ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA

PREDICCIÓN DE DEMANDA ELÉCTRICA MEDIANTE LA APLICACIÓN DE MODELOS ARIMA Y SARIMA EN LENGUAJE DE PROGRAMACIÓN R – CASO DE ESTUDIO EN LA EMPRESA ELÉCTRICA QUITO

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERA ELÉCTRICA

DAYANA ESTEFANIA HERRERA GRANDA

dayana.herrera@epn.edu.ec

DIRECTOR: DR.-ING. FABIÁN ERNESTO PÉREZ YAULI

Quito, Julio 2019

AVAL

Certifico que el presente trabajo fue desarrollado por Dayana Estefanía Herrera Granda, bajo mi supervisión.

Dr.-Ing. Fabián Ernesto Pérez Yauli
DIRECTOR DEL TRABAJO DE TITULACIÓN

DECLARACIÓN DE AUTORÍA

Yo, Dayana Estefanía Herrera Granda, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración dejo constancia de que la Escuela Politécnica Nacional podrá hacer uso del presente trabajo según los términos estipulados en la Ley, Reglamentos y Normas vigentes.

DAYANA ESTEFANIA HERRERA GRANDA

DEDICATORIA

Al tiempo por permitirnos encontrarnos en este momento y tener a mi familia en todo este arduo camino. Dedico este trabajo, todas las sonrisas vividas y por vivir a mi padre que me enseñó a jamás rendirme, a mi madre que su bondad trasciende en el mundo, a mis hermanos que me enseñaron la importancia de la constancia y de manera especial a quien hizo de este trabajo un viaje lleno de aprendizaje.

Dayana

AGRADECIMIENTO

Agradezco a la vida que se medió para experimentar y conocer los colores del vivir junto a mi familia que me enseñó a luchar contra todas las injusticias y crear un mundo mejor.

Por la magia de los amigos con los que se convierten las horas en minutos y los días grises en risas incontenibles. Por los maestros y en especial a mi director de tesis que día a día su esfuerzo se ve plasmado en el mañana.

Agradezco aquellos gestos caballerosos hacia mí que permitieron sobrellevar aquellos días en los que el sueño parecía vencerme y los golpes de la vida se sentían con el transcurrir de las horas.

A aquellas travesías hacia el origen infestadas de una multitud de aromas que inundaban en mi derredor, solamente interrumpidos momentáneamente por algún oasis dominical, viajes que tenían un destino, pero para mí no había rumbo, aventuras en las que lo incierto era el gobernante y los nuevos amigos y enseñanzas eran el presente.

Por aquellas vueltas y vueltas y vueltas que tiene la vida y uno tiene que tocar metal para echar buena raíz en tierra fértil con el propósito de dar jugosos frutos con el tiempo. Por las enseñanzas y las amistades que perduran en tiempo.

Por todos aquellos parajes, experiencias y vivencias únicos que la vida me ha permitido experimentar, por dormir placenteramente y despertar totalmente recuperada, por explorar rincones del horizonte desconocidos, por cambiar de perspectiva y no volver a ser jamás la misma de antes, por recibir la bienvenida a un mundo nuevo

ÍNDICE DE CONTENIDO

AVAL I

DECLARACIÓN DE AUTORÍA II

DEDICATORIA III

AGRADECIMIENTO IV

ÍNDICE DE CONTENIDO V

RESUMEN IX

ABSTRACT X

1. INTRODUCCIÓN 1

1.1 OBJETIVOS 1

1.1.1 OBJETIVO GENERAL 1

1.1.2 OBJETIVOS ESPECÍFICOS 1

1.2 ALCANCE 2

2. MARCO TEÓRICO 4

2.1 TÉRMINOS ELÉCTRICOS 4

2.2 MINERÍA DE DATOS 6

2.2.1 TAREAS DE LA MINERÍA DE DATOS 6

2.2.2 PROCESO DE MINERÍA DE DATOS 9

2.3 SERIES DE TIEMPO 12

2.3.2 AUTOCORRELACIÓN 13

2.3.3 AUTOCORRELACIÓN PARCIAL 13

2.3.4 RUIDO BLANCO 13

2.4 MÉTODOS DE PREDICCIÓN 14

2.4.1 REGRESIÓN LINEAL 14

2.4.2 REGRESIÓN POLINOMIAL 16

2.4.3 MODELOS AUTORREGRESIVOS AR 17
2.4.4 MODELOS DE PROMEDIOS MÓVILES MA 18
2.4.5 MODELOS AUTORREGRESIVOS DE PROMEDIO MÓVIL ARMA 19
2.4.6 MODELOS NO ESTACIONALES ARIMA 21

2.5 PRONÓSTICO 25
2.5.2 MÉTODOS DE ESTIMACIÓN DE ERRORES 26
2.5.3 MODELOS ESTACIONALES SARIMA 27

2.6 ESTIMACIÓN Y ORDEN DE SELECCIÓN 29
2.6.1 ESTIMACIÓN DE MÁXIMA VEROSIMILITUD 29
2.6.2 CRITERIO DE INFORMACIÓN 29

3. PRONÓSTICOS A MEDIANO PLAZO 31

3.1 CONOCIMIENTO PREVIO 31

3.2 ENERGÍA MENSUAL 32
3.2.1 PREPARACIÓN DE LOS DATOS DE ENERGÍA 32
3.2.2 MODELAMIENTO DE DATOS MENSUALES DE ENERGÍA..... 33
3.2.3 PRONÓSTICO 39
3.2.4 ANÁLISIS DE RESULTADOS DEL PRONÓSTICO MENSUAL DE ENERGÍA..... 41

3.3 POTENCIA ACTIVA MENSUAL 45
3.3.1 PREPARACIÓN DE LOS DATOS DE POTENCIA ACTIVA 45
3.3.2 MODELAMIENTO DE DATOS MENSUALES DE POTENCIA ACTIVA45
3.3.3 PRONÓSTICO 49
3.3.4 ANÁLISIS DE RESULTADOS DEL PRONÓSTICO MENSUAL DE POTENCIA ACTIVA..... 51

3.4 FACTOR DE CARGA MENSUAL 54

4. PRONÓSTICOS A LARGO PLAZO 56

4.1 ENERGÍA ANUAL 56
4.1.1 PREPARACIÓN DE LOS DATOS DE ENERGÍA ANUAL..... 56
4.1.2 MODELAMIENTO DE DATOS ANUALES DE ENERGÍA 56
4.1.3 PRONÓSTICO 61

4.1.4 ANÁLISIS DE RESULTADOS DEL PRONÓSTICO ANUAL DE ENERGÍA 62

4.2 POTENCIA ACTIVA ANUAL 64

4.2.1 PREPARACIÓN DE LOS DATOS DE POTENCIA ACTIVA ANUAL ... 64

4.2.2 MODELAMIENTO DE DATOS ANUALES DE POTENCIA ACTIVA 64

4.2.3 PRONÓSTICO 69

4.2.4 ANÁLISIS DE RESULTADOS DEL PRONÓSTICO DE POTENCIA ACTIVA ANUAL 70

4.3 POTENCIA REACTIVA ANUAL 72

4.3.1 PREPARACIÓN DE LOS DATOS DE POTENCIA REACTIVA ANUAL 72

4.3.2 MODELAMIENTO DE DATOS ANUALES DE POTENCIA REACTIVA 72

4.3.3 PRONÓSTICO 76

4.3.4 ANÁLISIS DE RESULTADOS DEL PRONÓSTICO ANUAL DE POTENCIA REACTIVA 77

4.4 FACTOR DE POTENCIA 79

4.5 PRONÓSTICO ANUAL DE POTENCIA ACTIVA Y ENERGÍA A TRAVÉS DEL PRONÓSTICO MENSUAL 81

4.6 FACTOR DE CARGA 82

4.7 ANÁLISIS DE METODOLOGÍAS DE PRONÓSTICO 83

4.8 ANÁLISIS COMPARATIVO TENDENCIAL 84

4.8.1 AGENCIA DE REGULACIÓN Y CONTROL DE ELECTRICIDAD 84

4.8.2 CENTRO NACIONAL DE CONTROL DE ENERGÍA 89

5. CONCLUSIONES Y RECOMENDACIONES 96

5.1 CONCLUSIONES 96

5.2 RECOMENDACIONES 97

6. REFERENCIAS BIBLIOGRÁFICAS 99

ANEXO A. DATOS DE ENERGÍA, POTENCIA ACTIVA Y REACTIVA 1

ANEXO B. COMPARATIVA SARIMA Y MACHINE LEARNING PARA ENERGÍA MENSUAL. 1

ANEXO C. Datos de energía, potencia activa y reactiva tabulados (Cd) 2

ANEXO D. Scripts para pronóstico en R (Cd) 2

ANEXO E. Video instructivo (Cd) 2

RESUMEN

El presente trabajo tiene por objetivo conseguir el pronóstico anual a largo plazo de energía, potencia activa y reactiva y el pronóstico mensual a mediano plazo de energía y potencia activa de la Empresa Eléctrica Quito, utilizando los datos históricos de demanda máxima coincidente y energía sin considerar cargas especiales como el ingreso del Metro de Quito.

La minería de datos es la técnica implementada en este trabajo en la cual una de las etapas a considerar es el modelamiento y pronóstico de los datos proporcionados por la Empresa Eléctrica Quito (EEQ).

El trabajo de implementación de los modelos ARIMA Y SARIMA se lo realizó en el lenguaje de programación R en la consola RStudio, para lo cual fue necesario realizar el adecuamiento de los datos, los cuales posteriormente fueron modelados a través de los modelos ARIMA y SARIMA. La selección del modelo requirió una fase de evaluación propia que se la realizó mediante pruebas de residuos.

La predicción de energía, potencia activa y reactiva se realizó luego de la obtención de sus respectivos modelos y se requirió realizar la fase de evaluación de acuerdo con lo dictaminado por la minería de datos.

La etapa de evaluación de los resultados obtenidos se lo realizó con diferentes criterios. El primero de ellos es un criterio estadístico procurando la obtención del error porcentual absoluto medio (MAPE) mínimo. El segundo criterio se basa en conceptos propios de distribución eléctrica, como el factor de potencia, empleando los valores anuales pronosticados de potencia activa y reactiva.

PALABRAS CLAVE: Análisis de series de tiempo, Demanda de electricidad, Minería de datos, pronóstico a mediano y largo plazo.

ABSTRACT

The objective of this work is to obtain the long-term annual forecast of energy, active and reactive power and the medium-term monthly forecast of energy and active power of Empresa Eléctrica Quito, using the historical data of coincident maximum demand and energy without considering special charges such as the entrance of the Quito Metro.

Data mining is the technique implemented in this work in which one of the stages to be considered is the modeling and forecasting of the data provided by Empresa Eléctrica Quito (EEQ).

The implementation work of the ARIMA and SARIMA models was carried out in the programming language R in the RStudio console, for which it was necessary to make the adaptation of the data, which were subsequently modeled through the ARIMA and SARIMA models. The selection of the model required an evaluation phase of its own, which was carried out by means of residue tests.

The prediction of energy, active and reactive power was made after obtaining their respective models and it was required to carry out the evaluation phase in accordance with what was dictated by data mining.

The evaluation stage of the results obtained was done with different criteria. The first one is a statistical criterion trying to obtain the minimum absolute percentage error (MAPE). The second criterion is based on specific concepts of electrical distribution, such as the power factor and coincidence factor, using the predicted annual values of active and reactive power.

KEYWORDS: Data Mining, Forecast Models, Energy market, Electricity Demand, Forecast in the medium and long term.

1. INTRODUCCIÓN

Este trabajo presenta la proyección de demanda de energía, potencia activa y reactiva para la Empresa Eléctrica Quito (EEQ) siendo ésta fundamental para la planificación y la operación de los sistemas eléctricos de distribución y su precisión tiene un gran impacto en el servicio eléctrico [1].

Una mala estimación de la demanda puede generar inversiones en proyectos que excedan la capacidad requerida o una subestimación que puede conllevar a un incumplimiento de la demanda, afectando el servicio de varios usuarios y su productividad [2].

La EEQ emplea modelos de regresión polinomial para sus pronósticos donde se verifica únicamente el coeficiente de determinación R^2. El implementar nuevos modelos para la estimación de la demanda, permitirá realizar una comparación de resultados a través de indicadores de error con el método actual, de esta manera se plantea obtener un mejor pronóstico con una visión más certera del crecimiento de la demanda. La construcción de los modelos de energía, potencia activa y reactiva en un software libre evitará costos de compra y mantenimiento de este servicio.

1.1 OBJETIVOS

1.1.1 OBJETIVO GENERAL

- Desarrollar modelos ARIMA y SARIMA en base a la información histórica recopilada y utilizando el Lenguaje de Programación R, para la predicción mensual y anual de la demanda eléctrica del sistema de la Empresa Eléctrica Quito.

1.1.2 OBJETIVOS ESPECÍFICOS

- Analizar diferentes métodos de predicción de demanda eléctrica, así como la base teórica que fundamenta el desarrollo y funcionamiento de los denominados modelos ARIMA y SARIMA utilizados también como herramientas de predicción.

- Recopilar los datos históricos disponibles de la demanda eléctrica de la EEQ para su posterior procesamiento a través de minería de datos que permita obtener la información adecuada para la etapa de modelamiento.

- Construir los modelos ARIMA y SARIMA para la predicción de la demanda eléctrica mensual y anual de la EEQ conforme lo establecido en el alcance, utilizando el lenguaje de programación R.

- Comparar los resultados obtenidos con los modelos propuestos y el método actual empleado por la EEQ y analizar la tendencia de potencia activa y energía con respecto a los datos históricos del CENACE y a los resultados del pronóstico realizado por ARCONEL.

1.2 ALCANCE

Para el desarrollo del presente trabajo, en primera instancia se recopilarán los datos históricos de energía, factor de potencia, potencia activa y reactiva que se encuentran almacenados en la base de datos de una empresa distribuidora en el Ecuador. En particular, se usarán los datos históricos que dispone la Empresa Eléctrica Quito hasta diciembre del 2016 como datos de aprendizaje. Luego de ello, se realizará la minería de datos correspondiente para identificar y eliminar inconsistencias, discrepancias o errores en los datos, para posteriormente generar el conocimiento (modelos).

A continuación, en el programa RStudio, se construirán modelos mensuales ARIMA y SARIMA con los datos históricos que fueron procesados, modelos que permitirán realizar una proyección mensual de potencia activa y energía hasta diciembre del 2018. Estos modelos serán desarrollados considerando una disminución del error porcentual absoluto medio (MAPE), del error cuadrático medio (RMSE) y del error absoluto medio (MAE), con respecto a los modelos vigentes de la Empresa Eléctrica Quito, lo que brindará mayor confianza en la predicción de la demanda eléctrica. Con los resultados del pronóstico de potencia activa y energía se obtendrá el factor de carga mensual, que será contrastado mediante el MAPE con los valores reales obtenidos hasta el mes de mayo de 2018 (información que actualmente se dispone), esto permitirá realizar un análisis de los resultados obtenidos con los valores históricos y con los valores esperados en base a la experiencia.

Adicionalmente, se construirán modelos ARIMA y SARIMA para el pronóstico anual de potencia activa, reactiva y energía, que tendrá un horizonte de 10 años (2017-2026), en base a los datos anuales históricos obtenidos hasta el año 2016 y a través del pronóstico mensual obtenido previamente.

La validación de los modelos mensuales se realizará con los datos reales de enero del 2017 hasta mayo del 2018, y para los modelos anuales se tomará en cuenta la información real del año 2017.

A partir de los resultados de potencia activa, reactiva y energía anual, se obtendrá el factor de potencia y de carga, tomando en cuenta su comportamiento histórico.

Considerando que la EEQ muestra una gran incidencia en el comportamiento del Sistema Nacional Interconectado, se realizará el análisis tendencial con los datos históricos mensuales de potencia y energía publicados por el Operador Nacional de Electricidad CENACE desde el año 2000 al 2016 y los modelos ajustados obtenidos en RStudio. De igual forma se comparará mediante un análisis tendencial con datos anuales del pronóstico de potencia y energía presentado en el Plan Maestro de Electricidad 2016-2025 de la Agencia de Regulación y Control de Electricidad ARCONEL.

2. MARCO TEÓRICO

En este capítulo se presentan conceptos fundamentales que preparan al lector y cimentan bases requeridas para la comprensión de este proyecto de titulación.

