes AR y MA del modelo.



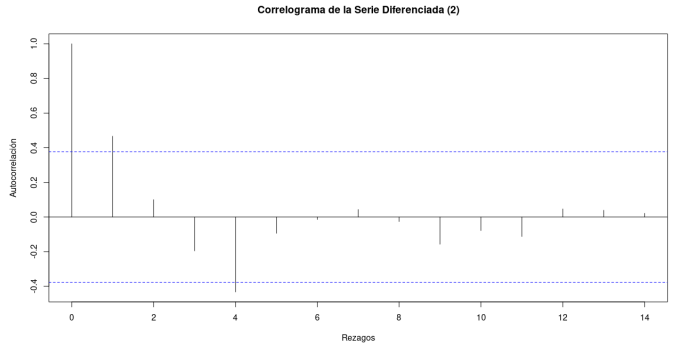
Gráfica 5.9: Correlograma de la Serie Diferenciada de los Casos por Años

A continuación, graficamos nuevamente la serie de tiempo después de aplicarle la segunda diferenciación, la cual puede apreciarse en la **Gráfica 5.10**.



Gráfica 5.10: Gráfica de la Serie Diferenciada (2) de los Casos por Años

Elaboramos el correlograma de los residuos, donde se observa que hay aún menos valores que sobrepasan las bandas de confianza (ver **Gráfica 5.11**). No obstante, para confirmar con mayor certeza que la serie es ahora estacionaria, aplicamos la prueba de Dickey-Fuller. Esta arrojó un p-valor de 0.01741, lo que nos permite rechazar la hipótesis nula de no estacionariedad. Por lo tanto, podemos concluir que la serie es estacionaria y proceder con los siguientes pasos para la construcción de nuestro modelo $ARIMA$.



Gráfica 5.11: Correlograma de la Serie Diferenciada (2) de los Casos por Años

Estimación

Se estiman ahora los parámetros:

```
ARIMA(2,2,0)

Coefficients:
         ar1      ar2
      0.5598  -0.1335
s.e.  0.1905   0.1975

sigma^2 = 5803:  log likelihood = -154.42
AIC=314.83   AICc=315.88   BIC=318.72
```

Analicemos en detalle lo que nos indican los resultados obtenidos anteriormente.

El modelo ajustado es un *ARIMA(2,2,0)* donde:

- $\boldsymbol{p = 2}$: el modelo incluye 2 términos autorregresivos (AR).
- $\boldsymbol{d = 2}$: se aplicaron dos diferenciaciones para hacer la serie estacionaria.
- $\boldsymbol{q = 0}$: no se incluyen términos de medias móviles (MA).
- $\boldsymbol{ar1 = 0.5598}$ y $\boldsymbol{ar2 = -0.1335}$: son los coeficientes de los términos autorregresivos.
- **s.e. (errores estándar)**: indican la incertidumbre de cada estimación, donde, los errores estándar son razonablemente pequeños, lo cual sugiere que los coeficientes son estadísticamente significativos.

$\boldsymbol{sigma^2 = 5803}$, es la varianza de los residuos del modelo. Menores valores suelen indicar mejor ajuste (pero deben compararse entre modelos en contextos similares).

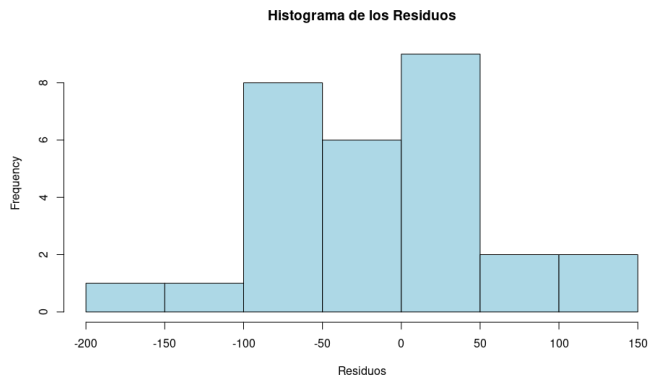
$\boldsymbol{Log - Likelihood = -154.42}$, mide qué tan bien se ajusta el modelo a los datos. Valores más cercanos a 0 indican mejor ajuste.

$\boldsymbol{AIC = 314.83}$, $\boldsymbol{AICc = 315.88}$ y $\boldsymbol{BIC = 318.72}$, estos son criterios de penalización que equilibran el ajuste del modelo y su complejidad. Valores más bajos indican un mejor modelo, en comparación con otros modelos ajustados a la misma serie.

Validación

Como primer paso en el proceso de validación del modelo, se verificó la normalidad de los residuos mediante la prueba de Jarque-Bera, la cual arrojó un p-valor de 0.8498, mayor al valor de referencia de 0.05. Este resultado nos indica que no se rechaza la hipótesis nula de normalidad en los residuos.

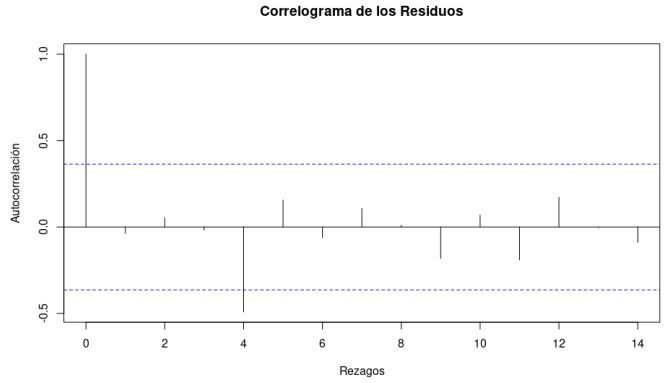
Adicionalmente, se analizó el histograma de los residuos (**Gráfica 5.12**), observándose una distribución aproximadamente simétrica y con forma de campana. Ambos elementos (principalmente la prueba de Jarque-Bera) sugieren que los residuos de nuestro modelo cumplen con el supuesto de normalidad.



Gráfica 5.12: Histograma de los Residuos del Modelo de los Casos por Años

Ahora, es necesario verificar si los residuos del modelo presentan autocorrelación. Para ello, se analizará el correlograma de los errores y, adicionalmente, se aplicará la prueba de Box-Ljung.

En la **Gráfica 5.13** se muestra el correlograma de los residuos del modelo. Se observa que la mayoría de las barras se encuentran dentro de los límites de confianza, lo que sugiere que no existe autocorrelación significativa en los residuos. Sin embargo, para tener una mayor certeza sobre este supuesto, se verificará la prueba de Box-Ljung.

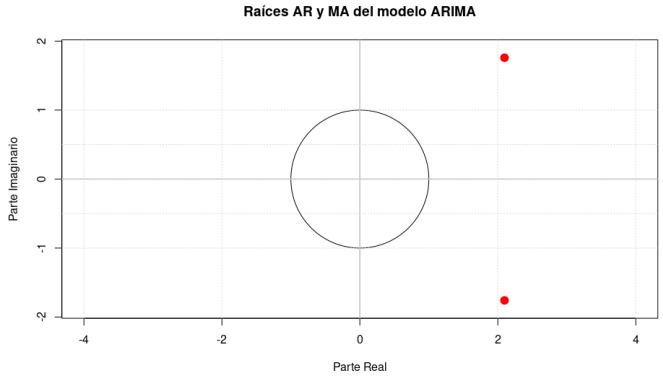


Gráfica 5.13: Correlograma del Modelo de los Casos por Años

```
    Box-Ljung test

data:  resid(modelo)
X-squared = 12.086, df = 10, p-value = 0.2794
```

Podemos observar que la prueba de Box-Ljung nos arrojó un p-value de 0.2794, lo cual indica que no existe evidencia estadísticamente significativa de autocorrelación en los residuos del modelo. Este resultado, junto con el análisis visual del correlograma de los errores, permite concluir que los residuos se comportan como ruido blanco, lo cual confirma que el modelo $ARIMA$ ha sido adecuadamente ajustado a los datos.

Para finalizar el proceso de validación, se realizó la gráfica de las raíces unitarias del modelo, que se muestra en la **Gráfica 5.14**. En ella, podemos observar que nuestro modelo es estacionario (no anticipante), ya que todas las raíces unitarias se encuentran fuera del círculo unitario. Sin embargo, no podemos concluir si el modelo es invertible o no, ya que no se incluyen raíces MA en nuestro análisis.



Gráfica 5.14: Raíces Unitarias del Modelo de los Casos por Años

Una vez verificado que nuestro modelo $ARIMA$ es el más eficaz para realizar la predicción que buscábamos, surge la pregunta: ¿cómo se representa matemáticamente este modelo?

En primer lugar, tenemos nuestro modelo $ARIMA(2,2,0)$, esto significa que:

- $p = 2$: parte autorregresiva (AR) de orden 2,
- $d = 2$: dos diferenciaciones (para lograr estacionariedad),
- $q = 0$: sin términos de medias móviles (MA).

Donde los coeficientes dados son:

- $\phi_1 = 0.5598 (ar1)$,
- $\phi_2 = -0.1335 (ar2)$.

Ahora, tenemos que la forma general del $ARIMA(2,2,0)$ es (Ver ecuación 4.1)):

$$(1 - \phi_1 L - \phi_2 L^2)(1 - 2L + L^2) y_t = a_t.$$

Donde:

- $\mathbf{(1 - 2L + L^2)}$ representa **diferenciar la serie dos veces** (hacerla estacionaria).
- $\mathbf{(1 - \phi_1 L - \phi_2 L^2)}$ representa la **parte AR** que modela la serie estacionaria.

Primero expandimos $(1 - 2L + L^2) y_t$ (Ver ecuación (4.4)), que es:

$$(1 - 2L + L^2) y_t = y_t - 2y_{t-1} + y_{t-2}.$$

Ahora aplicamos al operador AR:

$$(1 - \phi_1 L - \phi_2 L^2)(y_t - 2y_{t-1} + y_{t-2})$$

Expandimos:

$$y_t - 2y_{t-1} + y_{t-2} - \phi_1 (y_{t-1} - 2y_{t-2} + y_{t-3}) - \phi_2 (y_{t-2} - 2y_{t-3} + y_{t-4}).$$

Agrupamos:

$$y_t - (2+\phi_1)y_{t-1} + (1+2\phi_1-\phi_2)y_{t-2} - (\phi_1 - 2\phi_2)y_{t-3} - \phi_2 y_{t-4} = a_t$$

Ahora, sustituimos los valores de nuestros coeficientes:

$$y_t - (2+(0.5598))y_{t-1} + (1+2(0.5598) - (-0.1335))y_{t-2} - ((0.5598) - 2(-0.1335))y_{t-3} - (-0.1335)y_{t-4} = a_t$$

Por lo que tenemos que:

$$y_t - 2.5598y_{t-1} + 2.2531y_{t-2} - 0.8268y_{t-3} + 0.1335y_{t-4} = a_t.$$

Despejamos y_t para obtener la ecuación deseada, resultando en:

$$\boldsymbol{y_t = 2.5598y_{t-1} - 2.2531y_{t-2} + 0.8268y_{t-3} - 0.1335y_{t-4} + a_t},$$

con a_t un proceso de ruido blanco.

Predicción

Finalmente, realizamos la predicción utilizando nuestro modelo $ARIMA$, la cual se presenta en la **Figura 5.15**, junto con la gráfica correspondiente de dicha predicción.

	Point Forecast	Lo 95	Hi 95
2020	5815.992	5666.69124	5965.293
2021	5601.004	5190.69705	6011.311
2022	5385.317	4623.47296	6147.161
2023	5170.838	3987.45791	6354.217
2024	4957.128	3294.60673	6619.649
2025	4743.688	2551.67293	6935.703
2026	4530.296	1763.07077	7297.521
2027	4316.895	932.09836	7701.692
2028	4103.483	61.41698	8145.548
2029	3890.065	-846.74667	8626.876

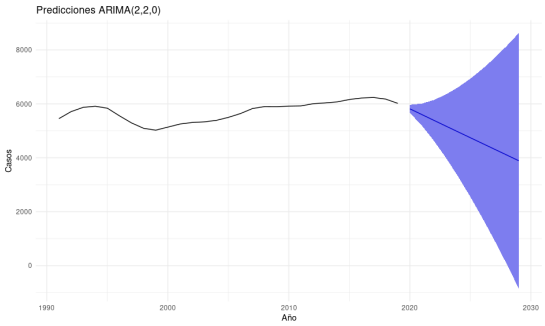


Figura 5.15: Predicción de los Casos por Años

Se presentan los pronósticos para los años 2020 a 2029 generados por el modelo ARIMA ajustado. La columna "Point Forecast" muestra el valor esperado para cada año, mientras que las columnas "Lo 95" y "Hi 95" corresponden a los límites inferior y superior del intervalo de confianza al 95 %, respectivamente.

También se observa que los intervalos de confianza se amplían conforme se avanza en el horizonte de predicción, lo que refleja un aumento en la incertidumbre asociada a las estimaciones futuras. Este comportamiento es característico de los modelos de series de tiempo y sugiere que las predicciones a corto plazo tienden a ser más confiables que aquellas realizadas a largo plazo.

Por ejemplo, para el año 2020, el valor pronosticado es 5,816 casos, con un intervalo de confianza entre 5,667 y 5,965, mientras que para el año 2029 el pronóstico disminuye a 3,890 casos, pero con un intervalo mucho más amplio, que va desde -847 hasta 8,627 casos, indicando mayor incertidumbre.

Finalmente, con el objetivo de evaluar la confiabilidad de las predicciones generadas por el modelo, se calcularon los errores de pronóstico comparando los últimos diez valores observados de la serie (correspondientes al periodo 2010–2019) con los valores pronosticados para ese mismo intervalo.

```
[1] "MAE: 710.527947370463"
[1] "RMSE: 772.063836151504"
[1] "MAPE: 13.1437925284302"
```

¿Qué nos indican estos resultados?

En primer lugar, el **MAE (Error Absoluto Medio)** señala que, en promedio, el modelo se equivoca por unas 710 unidades respecto al valor real de la serie. Es una medida simple y fácil de interpretar porque mantiene las mismas unidades que los datos originales.

Por otro lado, el **RMSE (Error Cuadrático Medio)** penaliza más fuertemente los errores grandes debido a la elevación al cuadrado. El modelo, en promedio, tiene un error cuadrático de aproximadamente 772 unidades.

Finalmente, el **MAPE (Error Porcentual Absoluto Medio)** revela que, en promedio, las predicciones del modelo se desvían un 13.14 % de los valores reales. De acuerdo con criterios usuales de interpretación del MAPE, un valor entre 10 % y 20 % se considera indicativo de una capacidad predictiva razonablemente buena. Este resultado sugiere que el modelo ofrece un nivel de precisión adecuado para propósitos descriptivos y de pronóstico a corto plazo.

5.2.2. Estudio de los Casos por Sexo

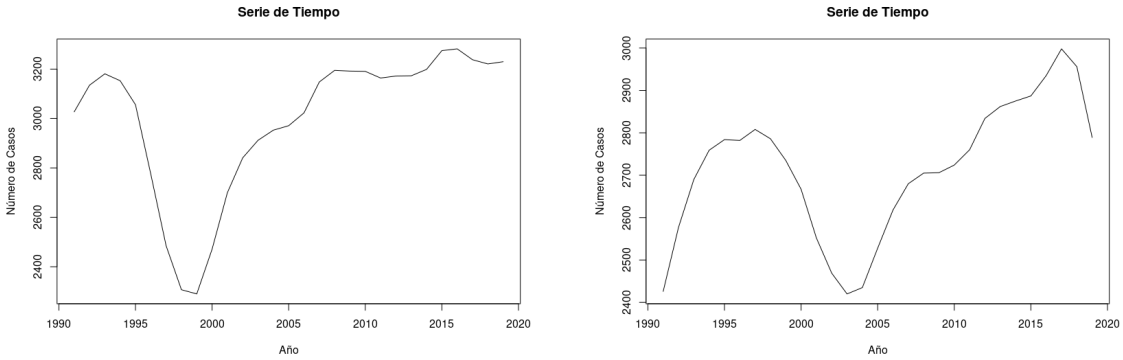
Otra manera de estudiar nuestros datos es a través del sexo. Afortunadamente, en la página donde se descargó la base de datos, había una opción para visualizar los valores desglosados por sexo, ya sea femenino o masculino. Este análisis nos permitirá determinar si hay una mayor prevalencia de esquizofrenia entre hombres o mujeres. La tabla con estos valores se muestra en la **Figura 5.16**, los cuales fueron ajustados de la misma manera que los casos por años para la elaboración de los modelos.

Año	Femenino	Masculino
1990	NA	NA
1991	2426	3027
1992	2577	3135
1993	2690	3181
1994	2759	3153
1995	2784	3056
1996	2782	2775
1997	2808	2485
1998	2786	2307
1999	2735	2290
2000	2667	2471
2001	2552	2700
2002	2469	2841
2003	2420	2912
2004	2435	2953
2005	2528	2971
2006	2618	3023
2007	2680	3148
2008	2705	3195
2009	2706	3192
2010	2724	3191
2011	2760	3164
2012	2834	3172
2013	2862	3173
2014	2875	3199
2015	2887	3275
2016	2935	3282
2017	2998	3238
2018	2956	3221
2019	2789	3230

Figura 5.16: Tabla de Casos por Sexo

Identificación

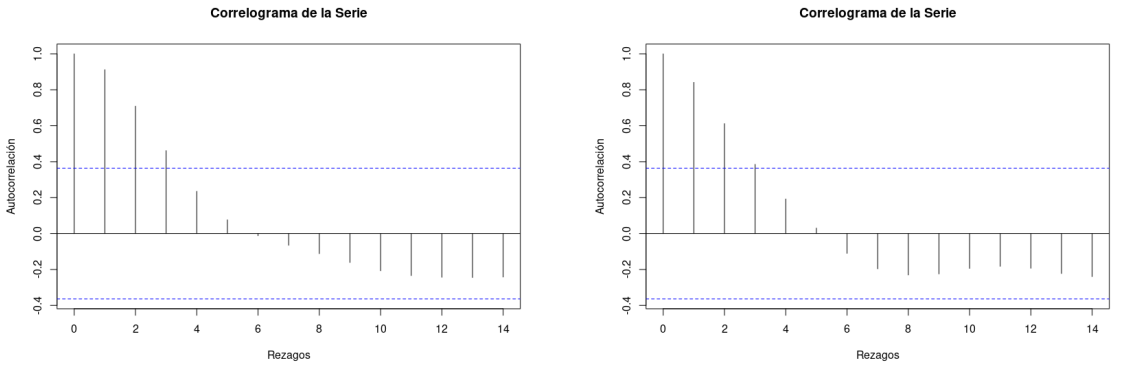
Primero, se generan las gráficas de las series de tiempo, las cuales se presentan en la **Gráfica 5.17**. En ellas se observa el comportamiento de los casos de esquizofrenia en el sexo masculino y femenino, respectivamente. A simple vista, ambas gráficas presentan tendencias similares, con caídas en determinados periodos de tiempo. En la serie de casos masculinos, la disminución se observa entre 1995 y 2000, mientras que en la serie de casos femeninos ocurre entre 2000 y 2005. Posteriormente, ambas series muestran un incremento en los casos, aunque en la serie femenina se aprecia una nueva disminución entre los años 2015 y 2019.



(a) Masculino (b) Femenino

Gráfica 5.17: Serie de Tiempo de los Casos por Sexo

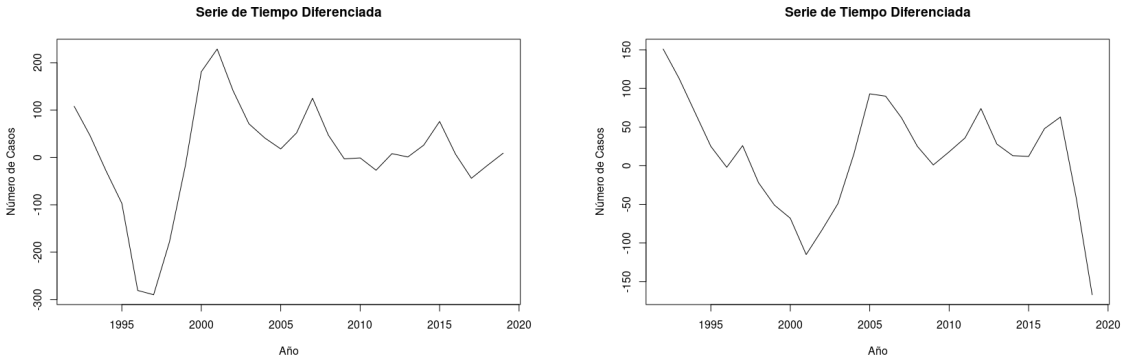
Ahora, para verificar la estacionariedad de las series de tiempo, se aplicó la prueba de Dickey-Fuller y se generó el correlograma de errrores (**Gráfica 5.18**) para cada una de las series de tiempo. Ambos métodos indicaron que las series no eran estacionarias, por lo cual fue necesario aplicar una diferenciación en ambas series de tiempo.



(a) Masculino (b) Femenino

Gráfica 5.18: Correlograma de la Serie de los Casos por Sexo

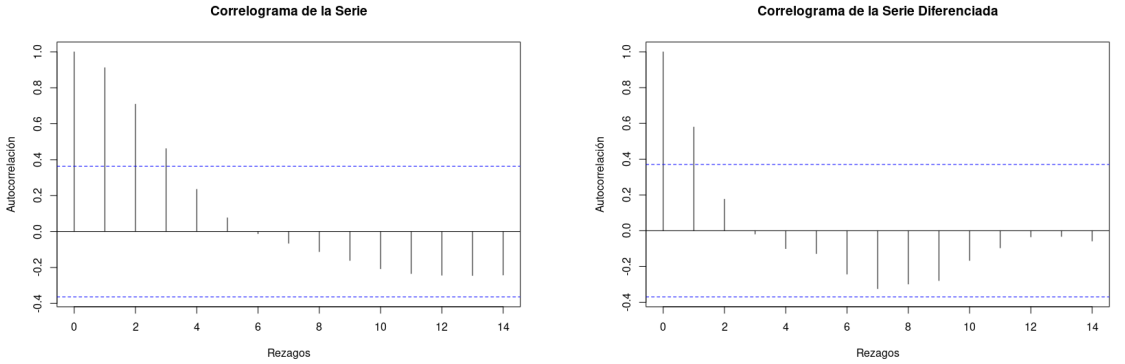
La **Gráfica 5.19** muestra las series de tiempo resultantes tras aplicar una diferenciación a ambas, con el fin de verificar si se alcanzó la estacionariedad esperada.



(a) Masculino (b) Femenino

Gráfica 5.19: Serie de Tiempo Diferenciada de los Casos por Sexo

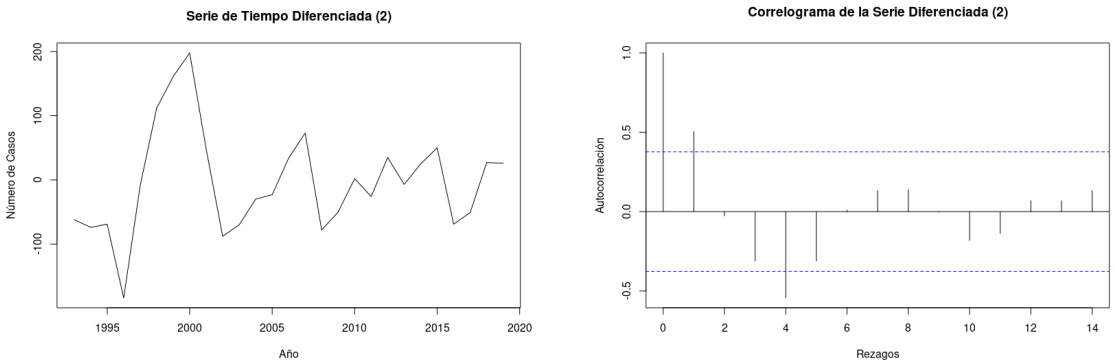
La **Gráfica 5.20** muestra los correlogramas de las series de tiempo correspondientes a los casos de esquizofrenia en el sexo masculino y femenino respectivamente. Según los resultados de las pruebas de Dickey-Fuller aplicadas a ambas series, se concluye que la serie de casos femeninos ya cumple con la condición de estacionariedad. En cambio, la serie de casos masculinos aún no la alcanza, por lo que será necesario aplicar una segunda diferenciación a esta última.



(a) Masculino (b) Femenino

Gráfica 5.20: Correlograma de la Serie Diferenciada de los Casos por Sexo

Se aplicó una segunda diferenciación a la serie de tiempo de casos masculinos, cuya gráfica resultante se muestra en la **Gráfica 5.21**, junto con su correlograma de errores. Tras esta transformación, y al obtener un p-valor menor a 0.05 en la prueba de Dickey-Fuller, podemos concluir que la serie de tiempo ha alcanzado la estacionariedad deseada.



(a) Gráfica (b) Correlograma

Gráfica 5.21: Segunda Diferenciación de la Serie de los Casos por Sexo Masculino

Estimación

```
ARIMA(2,2,0)

Coefficients:
         ar1      ar2
      0.6808  -0.3589
s.e.  0.1757   0.1740

sigma^2 = 4336:  log likelihood = -150.61
AIC=307.23   AICc=308.27   BIC=311.11
```

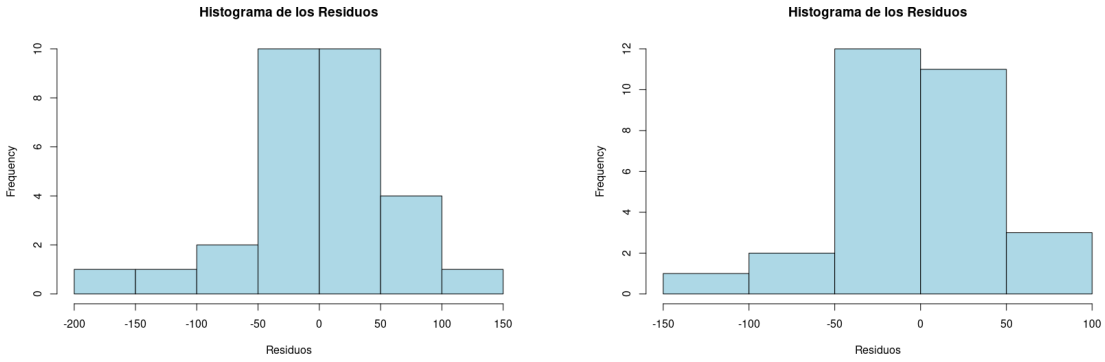
El modelo resentado corresponde a los casos de esquizofrenia en hombres y es un $ARIMA(2,2,0)$, este parece ajustarse razonablemente bien tras aplicar dos diferenciaciones, con coeficientes significativos y métricas aceptables. Se puede comparar con otros modelos (por ejemplo, con menos o más parámetros) para verificar si éste es el más adecuado.

```
ARIMA(1,1,0)

Coefficients:
         ar1
      0.8620
s.e.  0.1232

sigma^2 = 2240:  log likelihood = -147.9
AIC=299.8   AICc=300.28   BIC=302.47
```

El modelo anterior corresponde al análisis de los casos de esquizofrenia en mujeres. El modelo $ARIMA(1,1,0)$ indica que, tras aplicar una primera diferenciación, la serie se volvió estacionaria. El modelo obtenido parece razonablemente adecuado, sin embargo, podemos evaluar su desempeño en relación con otros modelos, considerando los valores de **AIC, AICc** y **BIC**, recordando que, en general, valores más bajos sugieren un mejor ajuste.

Validación

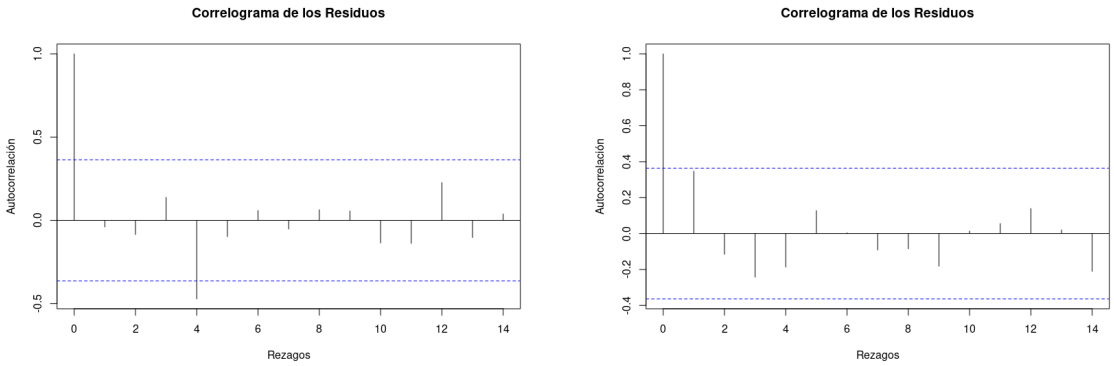
Como primer paso en la validación de nuestros modelos, se realizaron pruebas de normalidad aplicando el test de Jarque-Bera y analizando los histogramas de los residuos, los cuales se muestran en la **Gráfica 5.22**.