2.1 TÉRMINOS ELÉCTRICOS

Energía

La energía se define como la capacidad de realizar trabajo y se origina por el flujo de carga eléctrica. La energía eléctrica es una forma de energía muy valiosa debido a que puede ser transmitida, transformarse en otras formas de energía, no puede ser almacenada y para su entrega se realiza mediante un sistema de conductores y de control [3]. Se mide en vatios por hora [Wh].

Potencia instantánea

En un circuito de corriente alterna la potencia instantánea *p(t)* se define como el producto del voltaje instantáneo *v(t)* y la corriente instantánea *i*(*t*). El voltaje instantáneo y la corriente instantánea son sinusoidales y su valor promedio es cero, sin embargo la potencia instantánea en su valor promedio no es cero [4].

Potencia activa

La potencia activa se define como el valor medio de la integral de la potencia instantánea a lo largo de un periodo, que físicamente significa la potencia útil que se transmite. La potencia activa es el valor neto de la potencia a lo largo de cada periodo y su aprovechamiento se realiza al convertirse en potencia térmica o mecánica, es decir indica la transformación de energía de forma eléctrica a no eléctrica [5]. Medida en vatios [W].

Potencia reactiva

La potencia reactiva es una componente de la potencia instantánea contiene el término *sen θ* que es negativo para el circuito capacitivo y positivo para el circuito inductivo, este componente viaja hacia adelante y hacia atrás en la línea sin realizar ningún trabajo útil [5]. Tanto P como Q tienen las mismas dimensiones de vatios, pero para enfatizar el hecho de que Q representa una potencia no activa, se mide en términos de reactivos de voltamperios, es decir [VAr] [5].

Potencia aparente

La potencia aparente de un dipolo con ondas periódicas cualesquiera, semejantes a las de un régimen estacionario sinusoidal, es el producto de los valores eficaces del voltaje y de la corriente [6].

Triángulo de potencias

La Figura 2.1 representa la relación existente entre la potencia activa, reactiva y aparente. El ángulo formado entre la potencia activa y aparente se denomina θ.

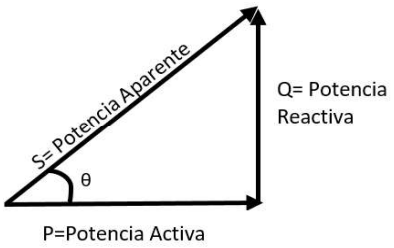

Figura 2.1. Triángulo de potencias [7].

Factor de potencia

El factor de potencia indica el grado de aprovechamiento que recibe el dipolo o circuito eléctrico para convertirla en potencia útil, es decir muestra la relación entre la potencia activa y aparente descrita en la Ecuación 2.1.

$$fp = \frac{P}{S} \tag{2.1}$$

Mediante el triángulo de potencias mostrado en la Figura 2.1 se puede encontrar la Ecuación 2.2 para determinar el valor del ángulo θ en función de la potencia activa y reactiva.

$$\theta = \tan^{-1}\frac{Q}{P} \tag{2.2}$$

El factor de potencia se define como el coseno del ángulo θ y es el ángulo por el cual la corriente de línea fundamental se retrasa con respecto al voltaje de la fuente medida entre la línea y neutro

$$\mathrm{fp} = \cos(\theta) \tag{2.3}$$

Mediante la Ecuación 2.2 y Ecuación 2.3 se escribe el factor de potencia en términos de potencia reactiva y activa.

$$\mathrm{fp} = \cos\left(\tan^{-1}\frac{Q}{P}\right) \tag{2.4}$$

La Ecuación 2.4 permite calcular el factor de potencia a través de los datos de potencia activa y reactiva.

Demanda

La demanda es la medida de la tasa de uso de energía eléctrica y generalmente se mide en vatios hora [Wh] y sus múltiplos [8].

Demanda máxima coincidente

La demanda máxima coincidente es la suma de las demandas individuales en un período de tiempo dado [9].

Factor de carga

El factor de carga (FC) se define como la relación entre la demanda promedio en un intervalo de tiempo t y la demanda máxima observada en ese intervalo de tiempo. Al considerar un intervalo de tiempo igual a un año se obtiene la Ecuación 2.5 [10].

$$\mathrm{FC} = \frac{Consumo\ de\ energía\ \left(\frac{MWh}{anual}\right)}{Potencia\ máxima\ (MW) \times 8760} \times 100\% \tag{2.5}$$

2.2 MINERÍA DE DATOS

La minería de datos o exploración de datos es el proceso de extracción de conocimiento y en términos simples es encontrar patrones útiles en un conjunto de datos. También se la conoce como el descubrimiento de conocimiento, autoaprendizaje y análisis predictivo [12]. Combinando el conocimiento estadístico, experiencia, tecnologías de bases de datos y las técnicas de aprendizaje automático para extraer información significativa y útil de los datos.

La minería de datos bajo un enfoque estadístico muestra modelos que describen como una o más variables de los datos están relacionadas con otras variables. la Figura 2.2 sintetiza el análisis predictivo donde las entradas son datos históricos y como salidas se obtiene el modelo y el pronóstico.

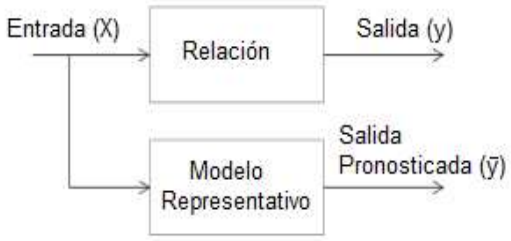

Figura 2.2. Modelo representativo para análisis predictivo [13].

2.2.1 TAREAS DE LA MINERÍA DE DATOS

Se puede agrupar los problemas de minería de datos en tareas de clasificación, regresión, análisis de asociación, detección de anomalías, series de tiempo y minería de texto como se muestra en la Figura 2.3 [14].

Los siguientes conceptos fueron tomados de [15]:

- Las técnicas de clasificación y regresión pronostican el valor de una variable en función de sus variables de entrada, por ejemplo, la decisión de aprobación de un préstamo.
- La agrupación identifica agrupaciones naturales en el conjunto de datos, estas asociaciones pueden utilizarse para la segmentación del mercado. El usuario final deberá investigar los motivos de las formaciones de estos grupos.
- El análisis de asociación comúnmente se lo realiza en un análisis comercial para identificar pares de artículos que se compran juntos, de modo que artículos específicos se pueden agrupar o colocar un junto a otro.

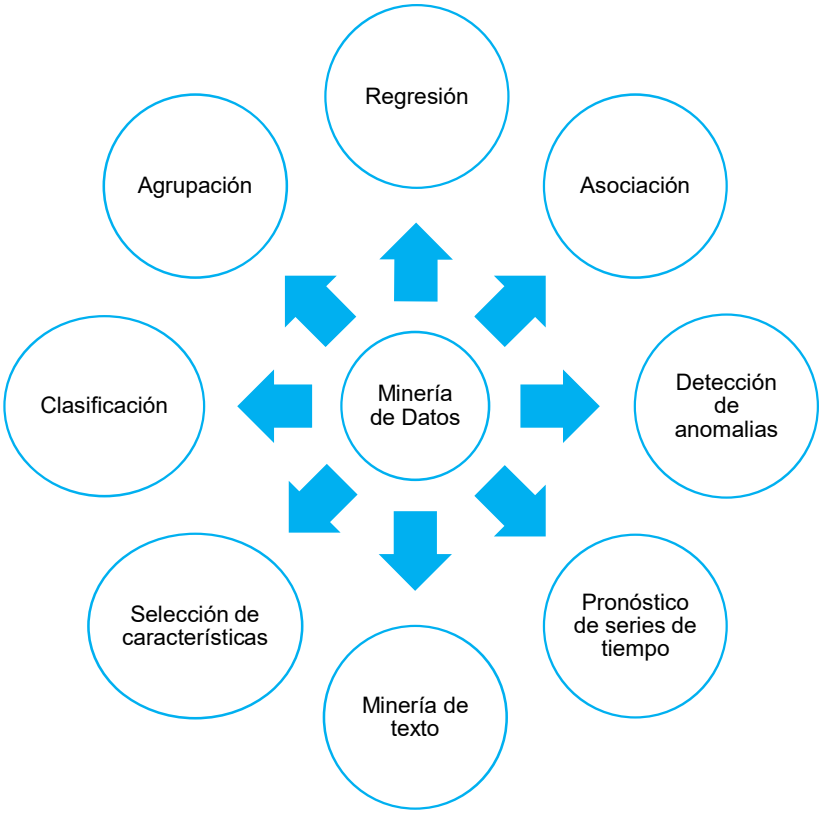

Figura 2.3. Tareas de la minería de datos [13].

- La identificación de valores atípicos o anomalías detecta datos que son significativamente diferentes de otros datos en un conjunto. La detección de fraude en tarjetas de crédito es la aplicación más empleada.

- El pronóstico de series de tiempo es un uso especial del modelado de regresión (los modelos pronostican un valor futuro de una variable en base a valores pasados).
- En la minería de texto los datos de entrada son texto (documentos, mensajes, correos electrónicos, etc.). Para la extracción de datos de texto, estos archivos se transforman en vectores de documentos donde cada palabra es un atributo.
- La selección de características procesa los atributos de un conjunto de datos disminuyéndolos y finalmente se mostrará solo los atributos que realmente sean de importancia.

La Tabla 2.1 sintetiza las tareas de minería de datos con técnicas algorítmicas de uso frecuente.

Tabla 2.1. Tareas de Minería de Datos y algoritmos de uso frecuente [15].

Tarea	Descripción	Algoritmo	Ejemplo
Clasificación	Predice si un dato pertenece a una clase predefinida. La selección puede fundamentarse en datos conocidos.	• Árbol de decisión • Red Neuronal • Modelos Bayesianos • Reglas de inducción • Vecino más cercano	Agrupar nuevos clientes en uno de los grupos de clientes ya conocidos.
Regresión	Predecir un dato numérico basándose en un conjunto de datos conocidos.	• Regresión Lineal • Regresión logística	Predecir la tasa de desempleo para el próximo año.
Detección de Anomalías	Predecir si un dato es atípico en comparación a otros puntos de datos en el conjunto de datos.	• Distancia basada en densidad • Factor atípico local basado (LOF).	Detección de transacciones de fraude en tarjetas de crédito.
Series de Tiempo	Predecir el valor futuro de la variable objetivo en función de los valores históricos de la variable.	• Suavizado Exponencial • ARIMA	Predicción de ventas, crecimiento de la demanda energética, fenómenos de crecimiento.

Asociación	Identificar asociaciones naturales dentro del conjunto de datos en función de las propiedades heredadas.	• k-means • Asociación basada en la densidad	Encontrar segmentos de clientes en una empresa basados en datos de transacciones, web y llamadas de clientes.
Análisis de Asociación	Reconocer relaciones dentro de un conjunto de elementos basados en datos de transacción.	• Algoritmo de crecimiento de patrones frecuentes (FP-Growth) • Algoritmo a priori	Encontrar oportunidades de venta cruzada para un minorista según el historial de compras de transacciones.

2.2.2 PROCESO DE MINERÍA DE DATOS

El objetivo de la minería de datos es encontrar relaciones y patrones útiles en los datos mediante un conjunto de actividades iterativas [15]. El procedimiento general implica:

1. Comprensión el problema.
2. Preparación de las muestras de datos.
3. Desarrollo del modelo que simule el comportamiento de los datos de entrenamiento.
4. Aplicación del modelo en un conjunto de datos y comparar los resultados con los del mundo real.
5. Implementación a la producción.

El modelo estándar de proceso más popular para minería de datos es CRISP-DM (Cross Industry Standard Process for Data Mining), la Figura 2.4 proporciona una visión general de este proceso y muestra que los pasos son dependientes unos de otros y empiezan con el conocimiento del negocio, algoritmo y herramienta de minería de datos [16].

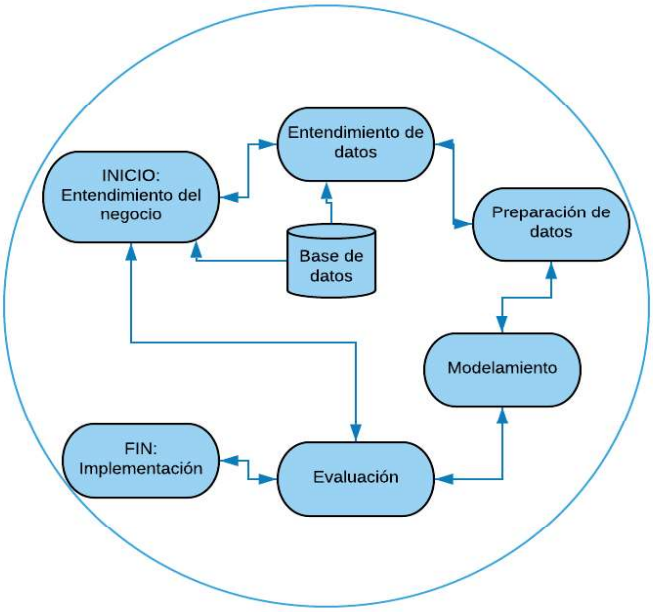

Figura 2.4. Proceso CRISP-M

El proceso de minería de datos representado en la Figura 2.5 comienza con una necesidad de análisis, una pregunta o un objetivo comercial.

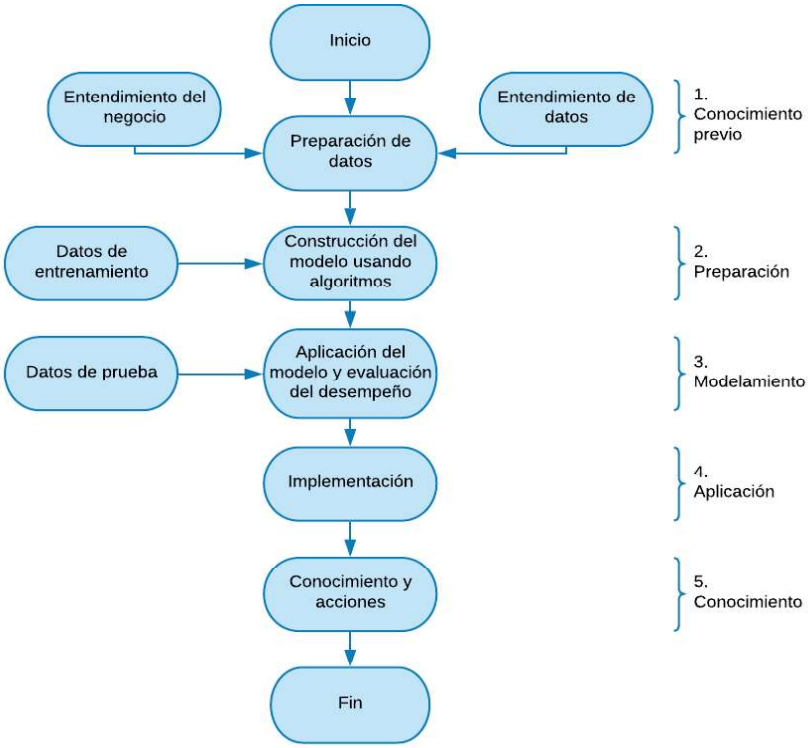

Figura 2.5. Proceso de Minería de Datos

Paso 1: el conocimiento previo es toda la información que ya se conoce sobre un tema. Este conocimiento previo ayuda a responder las preguntas:

- ¿cuál es el problema que se solucionará?
- ¿cómo encaja en el contexto comercial?
- ¿qué datos se necesitan para resolver este problema?

Sin un entendimiento claro de la problemática será imposible obtener el conjunto de datos correcto; el proceso de minería de datos es secuencial por lo que se deberá regresar al paso anterior y revisar las suposiciones, el enfoque y las tácticas. De manera obligatoria se debe cumplir el objetivo de todo el proceso.

Paso 2: se refiere a la preparación de datos, comúnmente se utiliza Microsoft Excel, aunque existe diferentes programas computacionales como Rapid Miner, Weka, Orange, Sisense, entre otros. La preparación del conjunto de datos para entrenamiento es una tarea ardua, muy rara vez los datos se encuentran disponibles en la forma requerida. Los algoritmos en su mayoría requieren estar estructurados en filas y sus atributos en columnas. Si los datos se encuentran de diferente manera se necesitará transformarlos aplicando funciones de pivote o transposición. También se deberá considerar datos que no se hayan capturado y se encuentren vacíos.