(a) Masculino (b) Femenino

Gráfica 5.22: Histograma de los Residuos del Modelo de los Casos por Sexo

No obstante, dado que los histogramas no permiten confirmar con total certeza el cumplimiento del supuesto de normalidad, se recurrió a las pruebas de Jarque-Bera. En el caso del modelo de casos masculinos, se obtuvo un p-valor de 0.6256, mientras que en el modelo de casos femeninos el p-valor fue de 0.4481. Dado que ambos valores p son suficientemente altos, no se rechaza la hipótesis nula de normalidad, por lo que se concluye que ambos modelos (tanto el de casos masculinos como el de casos femeninos) cumplen con el supuesto de normalidad.



(a) Masculino (b) Femenino

Gráfica 5.23: Correlograma del Modelo de los Casos por Sexo

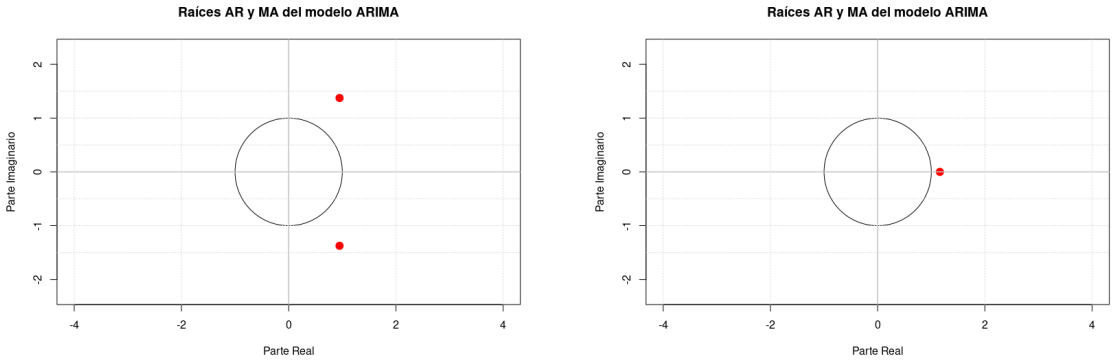
Como siguiente paso en el proceso de verificación, se graficaron los correlogramas de los errores de nuestros modelos, los cuales se presentan en la **Gráfica 5.23**.

En ambas gráficas podemos observar que no hay muchos valores que sobrepasen los límites de las bandas de confianza. Sin embargo, para tener una evaluación más precisa del cumplimiento

de los supuestos, se aplicará la prueba de Box-Ljung.

En la prueba de Box-Ljung aplicada al modelo de casos masculinos se obtuvo un p-valor de 0.3763, mientras que en el modelo de casos femeninos el p-valor fue de 0.4111. Por lo tanto, se concluye que no existe autocorrelación significativa en los residuos de ambos modelos.

Como paso final en el proceso de validación, se calcularon y graficaron las raíces unitarias de ambos modelos. Ahora, dado que no se incluyeron términos MA, solo se consideraron las raíces correspondientes a los componentes autorregresivos (AR). Las gráficas resultantes se pueden visualizar en la **Gráfica 5.24**.



(a) Masculino (b) Femenino

Gráfica 5.24: Raíces Unitarias del Modelo de los Casos por Sexo

En ambas gráficas se observa que las raíces unitarias correspondientes a los términos AR se encuentran fuera del círculo unitario, es decir, sus módulos son mayores que 1. Esto indica que ambos modelos cumplen con la condición de estacionariedad.

Finalmente, veamos la representación matemática de nuestros modelos $ARIMA$:

Masculino

Un modelo $ARIMA(2,2,0)$ se puede escribir como (Ver ecuación 4.1)):

$$(1-\phi_1 L-\phi_2 L^2)(1-L)^2 y_t = a_t.$$

Sustituimos los valores de ϕ_1 y ϕ_2:

$$(1-(0.6808)L-(-0.3589)L^2)(1-L)^2 y_t = a_t.$$

Veamos que $(1-L)^2 y_t$ es (Ver ecuación 4.4)):

$$(1-L)^2 y_t = y_t - 2y_{t-1} + y_{t-2},$$

llamamos a esto W_t, es decir:

$$W_t = y_t - 2y_{t-1} + y_{t-2}. \tag{5.1}$$

Recordemos que $AR(2)$ es:

$$(1-\phi_1 L-\phi_2 L^2). \tag{5.2}$$

Multiplicamos (5.1) por (5.2) y obtenemos:

$$W_t - 0.6808W_{t-1} + 0.3589W_{t-2} = a_t.$$

Sustituimos W_t, W_{t-1}, W_{t-2} por sus expresiones en términos de y_t:

$$y_t - 2y_{t-1} + y_{t-2} - 0.6808(y_{t-1} - 2y_{t-2} + y_{t-3}) + 0.3589(y_{t-2} - 2y_{t-3} + y_{t-4}) = a_t.$$

Distribuimos y agrupamos:

$$y_t - 2.6808y_{t-1} + 2.7205y_{t-2} - 1.3986y_{t-3} + 0.3589y_{t-4} = a_t.$$

Finalmente, despejamos y_t:

$$\boldsymbol{y_t = 2.6808y_{t-1} - 2.7205y_{t-2} + 1.3986y_{t-3} - 0.3589y_{t-4} + a_t},$$

con a_t un proceso de ruido blanco.

Femenino

Un modelo $ARIMA(1,1,0)$ se puede escribir como (Ver ecuación (4.1)):

$$(1 - \phi_1 L)(1 - L)y_t = a_t.$$

Sustituimos el valor de ϕ_1:

$$(1 - (0.8620)L)(1 - L)y_t = a_t.$$

Veamos que:

$$(1 - 0.8620L)(y_t - y_{t-1}) = a_t.$$

Aplicamos multiplicación de polinomios:

$$(y_t - y_{t-1}) - 0.8620(y_{t-1} - y_{t-2}) = a_t.$$

Ahora expandimos y agrupamos:

$$y_t - 1.8620y_{t-1} + 0.8620y_{t-2} = a_t.$$

Finalmente, despejamos y_t:

$$\boldsymbol{y_t = 1.8620y_{t-1} - 0.8620y_{t-2} + a_t},$$

con a_t un proceso de ruido blanco.

Predicción

En la **Figura 5.25** se muestra la predicción para los casos de esquizofrenia en hombres. El modelo proyecta un aumento leve y continuo en los valores futuros entre 2020 y 2029, mostrando para cada año un intervalo de confianza del 95 % que nos indica el rango probable donde se ubicarán los valores reales; estos intervalos se amplían con el tiempo, reflejando la creciente incertidumbre en las predicciones a medida que se avanza hacia el futuro, lo que implica que aunque se espera un crecimiento sostenido, la precisión disminuye en los años más lejanos.

	Point Forecast	Lo 95	Hi 95
2020	3247.011	3117.94450	3376.077
2021	3260.145	2890.85572	3629.434
2022	3267.764	2583.18284	3952.346
2023	3273.021	2231.80323	4314.239
2024	3278.648	1853.86261	4703.434
2025	3285.375	1451.26649	5119.484
2026	3292.719	1020.82512	5564.612
2027	3300.087	560.58991	6039.583
2028	3307.250	71.17426	6543.326
2029	3314.266	-445.48064	7074.012

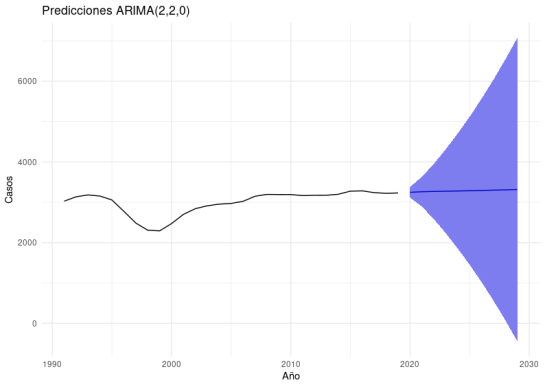


Figura 5.25: Predicción de los Casos por Sexo (Masculino)

Las predicciones realizadas con el modelo para los casos de esquizofrenia en mujeres se muestran en la **Figura 5.26**. El modelo muestra que los valores tienden a bajar con el tiempo. Los intervalos de confianza se hacen más amplios, lo que significa que la predicción es menos precisa conforme pasan los años. En resumen, se espera una disminución en los casos de esquizofrenia, pero con mayor incertidumbre en el futuro.

	Point Forecast	Lo 95	Hi 95
2020	2645.051	2552.2781	2737.823
2021	2520.971	2324.8943	2717.047
2022	2414.017	2102.8095	2725.224
2023	2321.826	1888.8106	2754.841
2024	2242.360	1684.1332	2800.587
2025	2173.863	1489.1834	2858.542
2026	2114.820	1303.9109	2925.728
2027	2063.926	1128.0103	2999.843
2028	2020.058	961.0358	3079.080
2029	1982.244	802.4714	3162.017

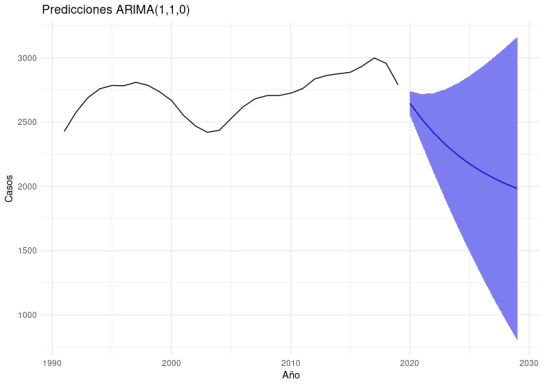


Figura 5.26: Predicción de los Casos por Sexo (Femenino)

Finalmente, calculamos los errores de las predicciones para evaluar la confiabilidad de los resultados.

```
[1] "MAE: 494.628560978403"
[1] "RMSE: 614.864168300565"
[1] "MAPE: 19.7570131565326"
```

Se observa que el modelo para el sexo masculino presenta un error moderado, con un margen de error aceptable pero no pequeño. Mientras que el MAPE indica que el modelo se equivoca en un 19.76 % respecto al valor real, lo que refleja un error relativo moderado.

```
[1] "MAE: 495.296620187283"
[1] "RMSE: 544.383766859766"
[1] "MAPE: 18.1043103885151"
```

Por otro lado, el modelo ajustado a la serie temporal de los casos de esquizofrenia en el sexo femenino presenta métricas que nos sugieren un error moderado. En particular, el MAPE es del 18.1 %, lo que implica que, en promedio, el modelo se desvía un 18.1 % respecto a los valores reales.

5.2.3. Estudio de los Valores por Edad

Una estrategia complementaria para el análisis de la base de datos consiste en construir modelos de predicción segmentados por rangos de edad. La base original clasifica las edades en los siguientes intervalos: **0-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54** y **55+ Años**. Ajustar un modelo ARIMA individual para cada uno de estos diez rangos resultaría poco práctico y dificultaría la interpretación comparativa de los resultados. Por ello, se decidió reducir la cantidad de grupos etarios mediante la agrupación de cada dos intervalos consecutivos, lo que permite trabajar con únicamente cinco modelos. Los nuevos rangos definidos son: **0–19, 20–29, 30–39, 40–49** y **50+ años**.

En la **Figura 5.27** se muestra la tabla con los valores correspondientes a los nuevos rangos de edad que utilizaremos para ajustar nuestros modelos.

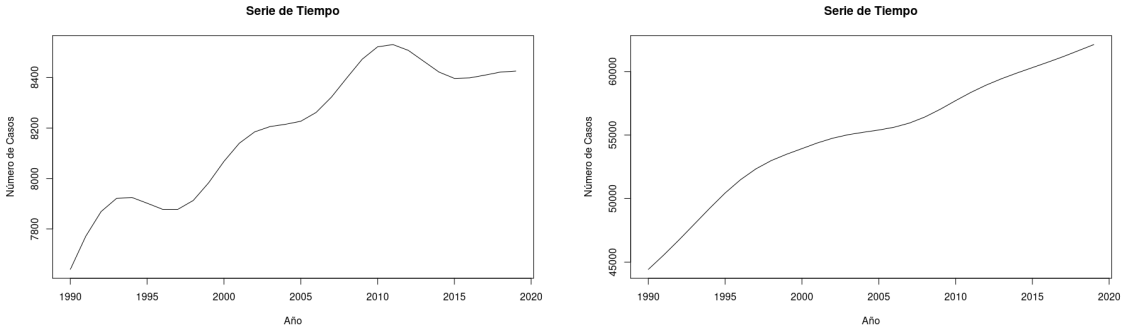
Años	0-19 Años	20-29 Años	30-39 Años	40-49 Años	50+ Años
1990	7640	44430	50850	33714	31275
1991	7771	45558	52907	34850	32276
1992	7869	46754	55037	36074	33340
1993	7921	47998	57171	37394	34461
1994	7925	49247	59235	38822	35628
1995	7902	50442	61163	40367	36823
1996	7878	51486	62904	41859	38127
1997	7878	52337	64507	43340	39485
1998	7913	52991	66022	44832	40882
1999	7983	53490	67501	46435	42256
2000	8068	53933	68990	48298	44059
2001	8140	54378	70497	49989	45051
2002	8185	54752	72043	51683	46702
2003	8206	55026	73637	53364	48464
2004	8215	55219	75262	55036	50353
2005	8227	55396	76891	56703	52367
2006	8262	55615	78462	58431	54455
2007	8323	55946	79870	60247	56667
2008	8399	56420	81054	62115	58965
2009	8473	57028	82026	63991	61333
2010	8522	57707	82911	65835	63791
2011	8531	58359	83838	67642	66320
2012	8508	58936	84790	69495	68967
2013	8465	59437	85726	71399	71704
2014	8422	59885	86636	73337	74525
2015	8397	60308	87537	75276	77449
2016	8399	60730	88432	77175	80448
2017	8410	61170	89386	78906	83548
2018	8422	61643	90436	80377	86719
2019	8426	62121	91587	81576	89906

Figura 5.27: Tabla de Casos por Edad

Es importante recordar que, al agrupar los datos por año y por género, se observaron tendencias crecientes, ya que los registros tienden a acumularse a lo largo del tiempo. Esto justificó el uso de diferencias temporales entre años para analizar la evolución en el número de casos documentados. Sin embargo, cuando los datos se agrupan por rangos de edad, este comportamiento no se presenta de la misma forma, ya que no se trata de una secuencia temporal, sino de categorías demográficas independientes. Cada grupo etario representa una población distinta, por lo que no es adecuado aplicar un análisis de diferencias como si existiera una progresión cronológica entre ellos. Por esta razón, los valores por edad muestran un comportamiento distinto, concentrándose los casos de esquizofrenia en ciertos grupos específicos, lo cual contrasta con las tendencias observadas en los análisis por año y por sexo.

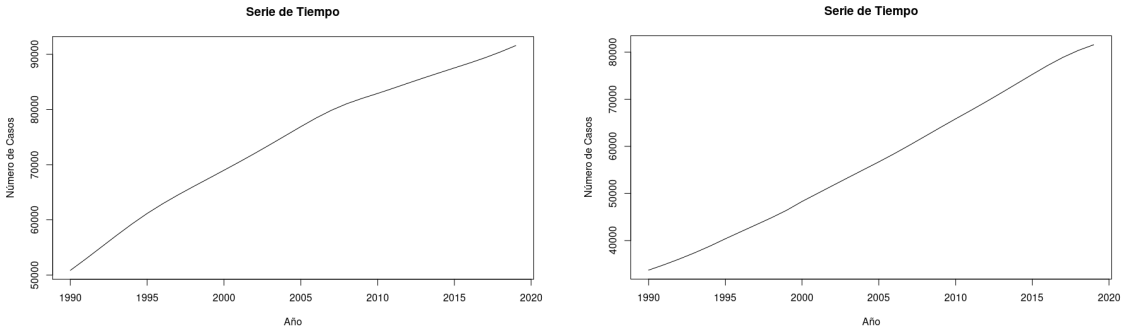
Identificación

Como parte del proceso de identificación, se graficaron las series de tiempo. Estas gráficas permiten observar cómo ha sido la evolución de los casos de esquizofrenia a lo largo de los años, según los diferentes grupos de edades. Estas se presentan en las **Gráficas 5.28, 5.29** y **5.30**.



(a) 0-19 Años (b) 20-29 Años

Gráfica 5.28: Serie de Tiempo de los Casos por Edad



(a) 30-39 Años (b) 40-49 Años

Gráfica 5.29: Serie de Tiempo de los Casos por Edad

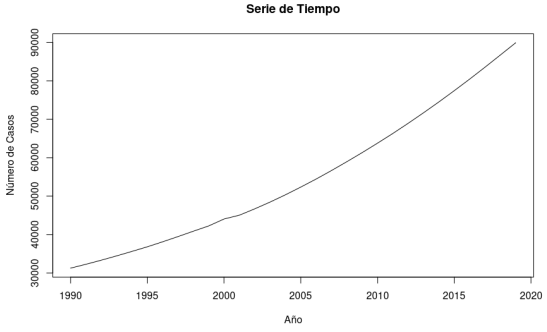
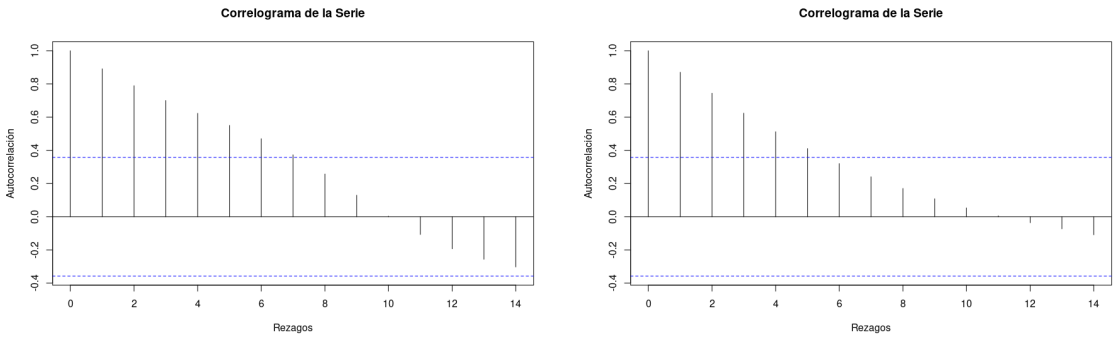


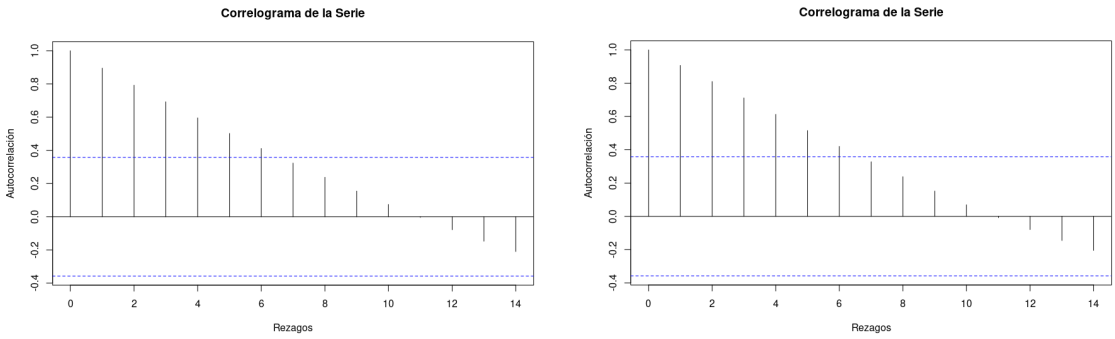
Figura 5.30: Serie de Tiempo de los Casos por Edad (50+ Años)

Como siguiente paso en la identificación, realizamos pruebas de estacionariedad utilizando la prueba de Dickey-Fuller y los correlogramas de todas las series de tiempo, en las cuales podemos afirmar que, cuanto menos valores superen la línea de confianza, mayor será la evidencia de que nuestra serie cumple con la estacionariedad. Estos correlogramas se presentan en las **Gráficas 5.31, 5.32** y **5.33**.



(a) 0-19 Años (b) 20-29 Años

Gráfica 5.31: Correlograma de la Serie de los Casos por Edad



(a) 30-39 Años (b) 40-49 Años

Gráfica 5.32: Correlograma de la Serie de los Casos por Edad

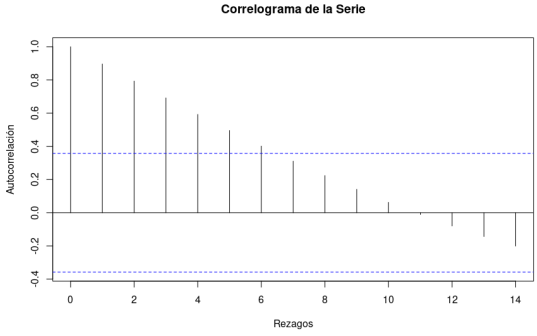
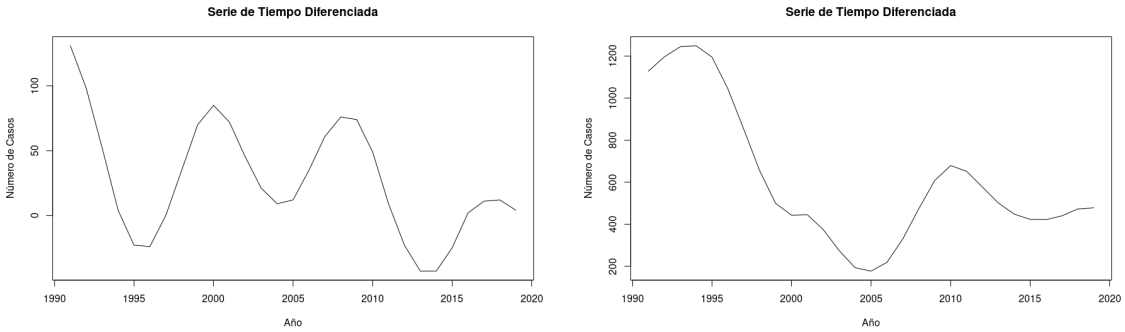


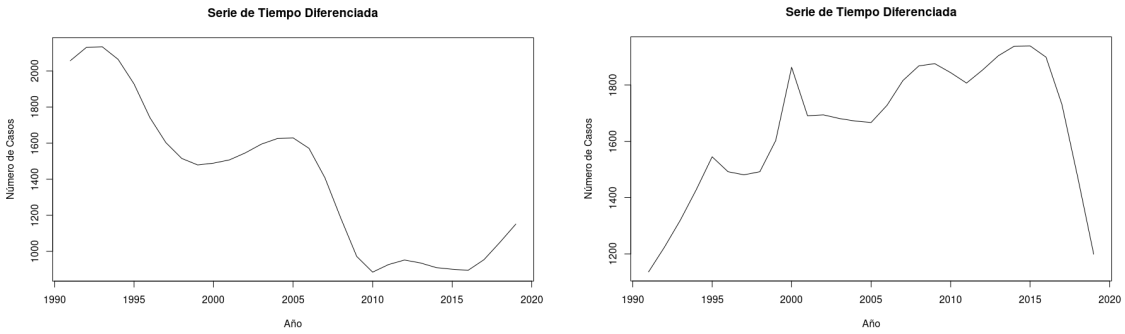
Figura 5.33: Correlograma de la Serie de los Casos por Edad (50+ Años)

Dado que todas nuestras series de tiempo requirieron una diferenciación para alcanzar la estacionariedad, en las **Gráficas 5.34, 5.35** y **5.36** se presentan las gráficas de las series resultantes luego de aplicar dicha transformación.



(a) 0-19 Años (b) 20-29 Años

Gráfica 5.34: Serie de Tiempo Diferenciada de los Casos por Edad



(a) 30-39 Años (b) 40-49 Años

Gráfica 5.35: Serie de Tiempo Diferenciada de los Casos por Edad

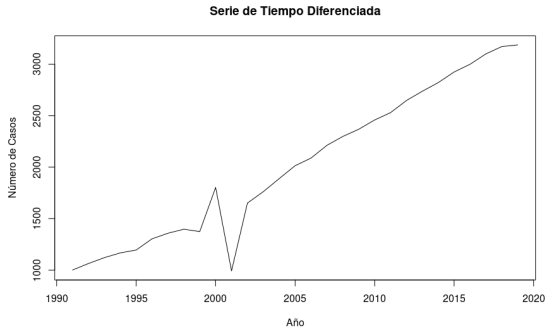
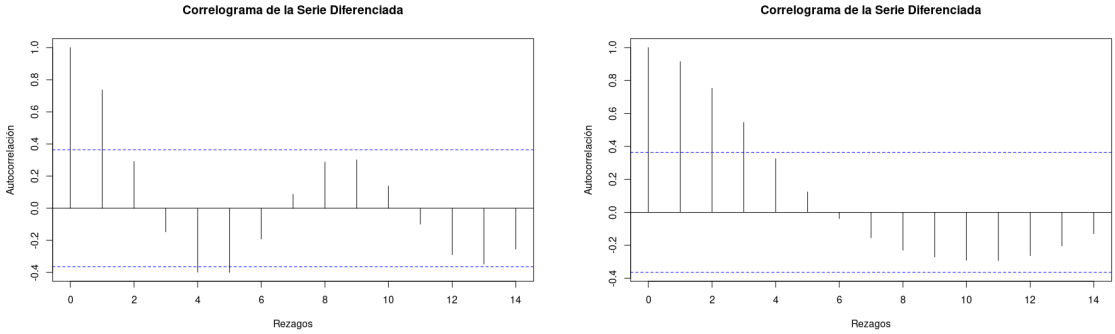


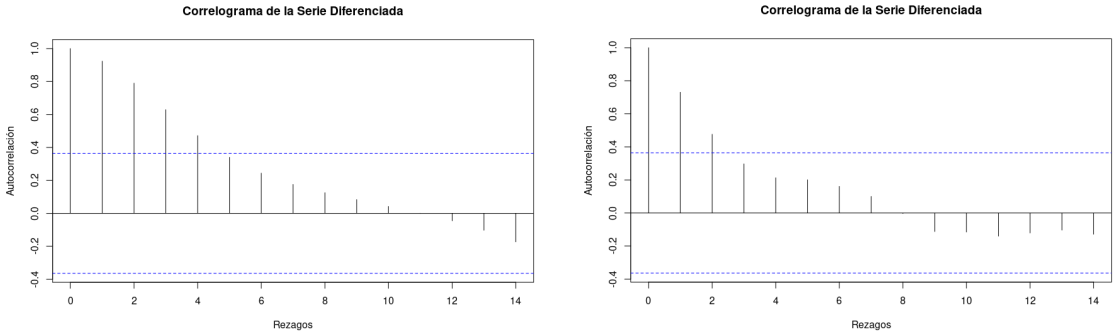
Figura 5.36: Serie de Tiempo Diferenciada de los Casos por Edad (50+ Años)

Ahora, en las **Gráficas 5.37, 5.38** y **5.39** se muestran los correlogramas correspondientes a las series diferenciadas.