Paso 3: es la parte más visible y de mayor discusión, aquí se realiza el modelado de los datos lo que implica la creación de modelos representativos a partir de un conjunto de datos o muestras para realizar pronósticos. La fase de modelado tiene ciertos pasos a seguir para la extracción de datos predictivos que se muestran en la Figura 2.6.

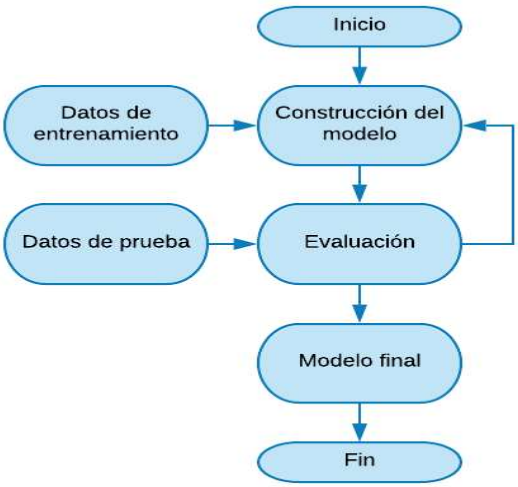

Figura 2.6. Pasos del modelamiento [13].

El desarrollo de un modelo estable requiere un conjunto de datos previamente preparado que es separado en dos grupos, el primero es llamado conjunto de datos de entrenamiento y se crea el modelo a partir de este. El segundo conjunto son los datos de prueba para verificar la validez del modelo.

Paso 4: Una vez obtenido el modelo final en el paso anterior, se implementará dicho modelo en el sitio de estudio.

Paso 5: La implementación arroja nueva información que retroalimenta al modelo y genera conocimiento acerca del objeto en estudio.

2.3 SERIES DE TIEMPO

Según [17]: Todo lo que sea observado secuencialmente en un determinado tiempo es llamado una serie de tiempo. Las mediciones generalmente serán realizadas en intervalos de tiempo fijos que pueden presentar relaciones con otras, es decir están correlacionadas y estas dependencias pueden emplearse para predecir valores futuros.

Descomposición

El análisis de una serie de tiempo es realizado a través de los componentes que influyen en ella, este procedimiento de identificación es llamado descomposición.

Tendencia

Una tendencia se presenta en un largo periodo de tiempo, los datos muestran un crecimiento o decrecimiento. Este componente se lo representa con la letra *T*.

Componente cíclico

Un ciclo ocurre cuando los datos se elevan o decaen en frecuencias que no son fijas. Usualmente un componente cíclico es difícil de identificar, por lo tanto, se le considera como parte de la tendencia. Su notación es la letra *C*.

Estacionalidad

El patrón estacional ocurre cuando una serie de tiempo es afectada por patrones estacionales como las estaciones del año que se mantienen más o menos estables y aparecen anualmente. Estos patrones pueden ser trimestrales, mensuales o semanales y pueden ser influenciados por eventos en el calendario como vacaciones y feriados.

Se lo representa con la letra S de su traducción en inglés “seasonal”.

Componente irregular

El componente irregular de una serie de tiempo se presenta por fluctuaciones impredecibles o aleatorias. Un componente irregular individual no aporta significativamente

a la predicción, pero la combinación de varios puede afectar considerablemente. Se representa con la letra *I* [17].

2.3.2 AUTOCORRELACIÓN

Mientras la correlación mide la relación lineal que existe entre dos variables, la autocorrelación mide la relación lineal existente entre valores rezagados de una misma serie de tiempo.

Las series de tiempo que presentan tendencia y estacionalidad pueden ser analizadas mediante los coeficientes de autocorrelación que deben ser graficados en un correlograma o función de autocorrelación (ACF). La estimación de los coeficientes de autocorrelación r_k se lo realiza mediante la Ecuación 2.6.

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \quad k = 0,1,2, \tag{2.6}$$

Donde

r_k= *coeficiente de autocorrelación para un retraso de* k *periodos*

$\bar{Y}$= *media de los valores de la serie*

Y_t= *observación en el periodo* t

Y_{t-k}= *observación* k *periodos anteriores o durante un periodo* $t - k$

n= *tamaño de la serie*

2.3.3 AUTOCORRELACIÓN PARCIAL

La correlación parcial entre dos variables se define como la correlación de dos variables mientras se controla una tercera variable, es decir, una medida de correlación entre las variables X e Y con una tercera variable Z como medida de la relación directa entre X e Y. El cálculo de la autocorrelación parcial entre X e Y dado Z sirve para identificar en qué medida la relación lineal entre X e Y puede deberse a la correlación de dos variables con Z [19].

2.3.4 RUIDO BLANCO

Una serie de tiempo que no muestra autocorrelación es llamada ruido blanco, los valores de ACF se esperan que sean aproximadamente cero. El modelo de ruido blanco es representado por la Ecuación 2.7.

$$Y_t = c + \varepsilon_t \tag{2.7}$$

Donde

Y_t = *observación en el periodo t*

c = *nivel general*

ε_t = *componente de error aleatorio*

La Figura 2.7 muestra los valores de autocorrelación obtenidos de una serie de tiempo de ruido blanco, cumpliendo con la definición de ruido blanco de no presentar autocorrelación.

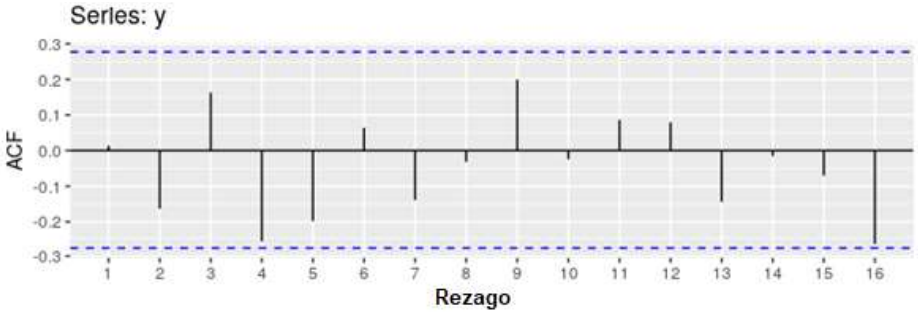

Figura 2.7. Valores de ACF de una serie de Ruido Blanco [20].

2.4 MÉTODOS DE PREDICCIÓN

Pronosticar es predecir el futuro tan precisamente como sea posible y se convierte en la base fundamental para la toma de decisiones.

Los métodos de pronóstico se pueden clasificar en cualitativos y cuantitativos.

Los métodos de pronósticos cualitativos se basan en la experticia de los individuos o grupos. Estos métodos generalmente no se basan en series de datos históricos y se basan únicamente en el "juicio" del pronosticador [21].

A continuación, se presentan algunas técnicas de predicción cuantitativas que generalmente emplean bases de datos históricos extensas y sus procedimientos son mecánicos.

2.4.1 REGRESIÓN LINEAL

El concepto básico de regresión lineal es la existencia de una relación lineal entre la variable a pronosticar *Y* con la serie de tiempo *X*. Al tener un conjunto de datos *X*-*Y* la línea que mejor se ajusta a ellos es la que minimiza la suma de cuadrados de las distancias entre estos puntos y la línea, esta distancia se mide verticalmente en el sentido de *Y*. El modelo estadístico de la regresión lineal es representado por la Ecuación 2.8.

$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \tag{2.8}$$

Donde

Y_t = *observación en el periodo t*

β_0 = *intersección con el eje Y o intercepto*

β_1 = *pendiente de la recta ajustada*

x_t = *periodo t*

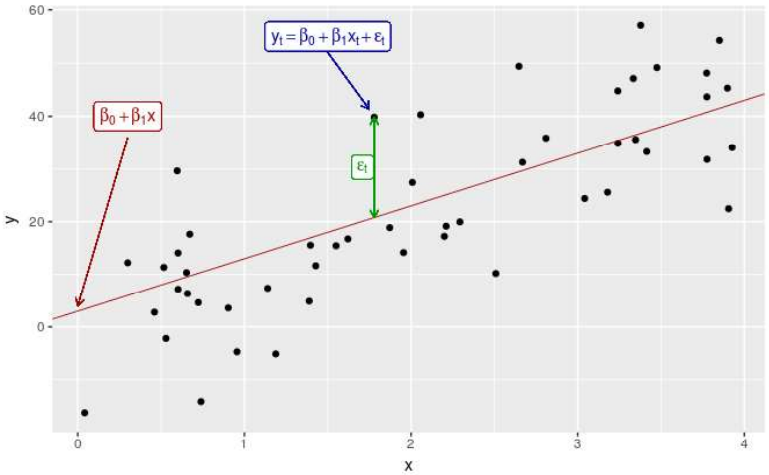

Figura 2.8. Modelo de regresión lineal y sus componente [20].

En la Figura 2.8 se observa que el componente ε_t no implica un error, pero es una desviación de la línea que se crea en el modelo.

Coeficiente de determinación

El coeficiente de determinación R^2 refleja el grado ajuste de un modelo con respecto a la variable que pretende modelizar y es se calcula a partir de la Ecuación 2.9 [22].

$$R^2 = \frac{\sum_{t=1}^{T}(\bar{\bar{Y}} - \hat{Y})^2}{\sum_{t=1}^{T}(Y - \hat{Y})^2} \tag{2.9}$$

Donde

$\bar{\bar{Y}}$ = *estimación de un modelo sobre lo que según las variables explicativas vale Y, pero no es el valor real de Y, sino una estimación de Y*

$\hat{Y}$ = *pronóstico de* Y

Este valor oscila entre 0 y 1, el mayor ajuste se obtendrá cuando valor se aproxime a 1 y este valor hará al modelo fiable.

Coeficiente de Pearson

El coeficiente de correlación *r* de Pearson mide el grado de asociación lineal entre dos variables, es decir mide la dispersión de los puntos en torno a una tendencia lineal subyacente y se determina mediante la Ecuación 2.10.

$$r = \frac{XY}{\sqrt{(XX)(YY)}} \tag{2.10}$$

Donde

r = coeficiente de Pearson

X = variable dependiente

Y = variable independiente

El valor de *r* puede situarse entre -1 y +1. La prueba de significación se hace con la hipótesis nula de que no hay asociación, $r = 0$. Una correlación será perfecta cunado el valor de *r* se encuentre en el rango de 1 a 0,96 [23].

2.4.2 REGRESIÓN POLINOMIAL

Los modelos de regresión polinomial tienen como variable independiente al tiempo cuadrado, tiempo al cubo, etc. El problema que presenta la regresión polinomial es la correlación perfecta que existe entre las variables del modelo, a medida que se agregan términos, las regresiones polinomiales de baja dimensión no son lo suficientemente flexibles como para capturar cambios repentinos de la pendiente, especialmente en intervalos irregulares. En grandes dimensiones, la regresión polinomial tiende a fallar debido a la correlación perfecta [24].

La forma de extender la regresión lineal a relaciones no lineales o polinomiales es reemplazando el modelo de la Ecuación 2.8 por la función polinomial de la Ecuación 2.11.

$$y_t = \beta_o + \beta_1 x_t + \beta_2 {x_t}^2 + \beta_3 {x_t}^3 + \cdots + \beta_d {x_t}^d + \varepsilon_t \tag{2.11}$$

Los coeficientes pueden ser fácilmente estimados empleando la regresión lineal de mínimos cuadrados. Se recomienda que *d* no sobrepase los valores de 3, debido a que la gráfica del polinomio puede volverse demasiado flexible y puede adoptar formas muy extrañas [25].

Estimación de mínimos cuadrados

El principio de estimación de mínimos cuadrados provee una manera de escoger estos coeficientes $\beta_o, \beta_1, \ldots, \beta_d$, de manera efectiva, mediante la minimización de la suma de los errores al cuadrado como se muestra en la Ecuación 2.12.

$$\sum_{t=1}^{T} \varepsilon_t^2 = \sum_{t=1}^{T} \left(y_t - \beta_o - \beta_1 x_t - \beta_2 {x_t}^2 - \cdots - \beta_d {x_t}^d\right)^2 \tag{2.12}$$

Donde

T = *tamaño de la serie de tiempo*

Encontrar los mejores estimadores de los coeficientes es adecuar el modelo a los datos o también se le puede llamar entrenamiento del modelo.

2.4.3 MODELOS AUTORREGRESIVOS AR

Los modelos autorregresivos formulan un pronóstico en función de observaciones o mediciones previas de la serie de tiempo. Los modelos autorregresivos AR son parte de los modelos de promedio móvil integrado autorregresivo ARIMA.

Un modelo autorregresivo de orden AR(p) puede ser escrito mediante la Ecuación 2.13.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{2.13}$$

Donde

y_t = *variable dependiente en el tiempo* t

c = *nivel constante de la serie o nivel general*

$y_{t-1}, y_{t-2}, \dots, y_{t-p}$ = *variable independiente en los retrasos de tiempo* $t-1, t-2, \dots, t-p$, *respectivamente, estas desempeñan el rol de variables independientes.*

$\phi_o, \phi_1, \phi_2, \cdots, \phi_p$ = *coeficientes*

Los modelos autorregresivos son adecuados para series de tiempo que presentan estacionariedad, si una serie de tiempo presentara tendencia deberá ser eliminada previamente empleando la diferenciación. La Figura 2.9 muestra dos series de tiempo diferentes ajustadas a través de los modelos AR (1) Y AR (2).

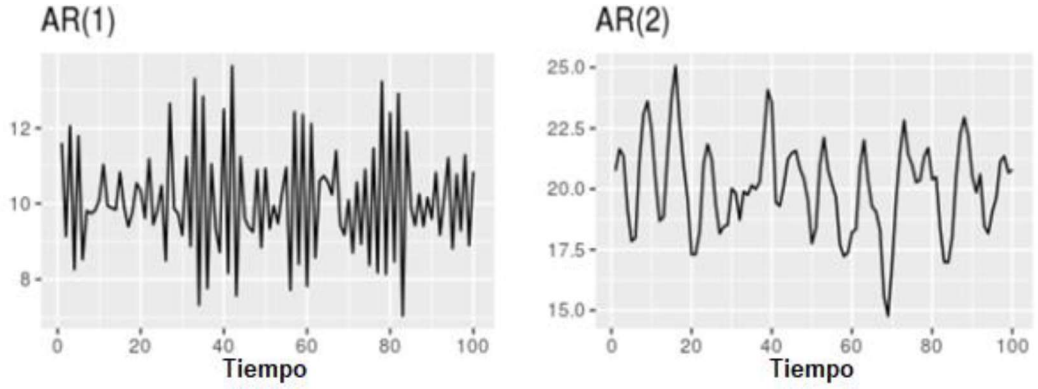

Figura 2.9. Modelos autorregresivos [20].

Al analizar el modelo AR (2) que es descrito mediante la Ecuación 2.14, se muestra el valor y_t o valor pronosticado en función únicamente de los valores observados dos periodos anteriores, es decir el pronóstico de un modelo AR (p) dependerá únicamente de sus p observaciones anteriores.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t \tag{2.14}$$

El nivel de la serie c depende únicamente de los valores de los coeficientes $\phi_1, \phi_2, \cdots, \phi_p$ y de la media de la serie de tiempo tal como se muestra en la Ecuación 2.15.

$$c = \left(1 - \sum_{i=1}^{p} \phi_i\right)\mu \tag{2.15}$$

Donde μ es la media de la serie de tiempo y puede ser determinada mediante la Ecuación 2.16

$$\mu = \frac{\sum_{i=0}^{n} y_i}{n} \tag{2.16}$$

2.4.4 MODELOS DE PROMEDIOS MÓVILES MA

La diferencia primordial existente entre los modelos autorregresivo AR y los modelos de promedios móviles MA radica en los valores empleados, en el primer caso son valores anteriores de la serie, mientras que el modelo de promedio móviles usa los errores anteriores del pronóstico en su modelo de regresión lineal tal como se muestra en la Ecuación 2.17.

$$y_t = \mu + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + \varepsilon_t \tag{2.17}$$

Donde

y_t = *variable dependiente en el tiempo* t

μ = *promedio o media de la serie de tiempo*

$e_{t-1}, e_{t-2}, \ldots, e_{t-p}$ = *errores en periodos anteriores*

$\theta_o, \theta_1, \theta_2, \cdots, \theta_p$ = *coeficientes*

Los modelos de promedio móvil (MA), emplean una combinación lineal de un número finito de errores de periodos anteriores. Al realizar la suma de los coeficientes $\theta_1, \theta_2, \cdots, \theta_p$ no necesariamente será 1 y estos pueden ser valores positivos o negativos.

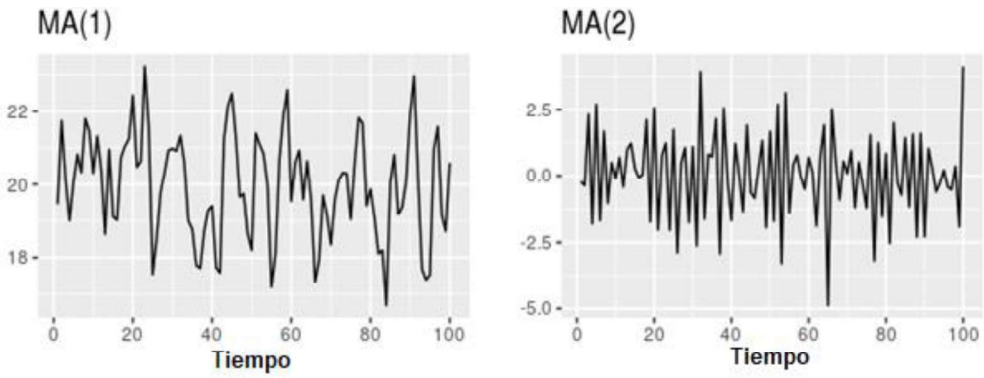

Figura 2.10. Ejemplo de modelos de promedio móvil [20].

En la Figura 2.10 muestra dos series de tiempo pertenecientes al modelo MA (1) y MA (2). Al cambiar los coeficientes $\theta_1, \theta_2, \cdots, \theta_p$ da como resultado un diferente patrón de la serie de tiempo.

En los modelos autorregresivos la variación del término del error ε_t, cambia únicamente la escala de la serie de tiempo y no afecta en absolutamente en nada a su patrón.

La Ecuación 2.18 fue obtenida mediante la Ecuación 2.17 considerando $q = 1$ para la obtención de un modelo MA (1) que depende del error anterior.

$$y_t = \theta_1 e_{t-1} + \varepsilon_t \tag{2.18}$$

2.4.5 MODELOS AUTORREGRESIVOS DE PROMEDIO MÓVIL ARMA

Los modelos autorregresivos de promedio móvil son aquellos que incluyen observaciones y errores de observaciones anteriores, es decir incluyen términos autorregresivos AR y de promedio móvil MA. Los modelos autorregresivos de promedio móvil serán identificados con notación ARMA(p, q), donde p es el orden del componente autorregresivo y q es el orden de la parte del promedio móvil y su modelo esta descrito mediante la Ecuación 2.19.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \tag{2.19}$$

Los modelos ARMA(p, q) abarcan a los modelos autorregresivos y de promedio móvil, es decir si obtiene un modelo, donde $q = 0$ se notaría como ARMA$(p, 0)$ y puede ser reducido a un modelo puro AR(p). De forma similar en un modelo ARMA(p, q) donde $p = 0$, podrá ser descrito como un modelo de promedio móvil de orden q, MA(q).

Los pronósticos que se generan a partir de un modelo ARMA(p, q), integran las observaciones actuales y pasadas de la respuesta, y, así como los errores actuales y pasados de los errores.

El proceso de identificación de modelos de promedio móvil autorregresivos se lo realiza mediante la observación del comportamiento de los valores de autocorrelación y autocorrelación parcial [27]. Los coeficientes ACF y PACF de los modelos AR, MA ya ARMA se resumen en la Tabla 2.2.

Tabla 2.2. ACF y PACF para modelos AR(p), MA(q) Y ARMA(p,q).

	ACF	**PACF**
AR(p)	Decaen exponencialmente	Es cero después de p
MA(q)	Es cero después de q	Decaen exponencialmente
ARMA (p,q).	Decaen exponencialmente o como una onda sinusoidal amortiguada a partir de un retraso *q*	Decaen exponencialmente o como una onda sinusoidal amortiguada a partir de un retraso *p*

Dickey Fuller

La prueba de Dickey Fuller permite determinar la existencia de una autocorrelación. Las autocorrelaciones indican la omisión de variables relevantes.

Mediante la Ecuación 2.13 se puede obtener el modelo autorregresivo AR(1).

$$y_t = c + \phi_1 y_{t-1} + \varepsilon_t \tag{2.20}$$

Mediante la Ecuación 2.20 se puede obtener la Ecuación 2.21 que permite entender la prueba de Dickey Fuller.

$$y_t = \mu + p y_{t-1} + \varepsilon_t \tag{2.21}$$

Donde μ y p son parámetros que se estiman. Debido a que p es un coeficiente de autocorrelación debe tomar valores entre $-1 < p < 1$, si $p = 1$, la serie que se encuentra en análisis es no estacionaria. Si el valor de p supera a la unidad, entonces la serie es explosiva (decrece rápidamente desmesuradamente).

La prueba ampliada de Dickey Fuller (ADF), permite considerar otros esquemas de autocorrelación, añadiendo términos diferenciados de la variable dependiente y en el lado derecho de la regresión [28].

La prueba de Dickey Fuller considera que μ no esta correlacionado, mientras que el ADF si contempla esta posibilidad.

2.4.6 MODELOS NO ESTACIONALES ARIMA

Diferenciación

Una serie que presente tendencia será llamada no estacionaria y sus propiedades dependen del tiempo en la que la serie sea observada. El ruido blanco es estacionario y no importa cuando se lo observe, este debe ser muy similar en cualquier periodo de tiempo.

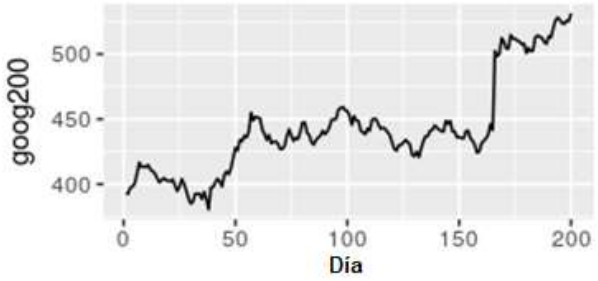

Figura 2.11. Precio de las acciones de Google durante 200 días consecutivos [20].

La Figura 2.11 representa una serie de tiempo no estacionaria y fácilmente se nota que se debe a la tendencia que presenta, por otra parte, la Figura 2.12 muestra una serie estacionaria que representa el cambio diario del precio de las acciones de Google durante 200 días obtenida mediante la diferencia entre las observaciones consecutivas, a este proceso se le conoce como diferenciación.

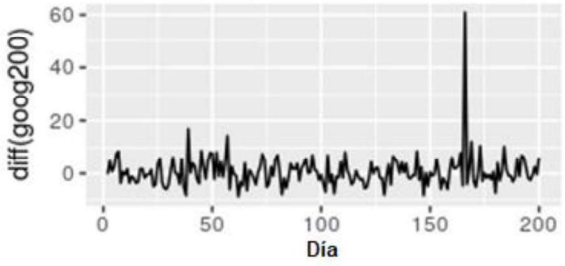

Figura 2.12. Cambio diario del precio de las acciones de Google durante 200 días consecutivos [20].

Transformaciones como el logaritmo y diferenciación pueden ayudar a estabilizar la varianza de la serie de tiempo y remover cambios en el nivel de la serie de tiempo, eliminando o reduciendo la tendencia.

Para identificar una serie estacionaria es conveniente emplear un correlograma con los valores de ACF. Para una serie estacionaria los valores de ACF decaerán a cero de manera rápida, mientras que los valores de ACF para una serie no estacionaria incrementarán lentamente.

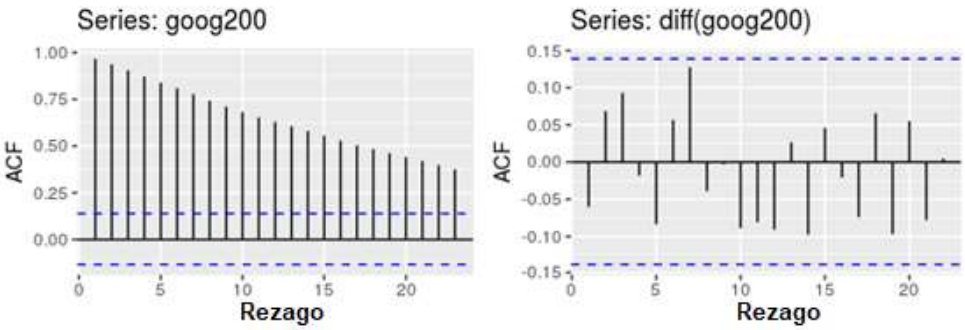

Figura 2.13. Valores de ACF para el precio (derecha) y el cambio diario (izquierda) de precio de las acciones de Google durante 200 días consecutivos [20].

La Figura 2.13 de la izquierda muestra el comportamiento de los valores ACF para una serie no estacionaria, la figura de la derecha corresponde a los valores de ACF de la primera serie luego de haber sido diferenciada (eliminada la tendencia).

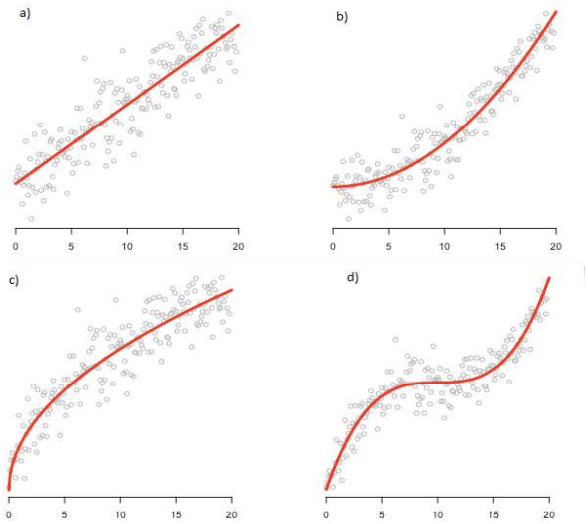

Figura 2.14. Series de tiempo y sus tendencias a) lineal, b) cuadrática c) exponencial y d) cúbica [29].

La Figura 2.14 muestra distintas series de tiempo con sus tendencias que deben ser removidas mediante la diferenciación, si se trata de una tendencia lineal se requiere diferenciar la serie una vez, una serie que presente una tendencia cuadrática debe ser diferenciada dos veces y una tendencia cúbica debe ser diferenciada tres veces. Para remover una tendencia exponencial se debe emplear el logaritmo a la serie. Para notar que una serie fue diferenciada se emplea la notación y'.

Una diferencia estacional es la diferencia entre una observación y su correspondiente observación del año anterior y se describe matemáticamente con la Ecuación 2.22.

$$y_t' = y_t - y_{t-m} \tag{2.22}$$

Donde

m = *número de estaciones*

Para el proceso de diferenciación es conveniente emplear el operador de retroceso B definido mediante la Ecuación 2.23

$$By_t = y_{t-1} \tag{2.23}$$

El operador B al interactuar con y_t tiene como efecto desplazar los datos un periodo hacia atrás. Al emplear dos veces el operador B sobre y_t, desplaza los datos dos periodos atrás.

$$B(By_t) = B^2(y_t) = y_{t-2} \tag{2.24}$$

Una única diferenciación lineal podría ser descrita como la diferencia entre el periodo la observación en el periodo actual t y la observación el periodo anterior t. La observación en el periodo anterior a su vez puede ser expresada mediante el operador de retroceso tal como muestra la Ecuación 2.24.

$$y_t' = y_t - y_{t-1} = y_t - By_t = (1 - B)y_t \tag{2.25}$$

Mediante la Ecuación 2.25 se observa que una primera diferenciación lineal es representada por el término $(1 - B)$. De manera análoga, al realizar una diferenciación sobre la Ecuación 2.25 $(y_t')'$, se obtiene una diferenciación de segundo orden y_t'' descrita mediante la Ecuación 2.26.

$$y_t'' = y_t' - y_{t-1}' \tag{2.26}$$

Reemplazo de la Ecuación 2.26 en la Ecuación 2.25.

$$y_t'' = y_t - 2y_{t-1} + y_{t-2} = (1 - 2B + B^2)y_t \tag{2.27}$$

Al emplear el operador de retroceso en la Ecuación 2.27 se obtiene la Ecuación 2.28.

$$y_t'' = (1 - B)^2 y_t \tag{2.28}$$

En general, una diferenciación de orden d puede ser escrita mediante la Ecuación 2.29.

$$(1 - B)^d y_t \tag{2.29}$$

Al combinar la diferenciación con los modelos autorregresivos y de promedio móvil, se obtendrá modelos no estacionales ARIMA.

Los modelos ARIMA toman su nombre de las siglas en inglés para Autoregressive Integrated Moving Average, donde la palabra *integrated* representa la diferenciación.

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t \qquad (2.30)$$

La Ecuación 2.30 muestra el modelo ARIMA donde y'_t es la diferenciación de la serie que puede ser realizada más de una vez. Los predictores en el lado derecho incluyen valores rezagados de y_t y errores. Esto constituye un modelo ARIMA(p, d, q), donde

p = *orden de la parte autorregresiva*

d = *grado de la diferenciación*

q = *orden de la parte de promedio móvil*

Al combinar componentes, se obtiene modelos más complejos y será más fácil emplear la notación de retroceso, de esta forma la Ecuación 2.30 que presenta una única diferenciación es descrita mediante operadores de retroceso como muestra la Ecuación 2.31.

$$(1 - B)y_t = c + (1 + \phi_1 B + \cdots + \phi_p B^p)y_t + (1 + \theta_1 B + \cdots + \theta_q B^q)e_t \qquad (2.31)$$

De forma general el modelo ARIMA(p, d, q) es descrito mediante la Ecuación 2.32 para una diferenciación de orden d.

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - B)^d y_t = c + (1 + \theta_1 B + \cdots + \theta_q B^q)e_t \qquad (2.32)$$

Donde el término $(1 + \phi_1 B + \cdots + \phi_p B^p)y_t$ representa la parte autorregresiva, $(1 - B)^d y_t$ son las diferenciaciones realizadas al modelo y $(\theta_1 B + \cdots + \theta_q B^q)e_t$ incorpora la parte de media móvil al modelo ARIMA.

En el lenguaje R es utilizada una parametrización ligeramente diferente como se muestra la Ecuación 2.33.

$$\left(1 - \phi_1 B - \cdots - \phi_p B^p\right)(y'_t - \mu) = (1 + \theta_1 B + \cdots + \theta_q B^q)e_t \qquad (2.33)$$

Donde

$y'_t = (1 - B)^d y_t$

μ = media o promedio de y_t'

2.5 PRONÓSTICO

El pronóstico de un punto en el tiempo se puede calcular mediante los siguientes tres pasos:

1. Expandir la ecuación del modelo ARIMA de tal manera que y_t se encuentre en el lado izquierdo y todos los demás términos estén a la derecha.
2. Reescribir la ecuación reemplazando t por $T + h$, donde h es el número de periodos para el pronóstico.
3. En el lado derecho de la ecuación, se reemplaza los valores futuros de las observaciones con sus pronósticos, los errores futuros son reemplazados con cero, y sus errores pasados con sus correspondientes residuos.

Se debe iniciar con $h = 1$, estos pasos se deben repetir para cada valor de h hasta que todos los puntos del pronóstico hayan sido calculados.

Se plantea como ejemplo el modelo ARIMA$(3,1,1)$,m descrito mediante la Ecuación 2.34.

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3)(1 - B)^1 y_t = c + (1 + \theta_1 B)e_t \qquad (2.34)$$

Al desarrollar la Ecuación 2.34 se obtiene la Ecuación 2.35.