(a) 0-19 Años (b) 20-29 Años

Gráfica 5.37: Correlograma de la Serie Diferenciada de los Casos por Edad



(a) 30-39 Años (b) 40-49 Años

Gráfica 5.38: Correlograma de la Serie Diferenciada de los Casos por Edad

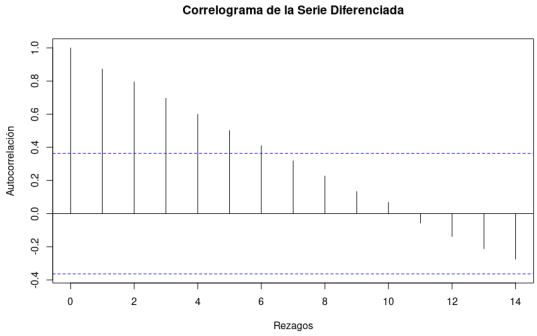
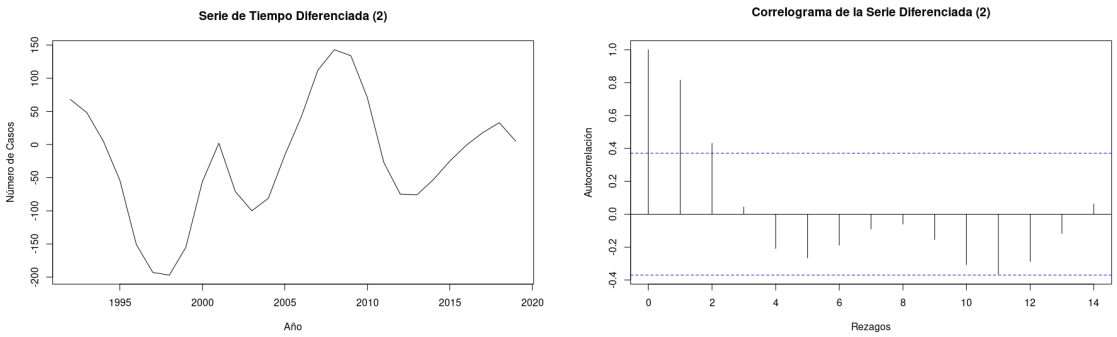


Figura 5.39: Correlograma de la Serie Diferenciada de los Casos por Edad (50+ Años)

A simple vista, los correlogramas de los errores de nuestras series no son suficientes para afirmar con total certeza que se cumple el criterio de estacionariedad, por ello, fue necesario aplicar la prueba de Dickey-Fuller. Los resultados de esta prueba indicaron que las series correspondientes a los rangos de edad **0-19** y **40-49 Años** ya cumplen con la estacionariedad requerida. En cambio, las series de los rangos **20-29, 30-39** y **50+ Años** aún no presentan estacionariedad, por lo que será necesario aplicar una segunda diferenciación en estos casos.

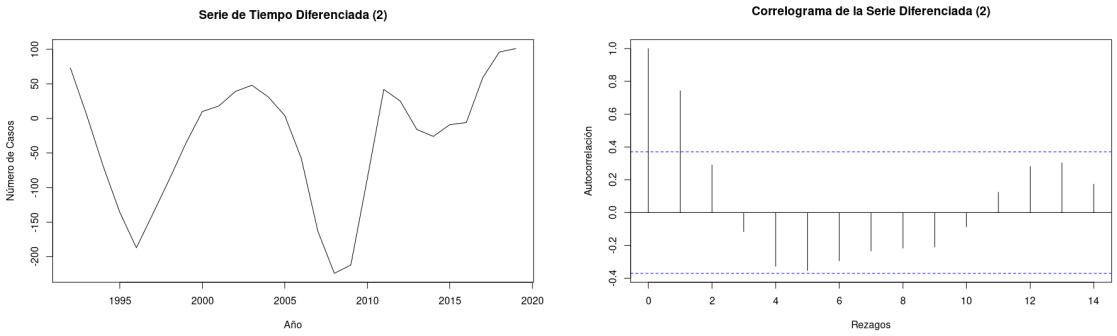
En la **Gráfica 5.40** se muestra la gráfica de la serie y su correlograma correspondientes a los casos de esquizofrenia en el grupo de **20-29 Años**, después de aplicar la segunda diferenciación. Donde, dado que el p-valor de la prueba de Dickey-Fuller resultó ser menor a 0.05, se concluye que la serie ahora si cumple con la condición de estacionariedad.



(a) Gráfica (b) Correlograma

Gráfica 5.40: Segunda Diferenciación de la Serie de los Casos por Edad (20-29 Años)

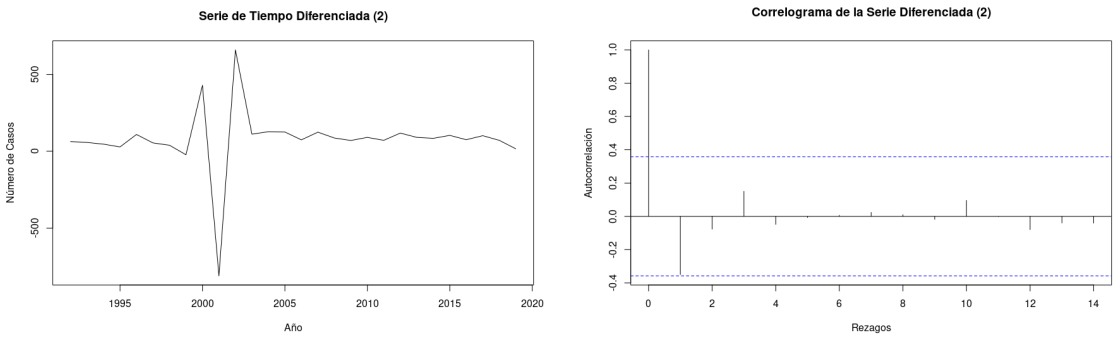
También se realiza una segunda diferenciación a la serie de tiempo de los casos de esquizofrenia en el grupo de **30-39 Años**, cuya gráfica y correlograma se muestran en la **Gráfica 5.41**. Luego de esta transformación, se aplica nuevamente la prueba de Dickey-Fuller, la cual confirma que la serie ahora sí es estacionaria.



(a) Gráfica (b) Correlograma

Gráfica 5.41: Segunda Diferenciación de la Serie de los Casos por Edad (30-39 Años)

Finalmente, se aplicó la segunda diferenciación a la serie de casos de esquizofrenia en el grupo de **50+ Años**, cuya gráfica y correlograma se presentan en la **Gráfica 5.42**. Tras realizar la prueba de Dickey-Fuller, se confirmó que la serie logró alcanzar la estacionariedad.



(a) Gráfica (b) Correlograma

Gráfica 5.42: Segunda Diferenciación de la Serie de los Casos por Edad (50+ Años)

Y, una vez comprobada la estacionariedad de nuestras series de tiempo, podemos proceder al siguiente paso en la construcción de los modelos $ARIMA$: la estimación.

Estimación

```
ARIMA(2,1,1)

Coefficients:
         ar1      ar2     ma1
      1.6581  -0.8696  0.6360
s.e.  0.0886   0.0899  0.1235

sigma^2 = 88.13:  log likelihood = -107.46
AIC=222.91   AICc=224.58   BIC=228.38
```

El modelo $ARIMA(2,1,1)$ está ajustado a la serie de tiempo de los casos en el grupo de edad de **0-19 Años**. Este modelo incluye dos coeficientes AR, un coeficiente MA y una diferenciación. La varianza del error ($sigma^2$) es 88.13, y el logaritmo de la verosimilitud es -107.46. Donde, recordemos que los valores más bajos de **AIC, AICc** y **BIC** indican un mejor ajuste del modelo, ya que estas métricas no solo evalúan la calidad del ajuste, sino que también penalizan la complejidad del modelo para evitar el sobreajuste.

```
ARIMA(1,2,1)

Coefficients:
         ar1     ma1
      0.7579  0.6932
s.e.  0.1156  0.1077

sigma^2 = 1503:  log likelihood = -140.9
AIC=287.81   AICc=288.81   BIC=291.8
```

El modelo $ARIMA(1,2,1)$ se aplicó a los casos correspondientes al rango de edad de **20-29 Años**, este indica que se aplicaron dos diferenciaciones a la serie original para lograr estacionariedad. Los coeficientes AR (0.7579) y MA (0.6932) son ambos significativos, y el

valor de $sigma^2 = 1503$ representa la varianza del error. Las métricas de ajuste, como **AIC** (287.81), **AICc** (288.81) y **BIC** (291.8), permiten comparar este modelo con otros, donde valores más bajos indican mejor ajuste con penalización por complejidad. En conjunto, este modelo capta la dinámica de la serie con un equilibrio razonable entre precisión y simplicidad.

```
ARIMA(2,2,1)

Coefficients:
         ar1      ar2     ma1
      1.2983  -0.6237  0.3701
s.e.  0.1818   0.1749  0.2328

sigma^2 = 1464:  log likelihood = -139.85
AIC=287.71   AICc=289.44   BIC=293.03
```

El modelo $ARIMA(2, 2, 1)$ ajustado indica que la serie fue diferenciada dos veces para lograr estacionariedad y se identificaron dos términos autorregresivos (AR) y uno de media móvil (MA). Los coeficientes estimados son $ar1 = 1.2983, ar2 = -0.6237$ y $ma1 = 0.3701$, todos con errores estándar moderados, lo que sugiere que cada término tiene un efecto importante en el comportamiento de la serie. Las métricas **AIC, AICc** y **BIC** nos permiten comparar este modelo con otros, siendo útiles para evaluar su calidad, donde valores más bajos indican mejor ajuste con menor complejidad. Este modelo corresponde al análisis de los casos de esquizofrenia en personas de **30-39 Años**.

```
ARIMA(1,1,1)

Coefficients:
         ar1     ma1
      0.9933  0.3966
s.e.  0.0084  0.1567

sigma^2 = 11067:  log likelihood = -177.65
AIC=361.31   AICc=362.27   BIC=365.41
```

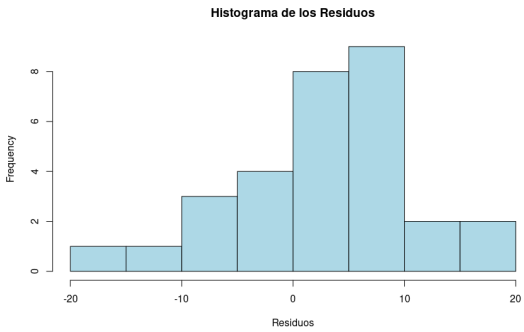
El modelo $ARIMA(1, 1, 1)$ se aplicó a los casos en el grupo de edad de **40-49 Años**. Este modelo usa una diferenciación y combina un término AR (0.9933) y MA (0.3966), mostrando un buen ajuste con coeficientes significativos. La varianza del error es 11,067, y las métricas **AIC, AICc** y **BIC** indican un ajuste razonable, útil para comparar con otros modelos.

```
ARIMA(1,2,1)

Coefficients:
          ar1     ma1
      -0.5379  0.1172
s.e.   0.2581  0.2800

sigma^2 = 43991:  log likelihood = -188.47
AIC=382.94   AICc=383.94   BIC=386.93
```

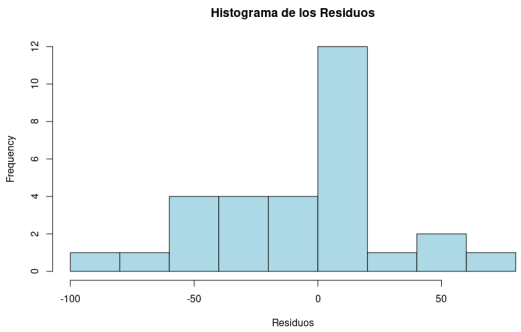
El modelo $ARIMA(1, 2, 1)$ aplicado a los casos del grupo de **50+ Años** utiliza una segunda diferenciación para alcanzar la estacionariedad y dos términos autorregresivos. Los errores estándar relativamente altos sugieren menor precisión en la estimación de los coeficientes. Además, la varianza del error es considerablemente elevada y las métricas **AIC, AICc** y **BIC** indican un ajuste menos favorable en comparación con modelos con valores más bajos.

Validación

Para el proceso de verificación, primero evaluamos la normalidad de los modelos mediante la prueba de Jarque-Bera y el análisis de sus respectivos histogramas, los cuales se presentan en las **Gráficas 5.43, 5.44** y **5.45**.

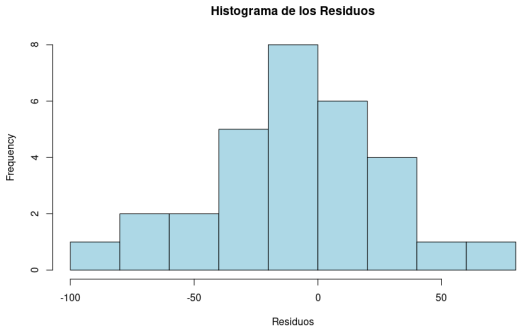


(a) 0-19 Años

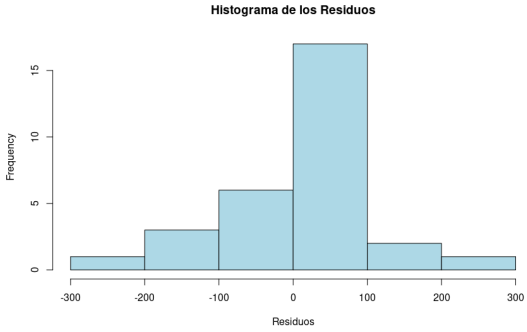


(b) 20-29 Años

Gráfica 5.43: Histograma de los Residuos del Modelo de los Casos por Edad



(a) 30-39 Años



(b) 40-49 Años

Gráfica 5.44: Histograma de los Residuos del Modelo de los Casos por Edad

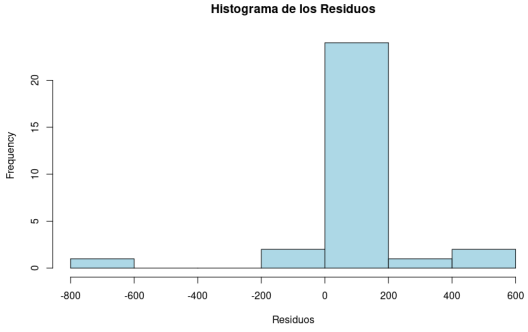


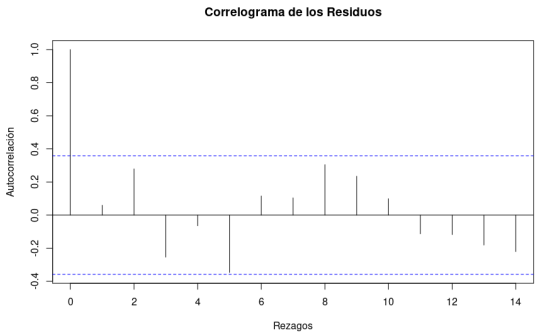
Figura 5.45: Histograma de los Residuos del Modelo de los Casos por Edad (50+ Años)

El histograma nos proporciona una forma visual de evaluar si los residuos cumplen con el supuesto de normalidad. Para que los residuos sean considerados normalmente distribuidos, la gráfica debe mostrar una forma aproximadamente simétrica y en forma de campana, similar a una distribución normal. Sin embargo, es importante recordar que la prueba de Jarque-Bera es una herramienta más confiable para verificar la normalidad, ya que el histograma puede no ser completamente preciso en algunos casos.