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - B + \phi_1 B^2 + \phi_2 B^3 + \phi_3 B^4)y_t = c + e_t + \theta_1 B e_t \qquad (2.35)$$

Se factora los operadores de retroceso de la Ecuación 2.35 y opera sobre y_t para obtener la Ecuación 2.36.

$$y_t - (1 + \phi_1)y_{t-1} + (\phi_1 - \phi_2)y_{t-2} + (\phi_2 - \phi_3)y_{t-3} + \phi_3 y_{t-4} = c + e_t + \theta_1 e_{t-1} \qquad (2.36)$$

Se despeja el valor de y_t como se solicitó en el paso 1, de esta manera se obtiene la Ecuación 2.37.

$$y_t = c + e_t + \theta_1 e_{t-1} + (1 + \phi_1)y_{t-1} - (\phi_1 - \phi_2)y_{t-2} - (\phi_2 - \phi_3)y_{t-3} - \phi_3 y_{t-4} \qquad (2.37)$$

Se reemplaza el valor de t por $T + 1$ en la Ecuación 2.37 y de esta manera se obtiene la Ecuación 2.38.

$$y_{T+1} = c + e_{T+1} + \theta_1 e_T + (1 + \phi_1)y_T - (\phi_1 - \phi_2)y_{T-1} - (\phi_2 - \phi_3)y_{T-2} - \phi_3 y_{T-3} \qquad (2.38)$$

Se debe asumir que la serie de datos a emplearse contiene datos hasta y_T, lo que ocasionaría que y_{T+1} será el siguiente valor de la serie de datos, es decir el valor pronosticado.

El paso 3 indica que los valores de error futuro deben ser remplazados por el valor de cero, en este caso e_{T+1} es un valor de error o residuo futuro como muestra la Ecuación 2.39.

$$y_{T+1} = c + \theta_1 e_T + (1+\phi_1)y_T - (\phi_1 - \phi_2)y_{T-1} - (\phi_2 - \phi_3)y_{T-2} - \phi_3 y_{T-3} \quad (2.39)$$

2.5.2 MÉTODOS DE ESTIMACIÓN DE ERRORES

Error Porcentual Absoluto Medio

Para evaluar el desempeño del pronóstico obtenido se emplea el error porcentual absoluto medio (MAPE) que mide el tamaño del error absoluto y los presenta en porcentaje como muestra la Ecuación 2.40. Este indicador es comúnmente utilizado para medir el grado de ajuste y generalización de los pronósticos realizados por los modelos seleccionados.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| * 100 \quad (2.40)$$

Donde

Y_i = *demanda de energía real en el periodo* i

$\hat{Y}_i$ = *demanda de energía prevista en el periodo* i

N = *número total de datos a considerar*

El método MAPE es el más adecuado para estimar el error relativo debido a que los errores pueden tener diferentes escalas [33].

Error absoluto medio

El cálculo del error absoluto medio (MAE) involucra el valor absoluto de todos los errores, obteniendo un error total que es dividido para el número de observaciones involucradas como se muestra en la Ecuación 2.41.

El MAE es una puntuación lineal lo que significa que todas las diferencias entre el valor real y estimado tienen la misma valoración [34].

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Y_i - \hat{Y}_i\right| \quad (2.41)$$

Raíz del error Cuadrático medio

La raíz del error cuadrático medio (RMSE) es una medida de desempeño frecuentemente utilizada para evaluar métodos de pronóstico de demanda. Difiere con la Desviación Media

Absoluta debido a que el RMSE amplifica y penaliza con mayor fuerza aquellos errores de mayor magnitud.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)}{N}} \tag{2.42}$$

La Ecuación 2.42 describe matemáticamente a la raíz del error cuadrático medio.

2.5.3 MODELOS ESTACIONALES SARIMA

Los modelos SARIMA son capaces de modelar una amplia gama de modelos estacionales. Un modelo SARIMA en su forma $(p, d, q)(P, D, Q)_m$ incluye datos estacionales $(P, D, Q)_m$ y a los modelos ARIMA(p, d, q), donde m es el número de observaciones por año.

La parte estacional del modelo consiste en términos muy similares a los componentes no estacionales, pero incorpora rezagos en periodos estacionales. Si se analiza un modelo SARIMA$(1,1,1)(1,1,1)_4$, emplea datos trimestrales ($m = 4$), y puede ser escrito mediante la Ecuación 2.43.

$$(1 - \Phi_1 B^4)(1 - B^4)(1 - \phi_1 B)(1 - B)y_t = c + (1 + \theta_1 B)(1 + \Theta_1 B^4)e_t \tag{2.43}$$

De manera general se puede escribir el modelo SARIMA mediante la Ecuación 2.44, donde se integra la parte no estacional (p, d, q) y la estacional $(P, D, Q)_m$.

Generalmente la Ecuación 2.44 es la empleada por los programas de estadística para representar a los modelos.

$$(1 - \Phi_1 B^m - \cdots - \Phi_P B^m)(1 - B^m)^D(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - B)^d y_t = c + (\theta_1 B + \cdots + \theta_q B^q)(1 + \Theta_1 B^m + \cdots + \Theta_Q B^m)e_t \tag{2.44}$$

La distinción de la parte SAR y SMA se encuentra en los rezagos estacionales de PACF y ACF. Por ejemplo, en un modelo SARIMA$(0,0,0)(0,0,1)_{12}$ tiene un pico en el rezago 12 en el gráfico de los coeficientes ACF y no existen otros picos significantes, mientras que en el gráfico de PACF se tiene un decaimiento exponencial en los rezagos estacionales 12, 24, 36…

De manera similar un modelo SARIMA$(0,0,0)(1,0,0)_{12}$ muestra un decaimiento exponencial en los rezagos estaciones en el gráfico de ACF, mientras que en el gráfico PACF muestra un único y significante pico en el rezago 12.

Residuos

Los residuos en los modelos de una serie de tiempo son los sobrantes luego del ajuste de un modelo como muestra la Ecuación 2.45.

$$e_t = y_t - \hat{y}_t \tag{2.45}$$

Los residuos son iguales a la diferencia entre las observaciones y su correspondiente valor ajustado.

Los residuos se emplean para la comprobar que el modelo construido es el adecuado y captura la información de los datos de entrenamiento. Un buen método de pronóstico obtiene residuos con las siguientes propiedades.

1. Los residuos no están correlacionados. Si existe correlaciones entre residuos, significa que la información que debió ser empleada en la construcción del modelo fue dejada de lado. El grafico de ACF deberá ser similar al ACF obtenido de una serie de ruido blanco.
2. Los residuos tienen una media de cero. Si los residuos tienen una media diferente de cero, el pronóstico es parcial.

Cualquier método de pronóstico que cumpla con estas propiedades puede ser mejorado. El comprobar estas propiedades es importante para ver si un método está usando toda la información disponible.

Se puede adicionar dos propiedades más, pero no son necesarias.

3. Los residuos tienen una varianza constante.
4. Los residuos están normalmente distribuidos.

Durbin Watson

La prueba desarrollada por Durbin y Watson es frecuentemente empleada para detectar la presencia de autocorrelación en los modelos de regresión. Se formulan dos hipótesis, la primera menciona $Ho: \rho = 0$ es decir no existe autocorrelación, la segunda $H1: 0 < |\rho| < 1$ existe autocorrelación [38].

El objetivo de la prueba de Durbin Watson es comprobar si la diferencia entre los residuos observados consecutivamente es muy grande o pequeña.

Una diferencia muy grande tendrá una autocorrelación negativa y una diferencia pequeña obtiene una autocorrelación positiva.

Esta prueba puede generar inconvenientes debido a que el resultado puede ser no concluyente y se debe emplear otros criterios [39].

2.6 ESTIMACIÓN Y ORDEN DE SELECCIÓN

2.6.1 ESTIMACIÓN DE MÁXIMA VEROSIMILITUD

Una vez que el orden modelo ha sido identificado, el siguiente paso es estimar los parámetros $c, \theta_1, \cdots, \theta_p, \phi_1, \cdots, \phi_q$.

Al estimar estos parámetros de los modelos SARIMA en RStudio se emplea el método de máxima verosimilitud (MLE). Esta técnica se encarga de encontrar los valores de los parámetros que maximicen la probabilidad de obtener los datos de entrenamiento. Para los modelos ARIMA, el MLE es muy similar a las estimaciones de mínimos cuadrados, que se obtienen mediante la minimización de la Ecuación 2.12.

En los modelos ARIMA el grado de dificultad para la estimación de los coeficientes es aún mayor que en la estimación de los coeficientes de los modelos regresivos. Para los valores de p, q y d RStudio trata de maximizar el logaritmo de la verosimilitud cuando se encuentra los parámetros estimados.

Se emplea la notación L para describir la verosimilitud, teniendo en cuenta que entre mayor sea el valor, será mejor el ajuste del modelo. L mide la bondad de ajuste del modelo.

2.6.2 CRITERIO DE INFORMACIÓN

El criterio de información de Akaike (AIC) descrito por la Ecuación 2.46, permite determinar el orden del modelo ARIMA.

$$AIC = -2\log(L) + 2(p + q + k + 1) \tag{2.46}$$

Donde

$k =$ *número total de rezagos de parámetros y los estados iniciales que han sido estimados*

El criterio AIC es la diferencia entre que también se ajusta el modelo $2\log(L)$ y la penalización por el número de parámetros $2(p + q + k + 1)$.

El criterio de información Bayesiana fue propuesto por Schwarz, este criterio se basa en la teoría de la información, pero en un contexto Bayesiano. La diferencia entre el criterio BIC y AIC es el grado de penalidad impuesto por el número de parámetros, siendo en el criterio BIC más fuerte que en el criterio AIC. El criterio BIC se puede calcular de con la Ecuación 2.47.

$$BIC = -2\log(L) + (p + q + k + 1)log(T) \tag{2.47}$$

El mejor modelo será aquel que obtenga menor BIC. Se sugiere emplear el criterio de BIC para series de datos de tamaño grande es decir mayor a 200 observaciones, y el criterio de AIC será empleado para series de datos pequeñas. A menudo los dos criterios producen el mismo resultado [18].

En RStudio el cálculo del valor de BIC se lo hace mediante la Ecuación 2.48.

$$BIC = AIC + k[log(T) - 2] \tag{2.48}$$

3. PRONÓSTICOS A MEDIANO PLAZO

Este capítulo presenta la metodología a seguir, para la minería de datos, construcción y validación del modelo para energía y potencia activa mensual en base a los datos proporcionados por Empresa Eléctrica Quito (EEQ). El proceso de minería de datos ha sido detallado en la sección 2.2.2.

3.1 CONOCIMIENTO PREVIO

El primer paso del proceso de minería de datos como se muestra en la Figura 2.4 es el conocimiento previo del negocio, a continuación, se relata brevemente la información obtenida en este paso.

El negocio de la Empresa Eléctrica Quito es el de proveer el servicio público de generación, subtransmisión y distribución de energía eléctrica a sus áreas de concesión. Se encuentra constituida en su mayor parte por capital público.

La EEQ dentro de sus políticas institucionales fomenta la cultura de planificación. La planificación técnica que desarrolla la EEQ es realizada a través de planes de expansión como una herramienta para su desarrollo donde se detalla un conjunto de programas y proyectos de inversión que apoyan a la expansión, mejora o reemplazo de sus equipos, instalaciones eléctricas y sistemas de gestión.

Los requerimientos actuales y futuros de su mercado son la base fundamental para realizar los planes de expansión. Los requerimientos a futuro están acorde al crecimiento de la demanda del sistema. Además, la Ley Orgánica del Servicio Público de Energía Eléctrica, Reglamentos, Regulaciones y Resoluciones emitidas por la Agencia de Regulación y Control de Electricidad y otras normas legales exhorta a las empresas de distribución tal como la Empresa Eléctrica Quito a desarrollar el plan de expansión en un horizonte de 10 años y debe ser actualizado anualmente [40].

Para proseguir al siguiente paso de la minería de datos se debe contestar las siguientes preguntas detalladas en la sección 2.2.2.

- ¿Cuál es el problema que se solucionará?

Se pretende obtener una adecuada planificación del sistema eléctrico de distribución en donde una mala estimación puede generar inversiones en proyectos que excedan la capacidad requerida o una subestimación que puede conllevar a un incumplimiento de la demanda, afectando el servicio de varios usuarios y su productividad.

- ¿Cómo encaja en el contexto comercial?

El conocimiento de la demanda eléctrica futura del sistema permitirá realizar inversiones adecuadas y evitar sanciones por incumplimiento del servicio.

- ¿Qué datos se necesitan para resolver este problema?

Se requieren datos históricos anuales y mensuales de potencia activa, reactiva, energía y factor de potencia.

3.2 ENERGÍA MENSUAL

Los registros de energía mensual proporcionados por la EEQ son medidos en MWh desde enero de 1983 hasta diciembre del año 2018 como muestra la Figura A.1. del Anexo A. (Base de datos proporcionada por la EEQ y analizada). Luego de realizar la minería de datos e identificar irregularidades se conformó el conjunto de datos mostrados en la Figura A.2. del Anexo A con los cuales se realizó el modelamiento. Los datos empleados para entrenamiento se comprenden desde enero de 1983 hasta diciembre de 2016, mientras que los datos desde enero de 2017 hasta mayo de 2018 fueron empleados como datos de validación.

En cumplimiento al alcance presentado para este proyecto de titulación se realizó el pronóstico desde mayo de 2018 hasta diciembre del mismo año.

Con el pronóstico obtenido se realizó la validación del modelo a través de la comparación de su MAPE con la metodología actual de la Empresa Eléctrica Quito.

3.2.1 PREPARACIÓN DE LOS DATOS DE ENERGÍA

La información proporcionada por la EEQ inicialmente se encuentra en Excel ordenada en columnas y sus atributos se encuentran en filas, se realizó una inspección visual al graficar la serie de tiempo que presentan cada año, para entender su evolución y comportamiento.

Se importó y declaró la serie de tiempo para su posterior modelamiento tomando en cuenta que RStudio trabaja con vectores. Los atributos de los datos proporcionados para el pronóstico mensual de energía son la fecha que consta del mes y año en la cual se obtuvo la medición y el segundo atributo es la medición energía en MWh medida en el mes.

Ecuador es un país que ha sufrido varios acontecimientos que han marcado su historia [41] derrocamiento de presidentes, erupciones, fenómeno del niño, dolarización, sismos, etc. Estos eventos se convierten en componentes irregulares en la demanda eléctrica y pueden llegar afectar al pronóstico como se mencionó en la sección 2.3.

Luego de realizar un análisis detallado de la información proporcionada por la EEQ existieron comportamientos anómalos provocados por eventos aislados, estos datos aportan errores al modelo debido a que los modelos ARIMA y SARIMA dependen únicamente de los datos históricos de una serie.

La Figura A.2. muestra la serie de datos donde los eventos aislados ya han sido removidos, esta nueva serie será la utilizada para la construcción del modelo ARIMA o SARIMA de la energía mensual para la Empresa Eléctrica Quito.

3.2.2 MODELAMIENTO DE DATOS MENSUALES DE ENERGÍA

A continuación, se realiza el proceso de modelamiento para los datos históricos de energía mensual entregados por la EEQ.

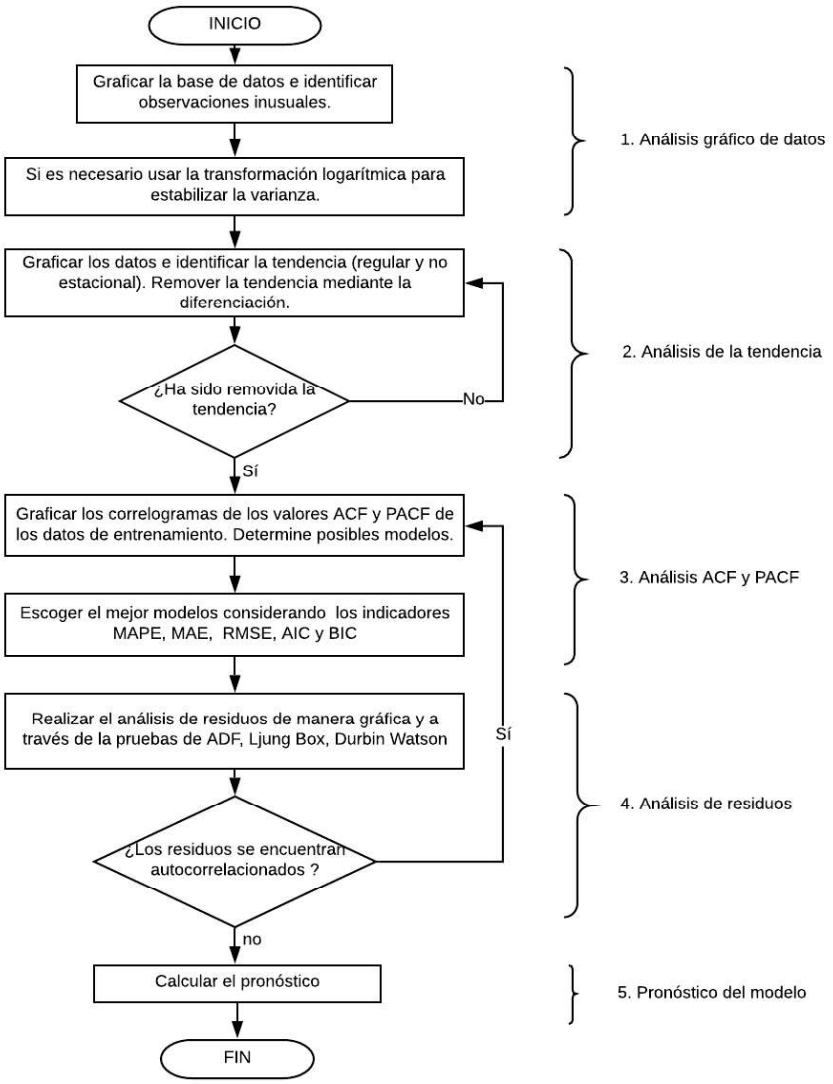

Figura 3.1. Diagrama de flujo del proceso de modelamiento [42].