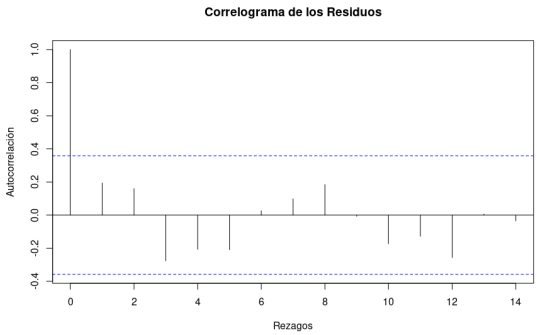
Ahora, de acuerdo con los resultados de las pruebas de Jarque-Bera aplicadas a cada uno de los modelos, se concluye que, todos nuestros modelos cumplen con los supuestos de normalidad, ya que los p-valores obtenidos fueron significativamente mayores a 0.05.

Continuando con el proceso de verificación, es necesario evaluar la significancia de los parámetros estimados. Para ello, realizamos los correlogramas de errores de nuestros modelos. Donde, a medida que disminuye el número de parámetros que superan la línea de confianza en el correlograma, esto indica que no hay autocorrelación significativa, lo cual es ideal para la validez de nuestro modelo. Los correlogramas correspondientes se presentan en las **Gráficas 5.46, 5.47** y **5.48**.

Además, también se aplicó la prueba de Ljung-Box a nuestros modelos para obtener un análisis mucho más preciso.

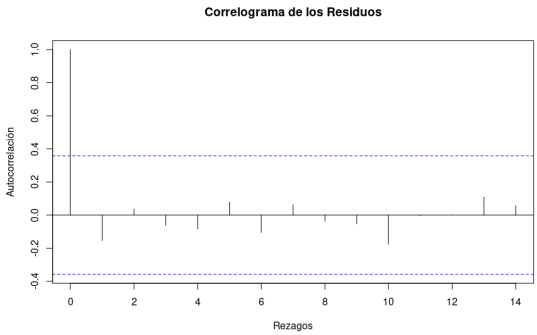


(a) 0-19 Años

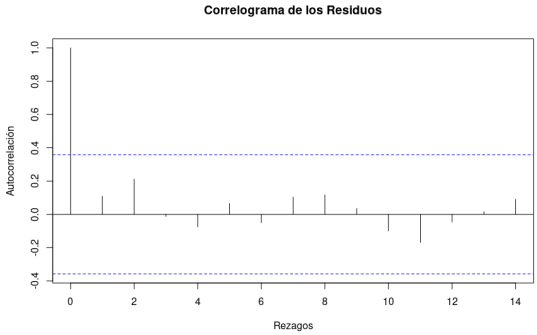


(b) 20-29 Años

Grfica 5.46: Correlograma del Modelo de los Casos por Edad



(a) 30-39 Años



(b) 40-49 Años

Gráfica 5.47: Correlograma del Modelo de los Casos por Edad

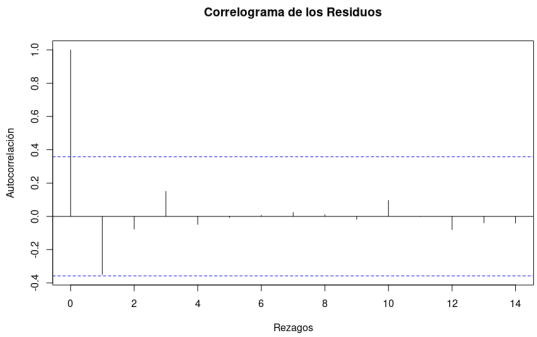
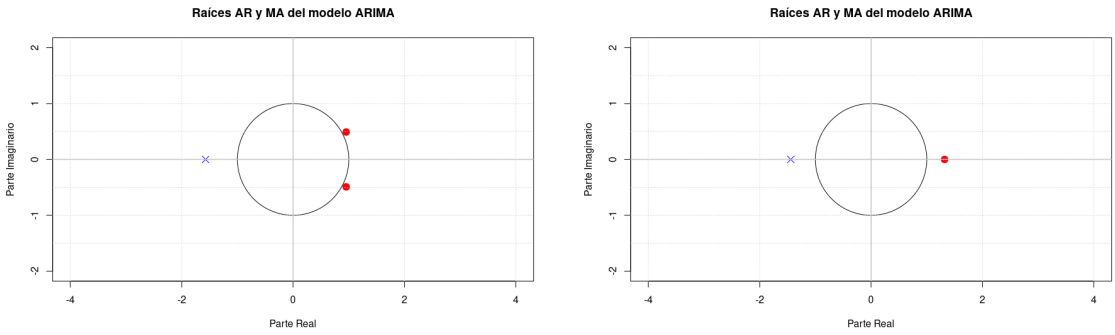


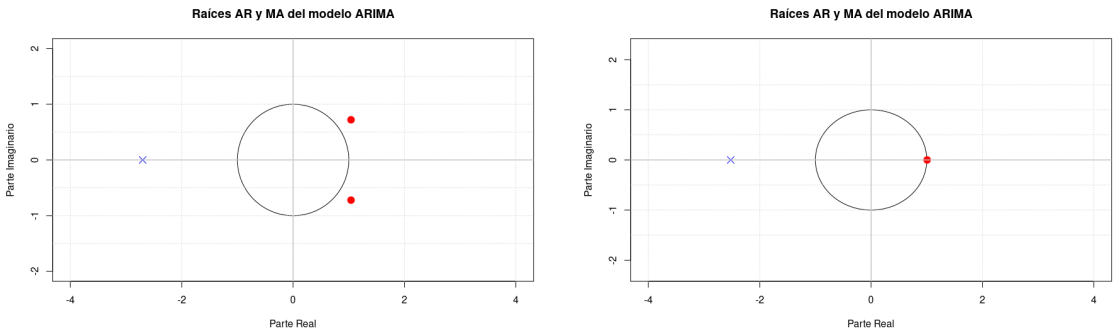
Figura 5.48: Correlograma del Modelo de los Casos por Edad (50+ Años)

Gracias a los correlogramas de los errores y a los resultados de la prueba de Ljung-Box, cuyos p-valores fueron significativamente mayores que nuestro nivel de significancia ($\alpha = 0.05$), podemos concluir que no existe autocorrelación significativa en los residuos de nuestros modelos. Esto indica que los modelos están bien especificados.



(a) 0-19 Años (b) 20-29 Años

Gráfica 5.49: Raíces Unitarias del Modelo de los Casos por Edad



(a) 30-39 Años (b) 40-49 Años

Gráfica 5.50: Raíces Unitarias del Modelo de los Casos por Edad

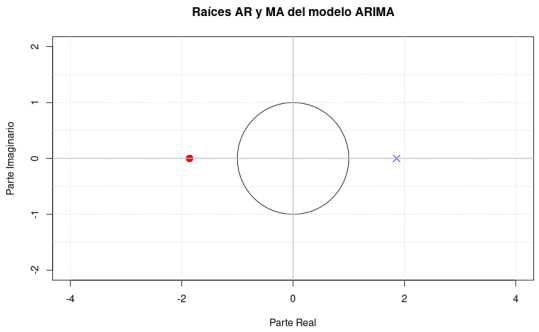


Figura 5.51: Raíces Unitarias del Modelo de los Casos por Edad (50+ Años)

Como paso final en nuestro proceso de verificación, se calcularon y graficaron las raíces unitarias de los modelos, las cuales se presentan en las **Gráficas 5.49, 5.50** y **5.51**. En ellas, se observa que tanto las raíces autorregresivas (AR), representadas por puntos rojos, como las de media móvil (MA), representadas por equis azules, se encuentran fuera del círculo unitario. Esto nos permite concluir que los cinco modelos son estacionarios e invertibles.

Ahora, que hemos revisado y verificado que cada uno de nuestros modelos cumple con los requisitos necesarios para realizar las predicciones deseadas, procederemos a analizar cómo se representan matemáticamente los modelos $ARIMA$.

En primer lugar, consideremos un modelo $ARIMA(2,1,1)$ aplicado a los casos de esquizofrenia en el grupo de edad de **0-19 Años**. Esto significa que:

- $p = 2$: parte autorregresiva (AR) de orden 2,
- $d = 1$: primera diferenciación,
- $q = 1$: parte de media móvil (MA) de orden 1.

Donde los coeficientes dados son:

- $\phi_1 = 1.6581 (AR1)$,
- $\phi_2 = -0.8696 (AR2)$,
- $\theta_1 = 0.6360 (MA1)$.

Ahora, tenemos que la forma general del $ARIMA(2,1,1)$ es (Ver ecuación (4.1)):

$$(1 - \phi_1 L - \phi_2 L^2)(1 - L) y_t = (1 + \theta_1 L) a_t.$$

Donde:

- $\boldsymbol{(1 - L)}$ representa **diferenciar la serie** (hacerla estacionaria).
- $\boldsymbol{(1 - \phi_1 L - \phi_2 L^2)}$ representa la **parte $\boldsymbol{AR}$** que modela la serie estacionaria.
- $\boldsymbol{(1 + \theta_1 L)}$ representa la **parte $\boldsymbol{MA}$** que modela el ruido/error.

Primero expandimos $(1 - L) y_t$ (ver ecuación (4.3)), que es:

$$(1 - L) y_t = y_t - y_{t-1}.$$

Ahora, aplicamos el operador AR:

$$(1 - \phi_1 L - \phi_2 L^2)(y_t - y_{t-1}).$$

Expandimos:

$$y_t - y_{t-1} - \phi_1(y_{t-1} - y_{t-2}) - \phi_2(y_{t-2} - y_{t-3}).$$

Agrupamos:

$$y_t - (1 + \phi_1)y_{t-1} + (\phi_1 - \phi_2)y_{t-2} + \phi_2 y_{t-3}.$$

Y esto es igual a:

$$(1 + \theta_1 L)a_t = a_t + \theta_1 a_{t-1}.$$

Ahora, sustituimos los valores de nuestros coeficientes:

$$y_t - (1 + (1.6581))y_{t-1} + ((1.6581) - (-0.8696))y_{t-2} + (-0.8696)y_{t-3} = a_t + (0.6360)a_{t-1}.$$

Por lo que tenemos que:

$$y_t - 2.6581y_{t-1} + 2.5277y_{t-2} - 0.8696y_{t-3} = a_t + 0.6360a_{t-1}.$$

Despejamos y_t para obtener la ecuación deseada, resultando en:

$$\boldsymbol{y_t = 2.6581y_{t-1} - 2.5277y_{t-2} + 0.8696y_{t-3} + 0.6360a_{t-1} + a_t},$$

con a_t un proceso de ruido blanco.

A partir del cálculo anterior, se procedió de la misma manera para obtener las ecuaciones correspondientes a los demás modelos $ARIMA$ estimados.

0-19 Años

$ARIMA(2,1,1) \rightarrow \boldsymbol{y_t = 2.6581y_{t-1} - 2.5277y_{t-2} + 0.8696y_{t-3} + 0.6360a_{t-1} + a_t}.$

20-29 Años

$ARIMA(1,2,1) \rightarrow \boldsymbol{y_t = 2.7579y_{t-1} - 2.5158y_{t-2} + 0.7579y_{t-3} + 0.6932a_{t-1} + a_t}.$

30-39 Años

$ARIMA(2,2,1) \rightarrow \boldsymbol{y_t = 3.2983y_{t-1} - 4.2203y_{t-2} + 2.5457y_{t-3} - 0.6237y_{t-4} + 0.3701a_{t-1} + a_t}.$

40-49 Años

$ARIMA(1,1,1) \rightarrow \boldsymbol{y_t = 1.9933y_{t-1} - 0.9933y_{t-2} + 0.3966a_{t-1} + a_t}.$

50+ Años

$ARIMA(1,2,1) \rightarrow \boldsymbol{y_t = 1.4621y_{t-1} + 0.0758y_{t-2} - 0.5379y_{t-3} + 0.1172a_{t-1} + a_t}.$

Predicción

En la **Figura 5.52** se presentan las predicciones del modelo para el rango de edad de **0-19 Años**. Este resultado muestra las predicciones puntuales anuales desde 2020 hasta 2029, junto con sus respectivos intervalos de confianza al 95 %. Por ejemplo, para el año 2020, se espera un valor de aproximadamente 8415.7, con un intervalo de confianza entre 8397.3 y 8434.1. A medida que avanzan los años, las predicciones tienden a disminuir ligeramente y los intervalos de confianza se amplían, lo que refleja un aumento en la incertidumbre de las estimaciones a largo plazo.

	Point Forecast	Lo 95	Hi 95
2020	8415.700	8397.300	8434.101
2021	8395.144	8331.802	8458.487
2022	8370.018	8239.080	8500.956
2023	8346.232	8133.702	8558.763
2024	8328.644	8031.257	8626.032
2025	8320.167	7944.567	8695.767
2026	8321.406	7881.182	8761.630
2027	8330.832	7842.523	8819.141
2028	8345.384	7824.685	8866.083
2029	8361.315	7820.509	8902.121

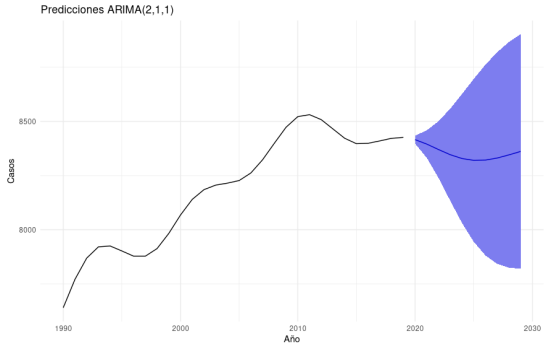


Figura 5.52: Predicción de los Casos por Edad (0-19 Años)

Las predicciones realizadas con el modelo para los casos en el grupo de edad de **20-29 Años** se presentan en la **Figura 5.53**. Este resultado presenta las predicciones anuales desde 2020 hasta 2029, con un crecimiento sostenido en los valores pronosticados, que pasan de 62,583.19 en 2020 a 66,439.69 en 2029. Los intervalos de confianza al 95 % se amplían progresivamente, lo que indica un aumento en la incertidumbre de las predicciones a medida que avanza el tiempo. Esto sugiere una tendencia creciente con menor precisión en los años más lejanos.