La Figura 2.6 de la sección 2.2.2 muestra de forma general el proceso de modelamiento, en base a esto la Figura 3.1 detalla la metodología que se emplea para la selección del modelo más adecuado.

Análisis gráfico de datos

La serie de datos históricos de energía está comprendida desde enero del año 1989 hasta diciembre del año 2016, la Figura 3.2 muestra esta serie de tiempo.

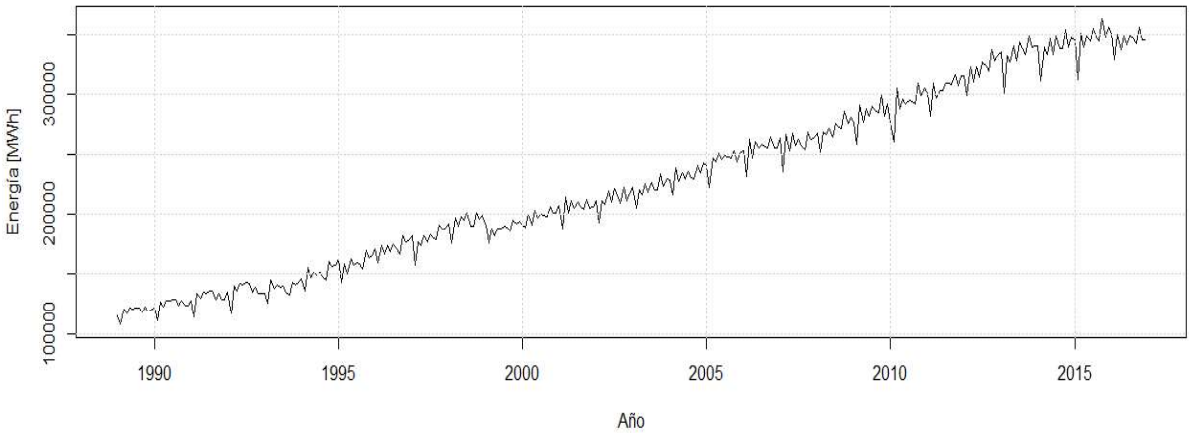

Figura 3.2. Datos de entrenamiento de energía mensual.

Los datos de entrenamiento pueden ser descompuestos en los componentes detallados en la sección 2.3 a través de la función *Decompose* de RStudio, de esta manera se consiguió la Figura 3.3.

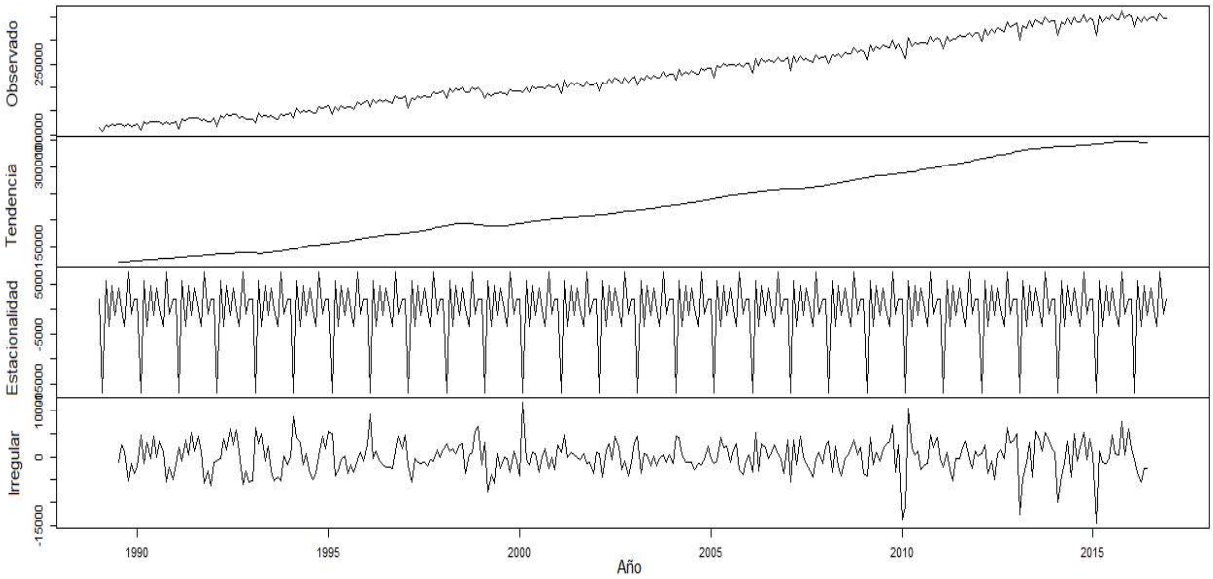

Figura 3.3. Datos de entrenamiento de energía mensual descompuesto en componentes aditivas.

La Figura 3.3 muestra la existencia de una tendencia significativa, por este motivo la serie es no estacionaria. Para comprobar lo expuesto por la Figura 3.3 se emplea la prueba de Dickey Fuller aumentada obteniendo como resultado un valor de $p = 0{,}2646$, no se puede rechazar la hipótesis nula, se confirma que se trata de una serie no estacionaria.

Análisis de la tendencia

Las series no estacionarias requieren de un proceso de diferenciación como se explicó en la sección 2.4.6. Se puede optar por una diferenciación regular, estacional o la combinación de las dos. Se seleccionará únicamente la que obtenga un valor menor de *p*.

Debido a la tendencia presentada en la Figura 3.3, se propone una diferenciación regular y una estacional. Al aplicar estas diferenciaciones a los datos de entrenamiento se obtiene la información organizada en la Tabla 3.1.

Tabla 3.1. Prueba de Dickey Fuller aumentada para datos de energía mensual.

Diferenciación	Modelo	R^2	Valor de *p*
Regular	$(p,1,q)(P,0,Q)_{12}$	0,4675	$< 2{,}2e^{-16}$
Estacional	$(p,0,q)(P,1,Q)_{12}$	0,4687	$< 2{,}2e^{-16}$

La Tabla 3.1 muestra los valores de *p* obtenidos en la prueba Dickey Fuller, los valores de *p* obtenidos por la diferencia regular y estacional son similares y muestra que la tendencia ha sido prácticamente removida en su totalidad. Es necesario emplear otro indicador para la selección, se plantea el valor del coeficiente de determinación R^2 que muestra la bondad de ajuste de un modelo a la variable que se pretende explicar, entre más cerca de 1 es mayor el ajuste, la diferenciación que obtenga un mayor valor de R^2 es la seleccionada, en base a este criterio se obtiene la diferenciación estacional.

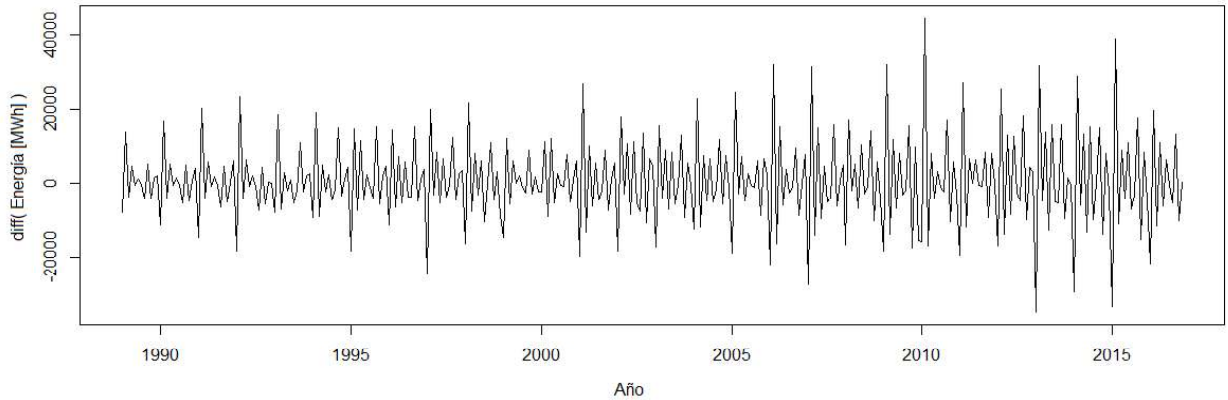

Figura 3.4. Diferenciación estacional de los datos de energía mensual.

La Figura 3.4 muestra la serie de energía mensual una vez que ha sido removida su tendencia, como se pueden observar los valores se mantienen alrededor de cero.

Análisis ACF y PACF

El siguiente paso consiste en analizar el orden del modelo que se encuentra en construcción, se grafica los valores de ACF y PACF de la serie ya diferenciada en la Figura 3.5.

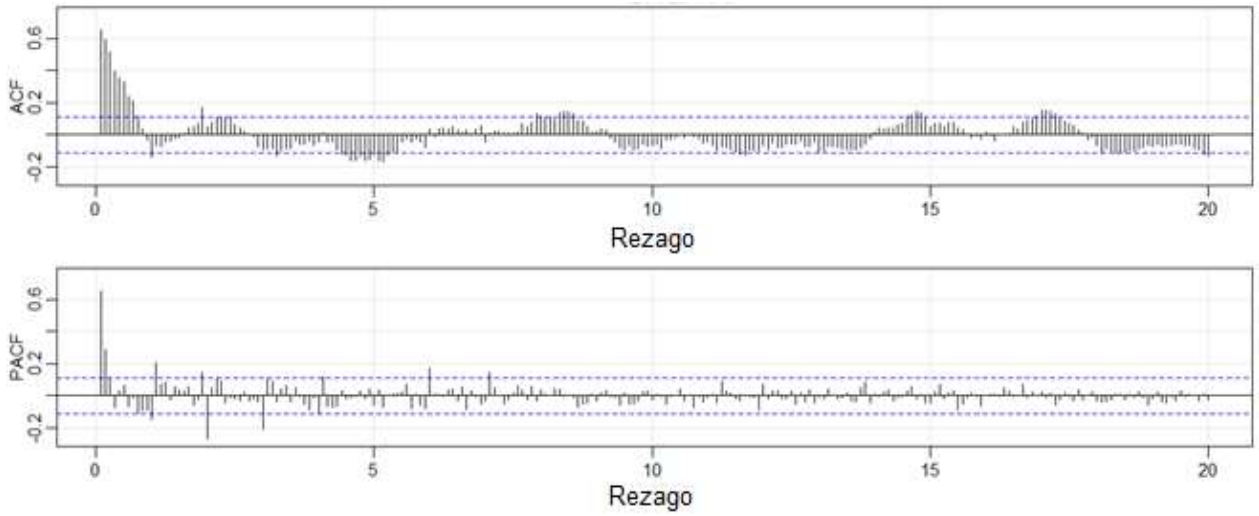

Figura 3.5. Valores de ACF y PACF para serie de energía mensual diferenciada.

La Figura 3.6 ha sido obtenida al delimitar el número de rezagos visibles de la Figura 3.5, esto permite una mejor visualización de los valores ACF y PACF.

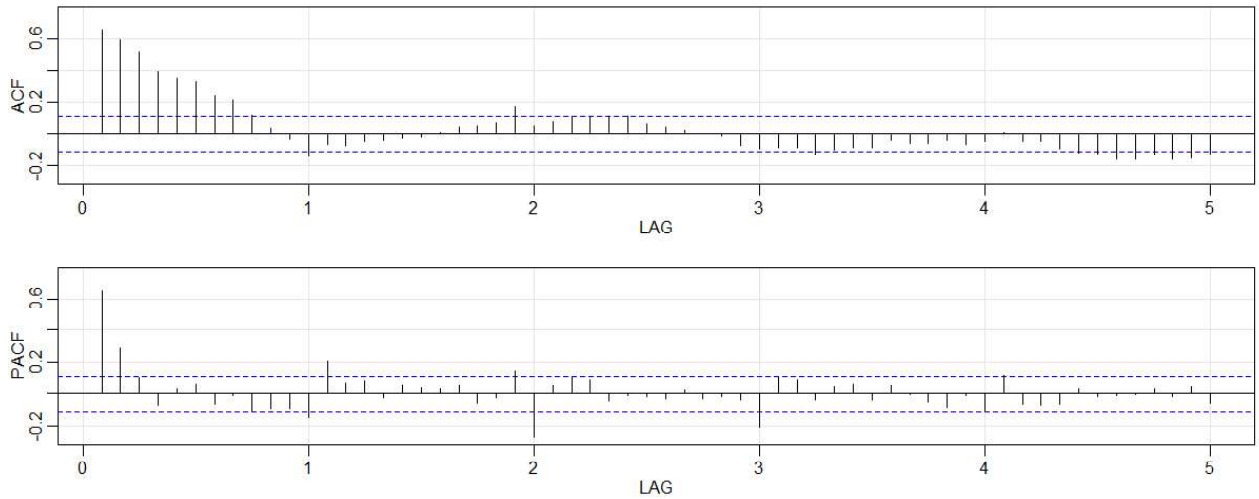

Figura 3.6. Valores de ACF y PACF para serie de energía mensual diferenciada acotada.

Los rezagos 1,2,3,4 . . . son rezagos estacionales y permiten construir el componente estacional (P, D, Q) del modelo SARIMA. De acuerdo con lo mostrado en la Figura 3.6,

estos rezagos van disminuyendo exponencialmente en los valores de PACF y los valores de ACF se vuelven insignificantes a partir del rezago 2.

Para la construcción de la parte no estacional (p, d, q) del modelo SARIMA se visualiza los rezagos que se encuentran en medio de los rezagos estacionales, debido a que los valores de ACF y PACF muestran una forma sinusoidal se tendrá valores de $p\ y\ q$, tal como se mencionó en la Tabla 2.2. En los valores de ACF la forma sinusoidal se empieza a formar a partir del rezago 1, mientras que en los valores de PACF no se puede asegurar un valor a simple vista por lo que se propone diferentes modelos.

Tabla 3.2. Resumen de los valores de ACF y PACF de la serie diferenciada.

	SAR (2)	ARMA(*p*,1)
ACF	Los valores son insignificantes a partir del rezago 2	Forma sinusoidal a partir del rezago 1
PACF	Decrece los valores	Forma sinusoidal a partir del rezago *p*

La Tabla 3.2 resume los resultados obtenidos de la observación de los correlogramas de autocorrelación y autocorrelación parcial con los cuales se proponen los siguientes modelos para su posterior evaluación.

Tabla 3.3. Modelos propuestos para datos mensuales de energía

Modelos	AIC	BIC	RMSE	MAE	MAPE (%)
SARIMA$(0,0,1)(0,1,2)_{12}$	6446,18	6465,09	4871,28	3709,55	1,6831
SARIMA$(1,0,1)(0,1,2)_{12}$	6265,34	6288,02	3637,69	2676,54	1,2145
SARIMA$(2,0,1)(0,1,2)_{12}$	6266,87	6293,34	3634,54	2675,16	1,2133
SARIMA$(3,0,1)(0,1,2)_{12}$	6267,35	6297,59	3625,11	2650,64	1,2031
SARIMA$(4,0,1)(0,1,2)_{12}$	6269,34	6303,37	3625,08	2650,19	1,2029

La Tabla 3.3 muestra varios indicadores de los modelos planteados, se distingue que los modelos presentan un mejor ajuste al incrementar el valor de *p*. Sin embargo, los modelos

SARIMA$(2,0,1)(0,1,2)_{12}$, SARIMA$(3,0,1)(0,1,2)_{12}$ y SARIMA$(4,0,1)(0,1,2)_{12}$ presentan valores similares en MAPE, para la selección del modelo se emplea el principio de parsimonia o razón de Ockam "*Es vano hacer más con lo que se puede hacer menos*", al aplicar este criterio a la predicción de series de tiempo, implica la preferencia de modelos que sean menos complicados siempre y cuando sus características sean similares [43]. Esta filosofía fue apoyada por Box y Jenkins donde se enfatiza la necesidad de emplear el principio de parsimonia en la selección de modelos ARMA [44].