	Point Forecast	Lo 95	Hi 95
2020	62583.19	62507.22	62659.17
2021	63033.40	62760.43	63306.38
2022	63474.53	62876.63	64072.44
2023	63908.78	62857.22	64960.35
2024	64337.82	62707.29	65968.34
2025	64762.90	62433.51	67092.29
2026	65184.99	62042.99	68326.99
2027	65604.80	61542.77	69666.84
2028	66022.90	60939.54	71106.26
2029	66439.69	60239.48	72639.90

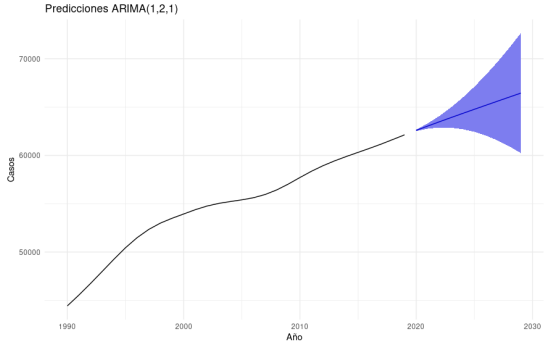


Figura 5.53: Predicción de los Casos por Edad (20-29 Años)

A continuación, en la **Figura 5.54** se muestra la predicción generada a partir del modelo desarrollado con los casos de esquizofrenia en el grupo de edad de **30-39 Años**. Este resultado presenta pronósticos anuales desde 2020 hasta 2029, mostrando un crecimiento sostenido en los valores estimados. Cada año incluye un intervalo de confianza al 95 %, que refleja la incertidumbre del pronóstico y se amplía con el tiempo, indicando menor precisión en las proyecciones más lejanas.

	Point Forecast	Lo 95	Hi 95
2020	92815.36	92740.35	92890.36
2021	94081.15	93795.97	94366.33
2022	95347.28	94691.12	96003.44
2023	96590.51	95412.06	97768.97
2024	97803.80	95982.78	99624.82
2025	98992.49	96445.60	101539.39
2026	100167.94	96843.05	103492.83
2027	101341.52	97206.42	105476.63
2028	102520.95	97551.73	107490.18
2029	103709.14	97881.67	109536.61

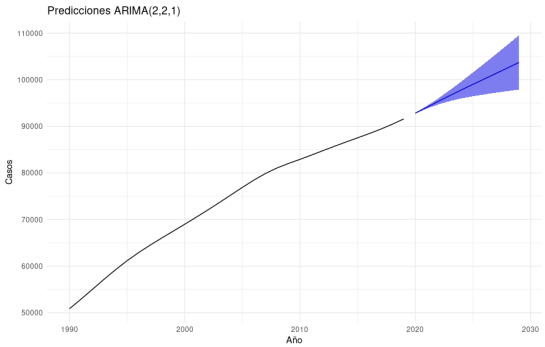


Figura 5.54: Predicción de los Casos por Edad (30-39 Años)

La **Figura 5.55** muestra las predicciones generadas por el modelo para los casos del grupo de edad de **25-29 Años**, donde se aprecia un aumento gradual en los valores estimados entre 2020 y 2029. Los intervalos de confianza se amplían con el paso del tiempo, lo que refleja una mayor incertidumbre en las proyecciones a largo plazo. Por lo tanto, podemos considerar que el modelo es adecuado para predicciones a corto plazo.

	Point Forecast	Lo 95	Hi 95
2020	82693.07	82486.88	82899.26
2021	83802.66	83268.48	84336.83
2022	84904.81	83961.54	85848.08
2023	85999.58	84580.50	87418.65
2024	87087.02	85134.52	89039.51
2025	88167.17	85630.09	90704.25
2026	89240.08	86072.12	92408.04
2027	90305.81	86464.54	94147.08
2028	91364.40	86810.57	95918.24
2029	92415.90	87112.91	97718.89

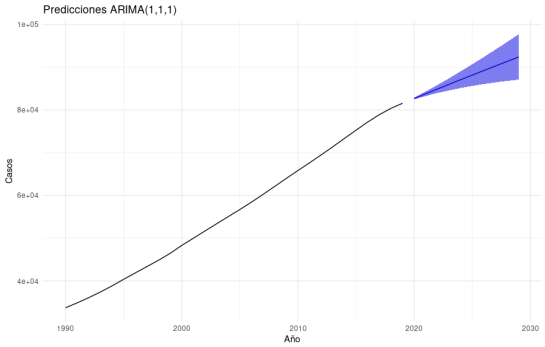


Figura 5.55: Predicción de los Casos por Edad (40-49 Años)

La **Figura 5.56** presenta las predicciones generadas por el modelo para el grupo de **50+ años**, en las que se observa una tendencia de crecimiento anual sostenido entre 2020 y 2029. Los intervalos de confianza al 95 % indican el rango en el que probablemente se ubicarán los valores reales, reflejando la incertidumbre asociada a las estimaciones. Aunque estos intervalos se amplían ligeramente conforme avanza el horizonte de pronóstico, como es común en modelos de series de tiempo, su incremento es más moderado en comparación con los modelos aplicados a otros grupos etarios, lo que sugiere una mayor estabilidad en las predicciones para este segmento poblacional.

	Point Forecast	Lo 95	Hi 95
2020	93089.23	92678.15	93500.31
2021	96274.49	95506.07	97042.90
2022	99458.65	98213.02	100704.28
2023	102643.41	100870.57	104416.24
2024	105827.84	103460.06	108195.62
2025	109012.45	105999.57	112025.33
2026	112196.97	108487.22	115906.71
2027	115381.53	110929.38	119833.68
2028	118566.07	113327.29	123804.85
2029	121750.62	115684.03	127817.21

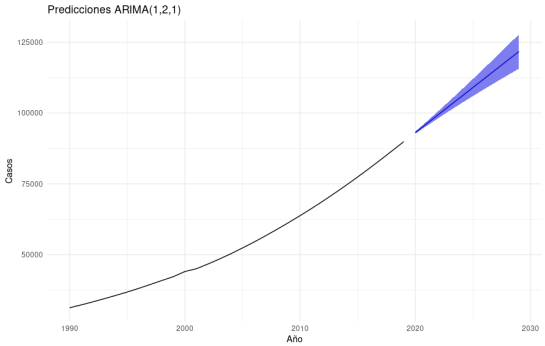


Figura 5.56: Predicción de los Casos por Edad (50+ Años)

Como paso final, se calcularon los errores de las predicciones con el objetivo de evaluar su grado de confiabilidad.

```
[1] "MAE: 485.484405034504"
[1] "RMSE: 499.264028742762"
[1] "MAPE: 6.18851832110809"
```

La predicción (**0-19 Años**) presenta un buen desempeño predictivo, con un **MAE** de 485.48 y un **RMSE** de 499.26, lo que indica errores moderados y sin grandes desviaciones. Además, el **MAPE** de 6.19 % sugiere que, en promedio, las predicciones se desvían poco respecto a los valores reales, mostrando una buena precisión relativa.

```
[1] "MAE: 15062.0011984714"
[1] "RMSE: 15171.3549740081"
[1] "MAPE: 30.7922269136736"
```

La predicción (**20-29 Años**) muestra un desempeño moderado, con un **MAE** de 15062 y un **RMSE** de 15171, lo que indica errores altos y consistentes. Además, el **MAPE** de 30.79 % refleja una desviación promedio considerable respecto a los valores reales, lo que sugiere una precisión relativa baja en las predicciones.

```
[1] "MAE: 38607.314399526"
[1] "RMSE: 38655.4358917971"
[1] "MAPE: 65.4770379780677"
```

La predicción (**30-39 Años**) presenta un **MAE** de aproximadamente 38,607, lo que indica que, en promedio, las predicciones se desvían de los valores reales en esa cantidad. El **RMSE** es similar, con un valor de 38,655, y penaliza más fuertemente los errores grandes. Finalmente, el **MAPE** es de 65.48 %, lo que implica que, en promedio, el modelo comete un error del 65.48 % respecto al valor real, lo cual sugiere un bajo nivel de precisión relativa y una alta variabilidad entre los valores pronosticados y observados.

```
[1] "MAE: 47829.3495740329"
[1] "RMSE: 47840.2467122528"
[1] "MAPE: 121.820088031693"
```

La predicción (**40-49 Años**) presenta errores considerables, con un **MAE** de 47,829 y un **RMSE** de 47,840, lo que indica un alto nivel de error promedio. Además, el **MAPE** supera el 121 %, lo que refleja una baja precisión relativa del modelo, con errores que en promedio duplican los valores reales. Si bien el modelo fue desarrollado siguiendo las reglas del enfoque $ARIMA$ para lograr la mejor configuración posible, estos resultados sugieren que podría ser necesario complementarlo con otras herramientas o métodos para mejorar su capacidad predictiva.

```
[1] "MAE: 70964.8250074766"
[1] "RMSE: 71188.1370838457"
[1] "MAPE: 194.981689300692"
```

La predicción (**50+ Años**) muestra errores absolutos altos y un **MAPE** muy elevado (casi 195 %), lo que indica un bajo desempeño en la precisión de las predicciones.

En conclusión, la aplicación del modelo ARIMA permitió analizar el comportamiento temporal de los casos de esquizofrenia registrados en México entre 1990 y 2019. A través de un proceso de identificación, estimación y validación del modelo, se lograron generar predicciones con un nivel de precisión aceptable, como lo indican las métricas de error obtenidas. Además, el análisis desagregado por sexo y edad ofreció una visión detallada sobre los grupos poblacionales más afectados, revelando patrones que no se observan en el análisis global. El estudio desarrollado en este trabajo de tesis puede servir como punto de partida para futuros estudios que analicen los datos con mayor detalle o que utilicen modelos más avanzados. Los códigos utilizados para la implementación de cada uno de los modelos presentados en este capítulo se encuentra disponible en el **Apéndice A**.

Conclusiones

El objetivo principal de esta tesis fue aplicar de manera adecuada el modelo $ARIMA$ a la base de datos seleccionada, centrada en los casos de esquizofrenia, haciendo énfasis en cada una de las etapas necesarias para obtener predicciones precisas: identificación, estimación, validación y pronóstico. A lo largo del estudio, se implementaron diversos modelos $ARIMA$ con distintos órdenes de autorregresión (AR), integración (I) y medias móviles (MA), ajustados según las características específicas de cada serie temporal analizada. Estos modelos se adaptaron en función del enfoque de análisis de los datos, ya fuera por año, por género o por grupo de edad. Por lo tanto, considerando esta subdivisión, se procederá a evaluar la eficacia del modelo y la calidad de las predicciones obtenidas en cada uno de estos casos.

Comenzamos con el modelo aplicado al análisis de la base de datos por años. En este caso, se utilizó un modelo $ARIMA(2, 2, 0)$, el cual fue seleccionado automáticamente mediante la función *auto.arima* en RStudio. Este modelo superó satisfactoriamente todas las pruebas de validación requeridas, lo cual respaldó su capacidad como el mejor modelo $ARIMA$ posible para esta serie temporal. Pero como una posible extensión de esta tesis, se podría considerar la estimación de otros modelos $ARIMA$ utilizando diferentes métodos de selección y evaluación, comparando sus desempeños con base en criterios como el **AIC, AICc** y **BIC**, a fin de identificar cuál ofrece el mejor ajuste.

Adicionalmente, al analizar los resultados obtenidos de los errores de predicción, se observa que los indicadores muestran un desempeño predictivo razonable, con errores dentro de un rango aceptable para aplicaciones prácticas. En conclusión, aunque el modelo y sus predicciones no son perfectos, ofrecen una base sólida y suficientemente precisa para su uso en estudios aplicados y futuros análisis.

En cuanto a los modelos desarrollados a partir de la base de datos dividida por género, se obtuvieron configuraciones distintas: un modelo $ARIMA(2, 2, 0)$ para el grupo masculino y un $ARIMA(1, 1, 0)$ para el femenino. Ambos modelos presentan una estructura específica que se ajusta a las características de cada serie temporal. Al analizar sus errores de predicción, se observó un desempeño aceptable, lo que indica que la elección del orden del modelo permitió capturar adecuadamente la dinámica de los datos en cada caso.

Por último, los cinco modelos construidos a partir de la base de datos dividida por grupos de edad mostraron mayor variabilidad en su comportamiento y resultados. En estos casos, se obtuvo un desarrollo más completo del proceso $ARIMA$, ya que los modelos incluyeron tanto parámetros autorregresivos (AR) como de media móvil (MA). Aunque la mayoría de estos modelos no presentaron dificultades significativas durante el proceso de validación, los errores de predicción mostraron un patrón creciente a medida que aumentaba el rango de edad

analizado. En particular, el modelo correspondiente al grupo de edad de **55+ Años** alcanzó un $MAPE$ de aproximadamente **195**, lo que indica una baja precisión en las predicciones.

Esto sugiere que, entre los cinco modelos analizados, solo el primero presentó un $MAPE$ considerablemente bajo, mientras que los restantes mostraron un desempeño menor en términos de precisión. En consecuencia, se podría considerar como una posible extensión de esta tesis la exploración de otros enfoques de modelado, con el objetivo de mejorar la capacidad predictiva en los grupos de edad donde el modelo $ARIMA$ mostró limitaciones.

Más allá del análisis técnico y de la precisión alcanzada por los modelos, resulta fundamental detenerse a reflexionar sobre el mensaje que transmiten las predicciones: en la mayoría de los escenarios analizados, se anticipa un aumento sostenido en los casos de esquizofrenia durante los próximos años. Este resultado, independientemente del modelo utilizado, subraya la urgencia de fortalecer las estrategias de prevención, diagnóstico temprano y tratamiento, así como de ampliar el acceso a servicios de salud mental. En resumen, de cumplirse las proyecciones, se requerirá una preparación adecuada para afrontar un problema de salud pública cuya magnitud podría volverse cada vez más apremiante.