El modelo seleccionado mediante el principio de Ockam es SARIMA$(2,0,1)(0,1,2)_{12}$, debido a que su valor MAPE no presenta gran diferencia con los modelos SARIMA$(3,0,1)(0,1,2)_{12}$, y SARIMA$(4,0,1)(0,1,2)_{12}$ que son modelos que incorporan mayor número de variables y complejidad.

Análisis de residuos

Una vez definido el modelo a emplearse, se debe realizar distintas pruebas para comprobar que éste capture la mayor cantidad de información relevante de los datos de entrenamiento.

La Figura 3.7 (a) muestra la media obtenida de los residuos del modelo SARIMA$(2,0,1)(0,1,2)_{12}$, manteniéndose alrededor de 0 de acuerdo con lo mencionado en la sección 2.5.3.

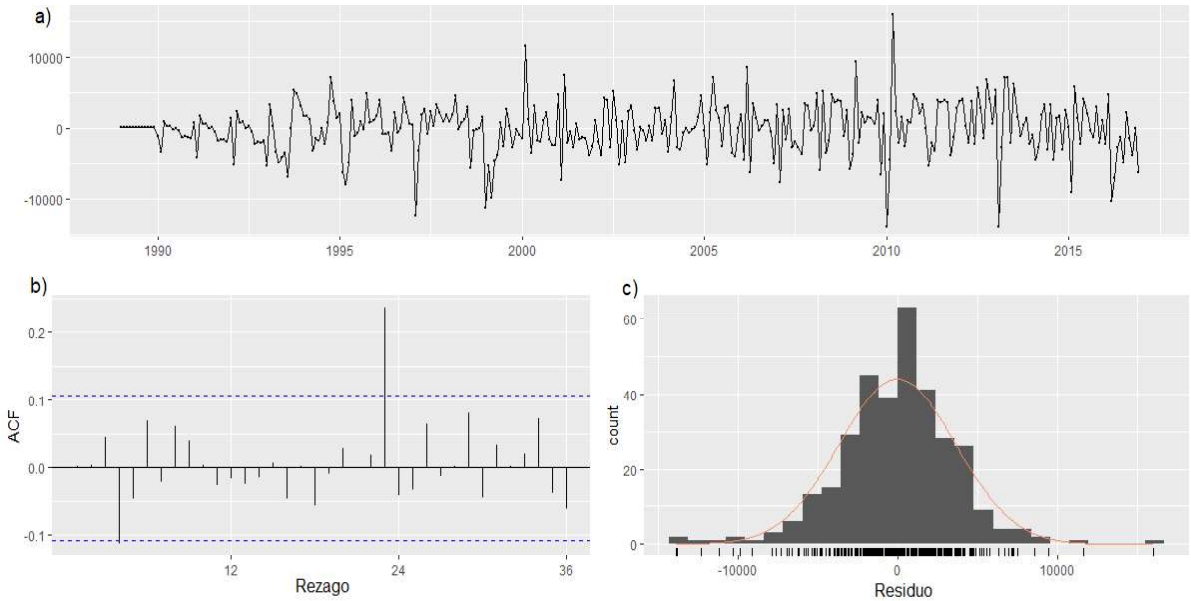

Figura 3.7. Análisis de residuos del modelo SARIMA$(2,0,1)(0,1,2)_{12}$ para energía mensual de la EEQ.

La Figura 3.7 (b) es el correlograma de los valores de autocorrelación obtenida de los residuos del modelo, los valores se acercan a los obtenidos por un proceso de ruido blanco. Finalmente, la Figura 3.7 (c) muestra que los residuos del modelo siguen una distribución normal. Cada una de las distintas graficas de la Figura 3.7 aporta a la conclusión de que el modelo SARIMA$(2,0,1)(0,1,2)_{12}$, ha capturado la mayor cantidad de información presentada en la serie de tiempo de demanda de energía mensual.

Sin embargo, se presentan dos pruebas más para desechar la hipótesis de correlación entre residuos del modelo. Los resultados obtenidos son escritos en la Tabla 3.4.

Tabla 3.4. Pruebas a residuos del modelo SARIMA$(2,0,1)(0,1,2)_{12}$ de energía mensual.

Prueba	**Valor de *p***
Ljung-Box	0,093465
Dickey Fuller aumentada	<2,2e-16
Durbin Watson **(Valor de *d*)**	1,9874

El modelo planteado en la prueba de Ljung-Box obtiene un valor de *p* mayor a 0.05 mostrando la independencia de los residuos [45].En la prueba de Dickey Fuller aumentada muestra un valor mínimo de *p*. Finalmente la prueba de Durbin Watson obtiene un valor $d = 1{,}9874$ muy cercano dos, lo que permite concluir que la correlación entre residuos es mínima.

3.2.3 PRONÓSTICO

Luego de haber construido el modelo mediante la observación y aplicación de pruebas se realiza el pronóstico del modelo SARIMA$(2,0,1)(0,1,2)_{12}$, obtenido con los datos mensuales de energía en MWh. Los resultados se muestran en la Tabla 3.5.

Tabla 3.5. Pronóstico de energía mensual, mediante el modelo SARIMA$(2,0,1)(0,1,2)_{12}$.

Fecha	**Energía pronosticada en R [MWh]**	
	2017	**2018**
Enero	347.541,40	355.998,40
Febrero	324.302,30	331.997,70

Marzo	350.644,30	360.389,70
Abril	340.788,20	350.740,70
Mayo	351.917,00	361.750,60
Junio	345.204,40	355.011,20
Julio	354.285,60	364.716,50
Agosto	351.658,30	361.270,80
Septiembre	348.397,20	358.385,00
Octubre	362.687,60	373.071,60
Noviembre	352.208,70	362.214,30
Diciembre	355.978,70	366.937,50
TOTAL	4’185.613,70	4’302.484,00
Tasa anual (%)	0,938	2,792

La Figura 3.8 reúne los datos históricos de energía mensual que sirvieron como datos de entrenamiento (negro), los valores pronosticados desde enero de 2017 a diciembre de 2018 se encuentran en color azul y el nivel de confianza para el intervalo de predicción se dispuso al 95%. Un nivel de confianza de 95 % indica que, si se tomaron 100 muestras aleatorias de la medición, los intervalos de confianza para aproximadamente 95 de las muestras incluirían la respuesta media. De manera similar, el intervalo de predicción indica que se puede estar un 95% seguro de que el intervalo contiene el valor de una nueva observación individual [45].

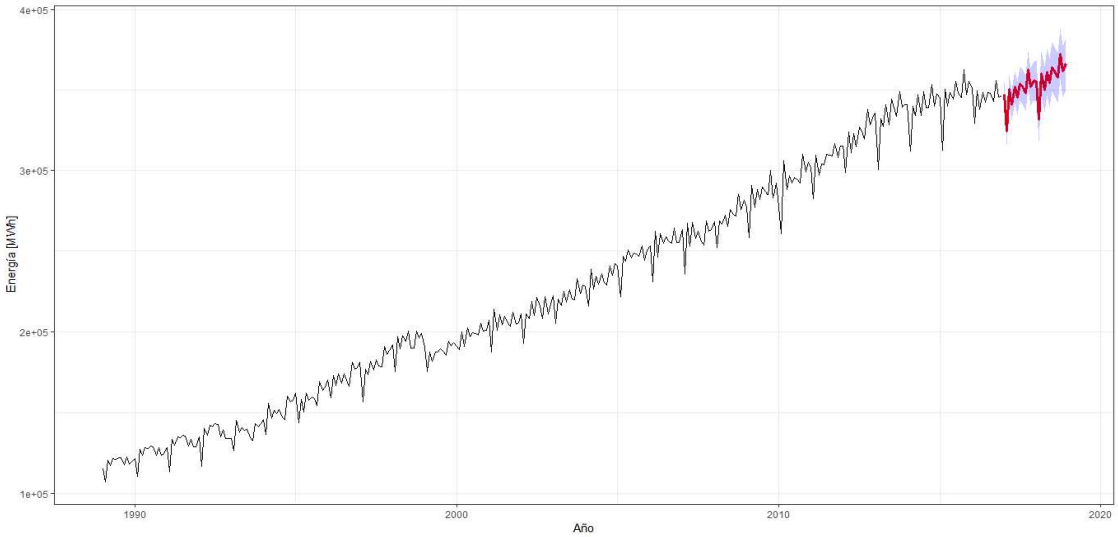

Figura 3.8. Pronóstico de energía mensual.

3.2.4 ANÁLISIS DE RESULTADOS DEL PRONÓSTICO MENSUAL DE ENERGÍA

El pronóstico a mediano plazo comprende en predecir dos años, la Empresa Eléctrica Quito proporcionó datos hasta diciembre del 2018. Los datos hasta el 2016 son utilizados como datos de aprendizaje mientras que los datos de enero del 2017 a mayo 2018 son empleados para la validación.

Los datos de aprendizaje están conformados desde enero de 1989 a diciembre de 2016. En la sección 3.2.2 se realizó la selección del modelo mensual de energía en base a la experticia y observación de pronosticador; RStudio a través de su paquete *forecast* permite emplear la función *auto.arima*, donde uno de sus parámetros permite configurar el criterio de selección entre AIC y BIC.

La Tabla 3.6 presenta los valores de MAPE con relación a los valores históricos y los ajustados del modelo, es decir mide el ajuste entre los valores obtenidos por el modelo y los reales. También muestra la relación entre los valores de pronosticados por el modelo y los datos de evaluación. Los resultados obtenidos fueron mediante el criterio de selección Akaike para este modelo debido a que presenta un menor valor de error MAPE, indicando que el modelo SARIMA$(2,0,1)(0,1,2)_{12}$ se ajusta de mejor manera a los datos mensuales de energía y presenta un mejor pronóstico.

Tabla 3.6. Criterio de selección para serie mensual de energía

Datos	MAPE [%]

	AIC	BIC
	$(2,0,1)(0,1,2)_{12}$	$(1,0,1)(0,1,2)_{12}$
Entrenamiento (históricos)	1,213	1,214
Evaluación (reales)	1,454	1,494

El modelo seleccionado para energía mensual mediante la función *auto.arima* fue el modelo SARIMA$(2,0,1)(0,1,2)_{12}$ y el criterio propio del pronosticador obtiene el mismo orden, sin embargo, existen pequeñas diferencias por los coeficientes empleados.

La Figura 3.9 muestra la relación entre los datos de entrenamiento en color azul y los valores entregados por el modelo en color rojo. El modelo SARIMA$(2,0,1)(0,1,2)_{12}$.captura la tendencia y estacionalidad a partir de los datos históricos mensuales de energía medidos en MWh.

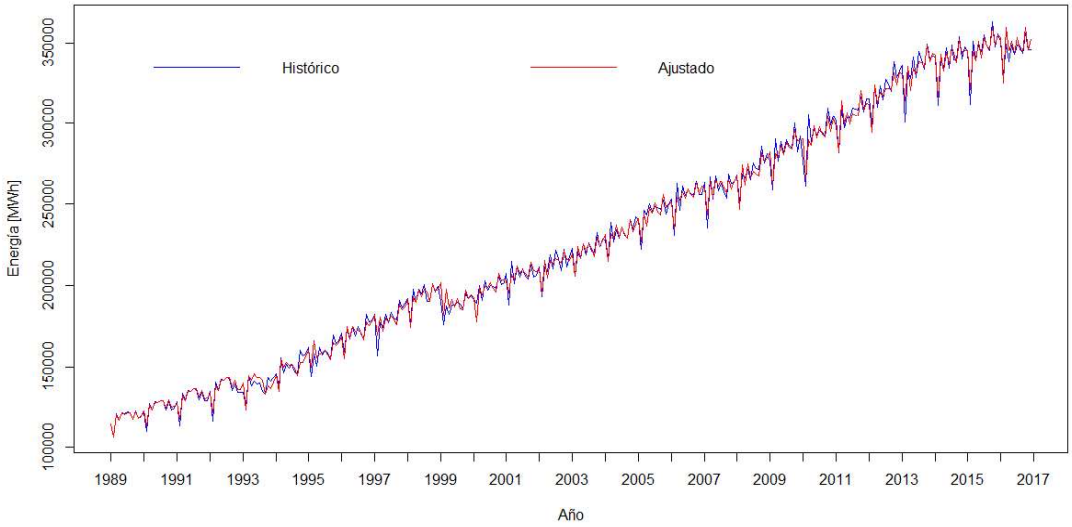

Figura 3.9. Comparativa entre datos de entrenamiento y valores ajustados del modelo SARIMA$(2,0,1)(0,1,2)_{12}$ para energía mensual.

Se realizó el cálculo del error porcentual absoluto medio y se presenta en la Tabla 3.7. El modelo SARIMA$(2,0,1)(0,1,2)_{12}$ elaborado a partir de los datos históricos de energía obtiene un MAPE de valor 1,140% para el año 2017, convirtiéndolo en un modelo adecuado para el pronóstico de energía mensual de la EEQ lo que permite que el modelo sea aceptado. El valor total real obtenido es de 4'228.000,30 MWh mientras que el valor total obtenido por el pronóstico del modelo desarrollado en lenguaje R es de 4'185.613,70,

manteniéndose mucho más cercano que el pronóstico obtenido por la metodología actual de la EEQ.

Tabla 3.7. Pronóstico al año 2017 de energía mensual

Fecha	Energía real [MWh]	Energía pronosticada en R [MWh]	Energía pronosticada por la EEQ [MWh]
Enero	356.814,70	347.541,40	357.584,30
Febrero	321.668,90	324.302,30	329.109,70
Marzo	366.928,80	350.644,30	360.234,10
Abril	344.504,70	340.788,20	350.006,50
Mayo	360.874,30	351.917,00	372.199,40
Junio	350.488,30	345.204,40	353.463,70
Julio	354.436,10	354.285,60	364.824,10
Agosto	352.652,60	351.658,30	361.151,20
Septiembre	348.406,20	348.397,20	358.299,90
Octubre	362.902,80	362.687,60	372.200,10
Noviembre	351.612,00	352.208,70	359.446,10
Diciembre	356.710,90	355.978,70	354.088,70
Total	4’228.000,30	4’185.613,70	4’292.607,80
MAPE (%)		1,140%	1,970%

La Tabla 3.8 muestra los resultados del pronóstico del año 2018 mediante el modelo SARIMA$(2,0,1)(0,1,2)_{12}$ propuesto y la metodología empleada por la EEQ. La sumatoria de los cinco primeros meses del año 2018 muestra que el modelo desarrollado en el lenguaje R se mantiene cercano al real y el MAPE obtenido por el modelo propuesto es inferior al obtenido mediante el pronóstico presentado por la EEQ. Cabe destacar que el valor real total del año 2018 fue de 4’230.002,446 y el obtenido en este trabajo de titulación es de

4’302.484,00 mostrado en la Tabla 3.5, señalando que el modelo SARIMA$(2,0,1)(0,1,2)_{12}$ ha obtenido y empleado de manera correcta la información proporcionada a través de la serie de tiempo y se coloca como una metodología viable para el pronóstico de energía mensual para la EEQ.

Tabla 3.8. Pronóstico al año 2018 de energía mensual y MAPE

Mes	Energía real [MWh]	Energía pronosticada en R [MWh]	Energía pronosticada por la EEQ [MWh]
Enero	364.408,40	355.998,40	369.430,60
Febrero	328.014,50	331.997,70	351.959,60
Marzo	370.215,20	360.389,70	371.093,70
Abril	357.420,40	350.740,70	361.071,10
Mayo	372.923,60	361.750,60	385.541,20
Total	1’792.982,10	1’760.877,10	1’839.096,20
MAPE (%)		2,208%	2,664%

Luego de la evaluación del modelo construido en lenguaje R y el pronóstico obtenido por la metodología actual de la EEQ, se realiza el pronóstico de los siguientes siete meses mostrados en la Tabla 3.9.