Además, es importante recordar que las predicciones dependen directamente de los datos disponibles y pueden verse afectadas por factores externos que no fueron considerados en el modelo. Por ello, para mejorar la precisión de los resultados, se recomienda incorporar información más actualizada y ampliar el análisis incluyendo variables adicionales, como factores ambientales, demográficos, regionales y socioeconómicos, que podrían influir en el comportamiento de la serie. De igual forma, como línea de trabajo futuro se podría comparar el desempeño de modelos ARIMA con técnicas de aprendizaje automático, como redes neuronales recurrentes (RNN) o modelos de árboles de decisión, con el fin de evaluar mejoras en la predicción.

Bibliografía

[1] Alanen, Y. O. (2003). *La esquizofrenia: Sus orígenes y su tratamiento adaptado a las necesidades del paciente*. Editorial H. Karnak Ltd.

[2] American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Association.

[3] Box, G. E. P. & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. Holden-Day.

[4] Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. (2008). *Time series analysis: Forecasting and control* (4th ed.). Wiley.

[5] Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. (2015). *Time Series Analysis: Forecasting and Control (5th ed.)*. Wiley.

[6] Brockwell, P. J. & Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. Springer Texts in Statistics.

[7] Cryer, J. D. & Chan, K. S. (2008). *Time series analysis with applications in R* (2^a *ed.)*. Springer.

[8] Escamilla-Orozco, R. I., et al. (2021). *Tratamiento de la esquizofrenia en México: Recomendaciones de un panel de expertos*. Gaceta Médica de México. https://doi.org/10.24875/gmm.m21000501

[9] González, M. P. (2009). *Análisis de series temporales: Modelos ARIMA*. España.

[10] Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.

[11] Hodrick, R. J., & Prescott, E. C. (1997). *Postwar U.S. Business Cycles: An Empirical Investigation*. Journal of Money, Credit and Banking.

[12] Kahn, R. S., et al. (2015). *Schizophrenia*. Nature reviews. https://doi.org/10.1038/nrdp.2015.67

[13] Martín García-Sancho, J. C. (Coord.), et al. (2009). *Guía de práctica clínica para el tratamiento de la esquizofrenia en centros de salud mental*. Servicio Murciano de Salud, Subdirección de Salud Mental.

[14] Mexican Congress. (2020, febrero 13). *Prevalencia de la esquizofrenia en México*. Gaceta Diputados. https://gaceta.diputados.gob.mx/Gaceta/64/2020/feb/20200213-V.html

[15] Milenio. (2022, julio 12). *En México más de un millón de personas tiene esquizofrenia: experto*. https://www.milenio.com/ciencia-y-salud/en-mexico-mas-de-un-millon-de-personas-tiene-esquizofrenia-experto

[16] Organización Mexicana de Especialistas en Salud Mental. (2020). *Déficit de especialistas y escasez de medicamentos para la esquizofrenia en México*. OEM. https://oem.com.mx/elsoldetoluca/ciencia-y-salud/mexico-tiene-un-deficit-de-especialistas-para-tratar-la-esquizofrenia-14239068

[17] Peña, D. (2005). *Análisis de series temporales*. Alianza editorial. Madrid.

[18] Romero, A. R. (2021-2022). *Apuntes de Macroeconometría.* Recuperado de *https* : *//randall − romero.github.io/econometria/*00 − *acerca.html*

[19] Rojas-Jiménez, K. (2022). *Ciencia de Datos para Ciencias Naturales.* Recuperado de *https* : *//bookdown.org/keilor_rojas/CienciaDatos/*

[20] Sánchez, C., Martínez, R., & Pérez, J. (2020). *Atención hospitalaria y consultas ambulatorias por esquizofrenia en México: análisis 2000–2016.* Salud Pública de México, 62(1), 72–79. https://www.scielosp.org/article/spm/2020.v62n1/72-79/es

[21] Secretaría de Salud. (2002, enero 5). *Esquizofrenia en México: atención y cifras.* Gobierno de México. https://www.salud.gob.mx/unidades/dgcs/sala_noticias/comunicados/2002-01-05-004-ESQUIZOFRENIA.html

[22] Shumway, R. H. & Stoffer, D. S. (2011). *Time series analysis and its applications: With R examples (*3^a *ed.).* Springer.

[23] Spiegel, M. R. (1991). *Estadística* (2^a ed., Hernández, H, Trad.; Abellanas, R, Rev. Téc.). McGraw-Hill/Interamericana de España.

[24] Institute for Health Metrics and Evaluation. (2025). *GBD Results Tool: Cause of death or injury – Schizophrenia.* University of Washington. https://tinyurl.com/72nb8437

[25] Wackerly, D. D., Mendenhall III, W. & Schaeffer, R. L. (2010). *Estadística matemática con aplicaciones (*7^a *ed.).* Cengage Learning Editores, S.A. de C.V.

[26] Wooldridge, J. M. (2020). *Introducción a la econometría: Un enfoque moderno* (4^a ed.). Cengage Learning.

[27] World Life Expectancy. (2020). *Schizophrenia mortality in Mexico.* https://www.worldlifeexpectancy.com/es/mexico-schizophrenia

Apéndice A

Para evitar redundancias en la sección descriptiva de la aplicación del modelo $ARIMA$, en este apéndice se incluirá el código completo en R utilizado para la implementación de los distintos modelos explicados en el **Capítulo 5**. Aunque los resultados obtenidos fueron diferentes, es importante señalar que se utilizó el mismo código en su ejecución, con solo algunos ajustes según los resultados obtenidos.

Recordemos que en el **Capítulo 4** se explicó que, para la construcción del modelo $ARIMA$, es necesario seguir cuatro pasos clave para obtener un pronóstico más preciso: identificación, estimación, validación y predicción. Estos mismos pasos fueron tomados en cuenta al desarrollar el código en R.

Por lo tanto, para explicar el código, utilizaremos estos pasos, lo que permitirá una comprensión más clara y precisa de cada fase del proceso.

A.1. Identificación

La identificación es el primer paso en la construcción del modelo. En este proceso, se llevan a cabo varios procedimientos que también serán explicados a continuación.

En primer lugar, se solicita la realización de la gráfica de la serie, pero para ello es necesario ejecutar algunos códigos previos.

```
excel = read.xlsx("Datos.xlsx", sheet=)
```

Este código lee el archivo de Excel que contiene nuestra base de datos, la cual fue previamente preparada para su uso en el análisis.

```
datos = data.frame(excel$Ano, excel$Valor)
```

En este código, definimos las variables que se utilizarán para la construcción de la serie de tiempo.

```
tsCasos = ts(datos$excel.Valor, start= 1990, end = 2019,
   frequency = 1)
```

Este código genera la serie de tiempo utilizando únicamente los valores almacenados en la variable "Valor". Además, se especifica que la serie abarca desde el año 1990 hasta 2019, con una frecuencia de 1, lo que indica que los datos se toman anualmente.

```r
plot(tsCasos)
```

Finalmente, con este código, se genera la gráfica de la serie de tiempo.

Además, en esta etapa de identificación, es necesario verificar la estacionariedad de la serie de tiempo, lo cual se realiza mediante la prueba de Dickey-Fuller y el correlograma de errores.

El correlograma de los errores se genera utilizando el siguiente código:

```r
#Calcular ACF sin graficar
acf_result <- acf(tsCasos, plot = FALSE)

#Graficar con etiquetas personalizadas
plot(acf_result,
     main = "Correlograma de la Serie",
     xlab = "Rezagos",
     ylab = "Autocorrelacion")
```

Por otro lado, la prueba de Dickey-Fuller se genera con el siguiente código:

```r
acf(tsCasos)
```

Con el cual, obtenemos un resultado como el siguiente:

```
    Augmented Dickey-Fuller Test

data:  Df2
Dickey-Fuller = -2.665, Lag order = 3, p-value = 0.03172
alternative hypothesis: stationary
```

En caso de que la serie no sea estacionaria, es necesario aplicar diferencias hasta que cumpla con los criterios de estacionariedad. Este proceso se lleva a cabo con el siguiente código:

```r
Df = diff(tsCasos)
```

Además, tras aplicar cada diferenciación necesaria (una o dos), es importante volver a graficar la serie de tiempo y generar nuevamente el correlograma, ahora con la serie ya diferenciada.

Y, una vez comprobada la estacionariedad de nuestra serie de tiempo, podemos proceder al siguiente paso en la construcción del modelo $ARIMA$: la estimación.

A.2. Estimación

En el paso de estimación, es necesario calcular los parámetros del modelo $ARIMA$. Para ello, se utilizó el siguiente código:

```r
modelo <- auto.arima(tsCasos,stepwise=FALSE,approximation=
    FALSE)
```

El cual genera un resultado similar al siguiente:

```
ARIMA(2,2,1)

Coefficients:
         ar1      ar2      ma1
      1.4253  -0.7406  -0.6476
s.e.  0.1828   0.1434   0.2593

sigma^2 = 8182:  log likelihood = -176.72
AIC=361.44   AICc=363.04   BIC=367.05
```

Este código proporciona con precisión el resultado del modelo $ARIMA$ aplicable a nuestra serie de tiempo. Dicho resultado se utilizó en la explicación de cada uno de los modelos desarrollados en esta tesis, por lo que sería redundante detallar nuevamente cada uno de los valores obtenidos.

A.3. Validación

El siguiente paso consiste en validar el modelo $ARIMA$ estimado en el paso anterior, con el fin de comprobar su precisión y confiabilidad.

Como primer paso, se lleva a cabo una prueba de normalidad sobre los residuos del modelo. Para ello, se utiliza la prueba de Jarque-Bera, implementada mediante el siguiente código:

```
jarque.bera.test(residuals(modelo))
```

El cual nos genera un resultado similar al siguiente:

```
    Jarque Bera Test

data:  residuals(modelo)
X-squared = 3.9785, df = 2, p-value = 0.1368
```

Como complemento a la prueba de Jarque-Bera, para verificar si los residuos de nuestro modelo cumplen con el supuesto de normalidad, también se generó un histograma, realizado con el siguiente código de R:

```
hist(residuals(modelo), main="Histograma de los Residuos", xlab
   ="Residuos", col="lightblue", border="black")
```

Continuando con el proceso de verificación, ahora es necesario determinar si los parámetros estimados son significativos. Para ello, se debe realizar el correlograma de los errores del modelo, el cual fue generado con el siguiente código:

```
acf_result <- acf(resid(modelo), plot = FALSE)

# Graficar con etiquetas personalizadas
plot(acf_result,
     main = "Correlograma de los Residuos",
     xlab = "Rezagos",
     ylab = "Autocorrelacion")
```

Además, para obtener un resultado más preciso, se realizó la prueba de Box-Ljung, la cual se llevó a cabo utilizando el siguiente código:

```
Box.test(resid(modelo), lag = 10, type = "Ljung-Box")
```

El cual generó un resultado como el siguiente:

```
    Box-Ljung test

data:  resid(modelo)
X-squared = 3.9942, df = 10, p-value = 0.9476
```

Finalmente para terminar con el proceso de verificación, generamos la gráfica de nuestras raíces unitarias para comprobar si el modelo es estacionario e invertible. Esta gráfica se elaboró utilizando el siguiente código:

```
#Para obtener coeficientes AR y MA
ar_coefs <- c(1, -coef(modelo)[1:2])  #Coeficientes AR
ma_coefs <- c(1, coef(modelo)[3])   #Coeficientes MA

#Calcular raices del polinomio caracteristico
ar_roots <- polyroot(ar_coefs)  #Raices del componente AR
ma_roots <- polyroot(ma_coefs)  #Raices del componente MA

#Crear el circulo unitario
theta <- seq(0, 2*pi,length.out = 500)
x <- cos(theta)
y <- sin(theta)

#Dibujar el circulo unitario
plot(x, y, type = "l", lty = 7, col = "black",
     xlab = "Parte Real", ylab = "Parte Imaginario",
     asp = 1, xlim = c(-4, 4), ylim = c(-1, 1),
     main = "Raices AR y MA del modelo ARIMA")

#Raices AR (Circulos rojos)
points(Re(ar_roots), Im(ar_roots), col="red", pch=19, cex=1.5)

#Raices MA (Cruz azul)
points(Re(ma_roots), Im(ma_roots), col="blue", pch=4, cex=1.5)

#Agregar ejes x e y
abline(h = 0, v = 0, col = "grey", lwd = 1.5)

#Agregar cuadricula
grid(col = "lightgray", lty = "dotted")
```

A.4. Predicción

Como paso final, se realiza la predicción. Para ello, utilizamos el siguiente código:

```
prono = forecast(modelo,level=c(95), h=10)
```

Este código generó las predicciones esperadas, abarcando un período de diez años en el futuro y con un nivel de confianza del 95 %.

```
     Point Forecast    Lo 95    Hi 95
2022       39302.70 39125.42 39479.98
2023       39060.20 38536.83 39583.58
2024       38789.63 37771.93 39807.32
2025       38605.61 36998.37 40212.85
2026       38565.78 36330.25 40801.32
2027       38667.34 35807.85 41526.84
2028       38863.65 35406.69 42320.61
2029       39090.29 35064.59 43115.98
2030       39289.97 34712.59 43867.36
2031       39428.79 34298.48 44559.11
```

Además, como complemento, se generó la gráfica del pronóstico utilizando el siguiente código:

```
autoplot(prono) +
  labs(title = "Predicciones ARIMA(p,d,q)",
       x = "Ano",
       y = "Casos") +
  theme_minimal()
```

Finalmente, para evaluar la precisión de nuestras predicciones, calculamos los errores de predicción utilizando los siguientes códigos:

```
#Calcular los errores de la prediccion
Totales = ts(datos$excel.Valor, start= 2010, end = 2019,
   frequency = 1)
valores_reales <- as.numeric(Totales)
valores_predichos <- as.numeric(prono$mean)

mae <- mean(abs(valores_reales - valores_predichos))
rmse <- sqrt(mean((valores_reales - valores_predichos)^2))
mape <- mean(abs((valores_reales - valores_predichos) / valores
   _reales)) * 100

print(paste("MAE:", mae))
print(paste("RMSE:", rmse))
print(paste("MAPE:", mape))
```

El código anterior nos arrojó resultados como los siguientes:

```
[1] "MAE: 1057.53975142231"
[1] "RMSE: 1149.75522566852"
[1] "MAPE: 19.2493926545431"
```