Tabla 3.9. Pronóstico del modelo mensual de energía Junio-Diciembre 2018

Fecha	Energía pronosticada en R [MWh]	Energía real [MWh]
Junio	355.011,235	350.448,356
Julio	364.716,501	354.436,085
Agosto	361.270,77	352.652,611
Septiembre	358.384,966	348.406,170
Octubre	373.071,582	362.902,778

Noviembre	362.214,294	351.616,352
Diciembre	366.937,452	358.708,698

3.3 POTENCIA ACTIVA MENSUAL

3.3.1 PREPARACIÓN DE LOS DATOS DE POTENCIA ACTIVA

Los atributos proporcionados para el pronóstico mensual de potencia activa es la fecha que consta del mes y año en la cual se obtuvo la medición y el segundo atributo es la medición del valor máximo de potencia activa en MW medido en el mes.

3.3.2 MODELAMIENTO DE DATOS MENSUALES DE POTENCIA ACTIVA

La serie de potencia activa suministrada por el Empresa Eléctrica Quito en la Figura A.3. del Anexo A, presentan grandes variaciones y no existe un patrón tan definido como el encontrado en la serie de energía mensual, debido a que los valores de potencia activa son los máximos obtenidos en un mes y su obtención es susceptible a errores humanos. La Figura A.4. del Anexo A muestra una serie mensual de potencia activa en la cual se ha desechado ciertos componentes irregulares.

Análisis grafico de datos

La serie de datos para entrenamiento para el modelo potencia activa está comprendida desde enero del año 1996 hasta el diciembre del año 2016. La Figura 3.10 muestra la serie de tiempo formada con estos datos.

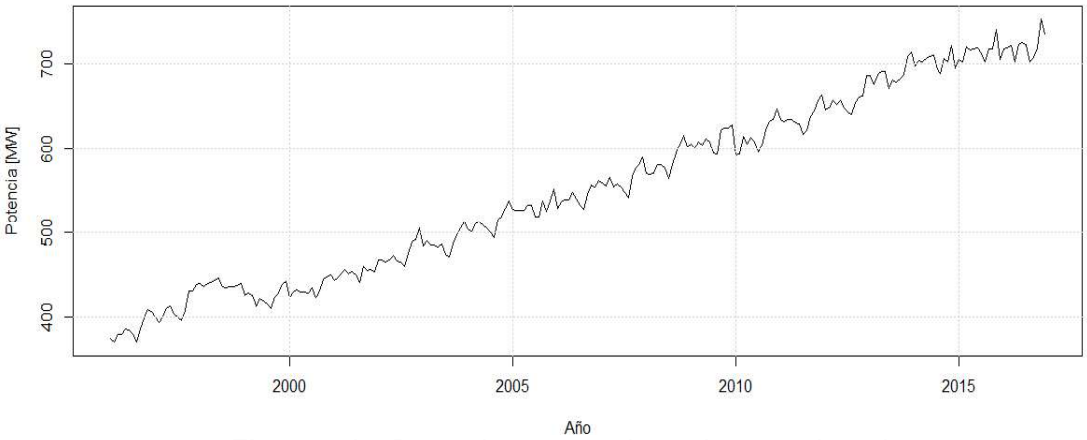

Figura 3.10. Datos de entrenamiento de potencia activa.

La Figura 3.11 muestra la descomposición de la serie mensual de potencia activa en sus componentes tendencia, estacionalidad y componente irregular. El componente de tendencia es grande por lo que se ve la necesidad de diferenciar la serie para remover la tendencia existente.

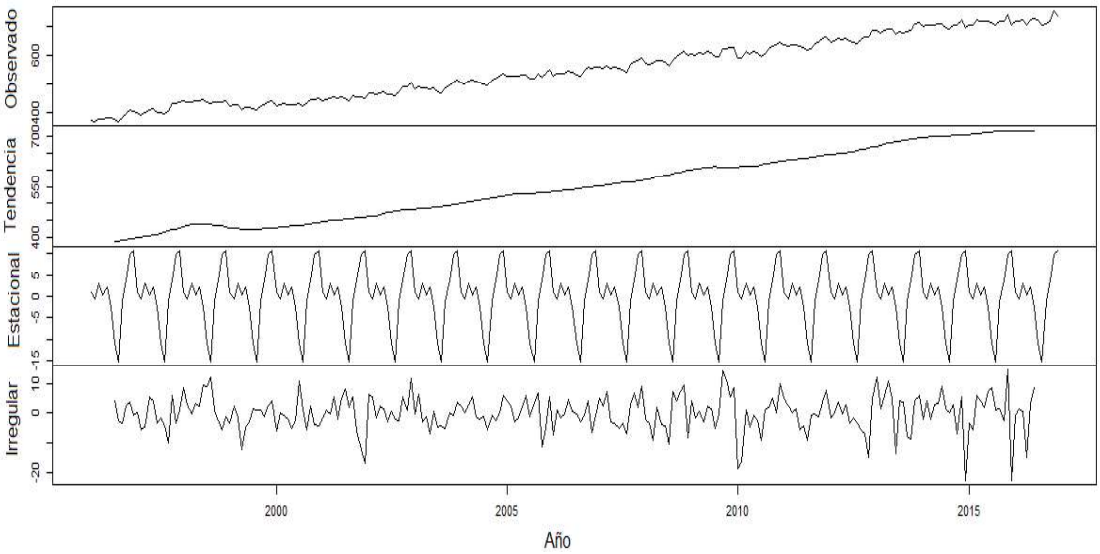

Figura 3.11. Datos de entrenamiento de potencia activa mensual descompuestos en componentes aditivas.

Análisis de la tendencia

A pesar de la visualización de la existencia de tendencia en la serie mensual de potencia activa se realiza la prueba de Dickey Fuller aumentada, los resultados sugieren que se trata de una serie no estacionaria con un valor de $p = 0{,}0425$. Por este motivo se requiere diferenciar la serie mensual de potencia.

Tabla 3.10. Prueba de Dickey Fuller aumentada para datos de potencia activa mensual.

Diferenciación	Modelo	R^2	Valor de *p*
Regular	$(p,1,q)(P,0,Q)_{12}$	0,1667	$4{,}489e^{-10}$
Estacional	$(p,0,q)(P,1,Q)_{12}$	0,5756	$< 2{,}2e^{-16}$

La Tabla 3.10 muestra los valores obtenidos al realizar una diferenciación regular y una estacional, al comparar estos valores la diferencia regular obtiene un menor valor de *p* y un mejor ajuste, esto permite concluir que la mejor opción es la diferenciación regular.

La Figura 3.12 muestra la serie luego de haber sido diferenciada regularmente ($d = 1$). La serie se presenta sin tendencia, es decir se constituye como una estacionaria.

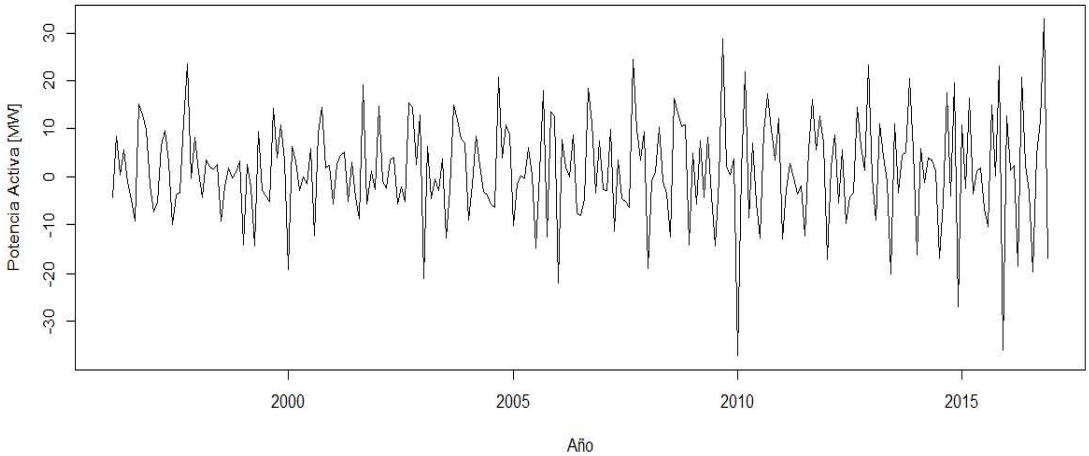

Figura 3.12. Diferenciación regular de potencia activa mensual.

Análisis ACF y PACF

Para el reconocimiento del orden del modelo ARIMA o SARIMA, la serie debe ser estacionaria y se grafica sus coeficientes de autocorrelación y autocorrelación parcial tal como se encuentran en la Figura 3.13.

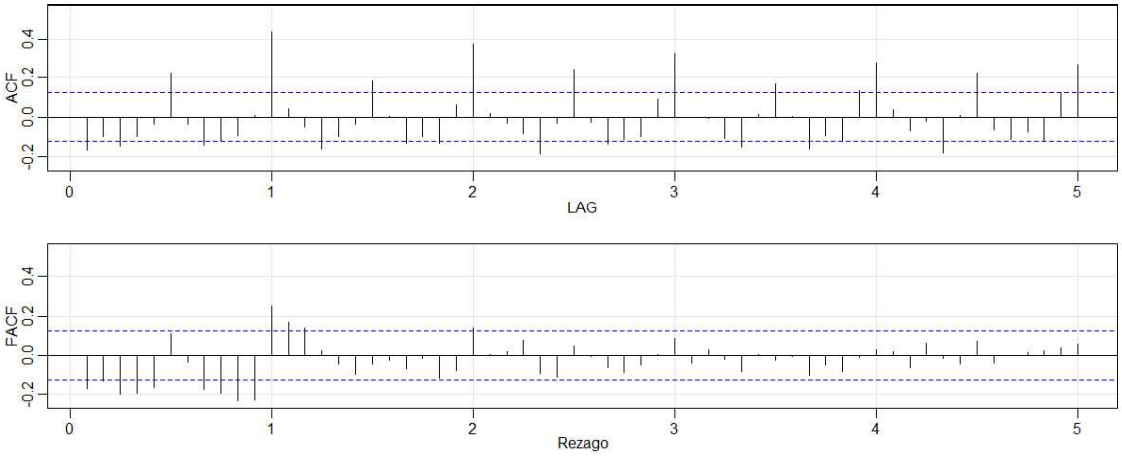

Figura 3.13. Correlograma de ACF y PACF de potencia activa mensual diferenciada.

Los rezagos estacionales en el correlograma de autocorrelación van decreciendo de forma exponencial mientras que en el correlograma de valores de PACF a partir del tercer rezago sus valores son insignificantes, estos detalles lo convierten en un modelo SARMA$(3,0,0)_{12}$.

Los rezagos no estacionales muestran un decrecimiento y varios de sus valores son diferentes de cero en sus correlograma de PACF y ACF, dando paso a un modelo ARMA con un valor p y q no evidente, por lo que se proponen varios modelos para la selección.

Tabla 3.11. Modelos propuestos para datos mensuales de potencia activa

Modelos	AIC	BIC	RMSE	MAE	MAPE (%)
SARIMA$(2,1,1)(3,0,0)_{12}$	1775,14	1803,34	7,889	5,862	1,075714
SARIMA$(1,1,2)(3,0,0)_{12}$	1775,25	1803,45	7,897	5,8729	1.078231
SARIMA$(2,1,2)(3,0,0)_{12}$	1777,39	1809,11	7,899	5,8722	1.078227

Los valores mostrados en la Tabla 3.11 obtenidos por los modelos propuestos son satisfactorios, mostrando que el modelo SARIMA$(2,1,1)(3,0,0)_{12}$, obtiene un mejor ajuste y menor error con respecto a los datos mensuales de entrenamiento de potencia activa, los modelos con mayor orden presentan un error pequeño pero mayor al modelo seleccionado, adicional es preferible mantener un orden menor.

Análisis de residuos

Una vez definidos todos los parámetros del modelo de potencia activa mensual, y en base a diferentes pruebas que permitieron seleccionar el modelo SARIMA$(2,1,1)(3,0,0)_{12}$ como el más apropiado, es necesario cerciorar que el modelo emplee la mayor cantidad de información disponible.

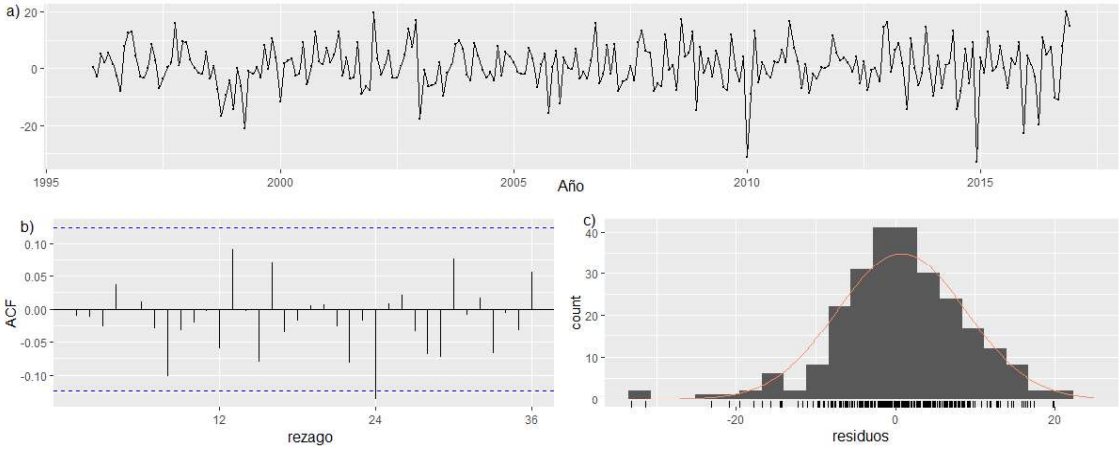

Figura 3.14. Análisis grafico de residuos para potencia activa mensual.

La Figura 3.14 a) evidencia que la media de los residuos obtenidos del modelo SARIMA$(2,1,1)(3,0,0)_{12}$ se mantiene alrededor de cero, mientras que la Figura 3.14 b) se forma con los valores de ACF de los residuos del modelo, estos valores forman un correlograma semejante al obtenido de un proceso de ruido blanco. Finalmente, la Figura 3.14 c) muestra que los residuos del modelo siguen una distribución normal. Todos estos

gráficos señalan que el modelo seleccionado es el adecuado para la serie mensual de potencia activa, debido a que ha capturado la mayor cantidad de información para el modelamiento.

Tabla 3.12. Pruebas a residuos del modelo SARIMA$(2,1,1)(3,0,0)_{12}$ de potencia activa mensual.

Prueba	**Valor de *p***
Ljung-Box	0,8062
Dickey Fuller aumentada	<2,2e-16
Durbin Watson **(Valor de *d*)**	2.0031

La Tabla 3.12 resume los resultados obtenidos por el modelo seleccionado en la prueba de Ljung-Box, el valor de *p* es mayor a 0.05 mostrando la independencia de los residuos [45]. La prueba de Dickey Fuller aumentada obtiene un valor mínimo de *p*. Finalmente la prueba de Durbin Watson obtiene un valor de $d = 2.0031$ lo que permite rechazar la hipótesis de la existencia de autocorrelación entre residuos.

3.3.3 PRONÓSTICO

Luego de realizar distintas pruebas al modelo y corroborar que los resultados son satisfactorios, se procede con el pronóstico que se muestra en la Tabla 3.13.

Tabla 3.13. Pronóstico de potencia activa, mediante el modelo SARIMA$(2,1,1)(3,0,0)_{12}$.

Mes	**Potencia activa pronosticada en R [MW]**	
	2017	**2018**
Enero	726,04	748,017
Febrero	721,25	748,567
Marzo	734,73	754,321
Abril	719	747,666
Mayo	729	755,899

Junio	733,239	757,842
Julio	725,79	754,347
Agosto	708,49	744,847
Septiembre	709,6	752,155
Octubre	717,6	756,587
Noviembre	764,06	776,097
Diciembre	750,78	759,082
Máximo	764,06	776,097
Tasa de crecimiento (%)	1,658	1,575

La Figura 3.15 muestra los valores históricos que fueron empleados como datos de entrenamiento en color negro. El pronóstico se realizó para 24 meses, es decir dos años desde enero del 2017 a diciembre de 2018 lo cual se muestra de color azul. Finalmente, en color gris se muestra el intervalo de confianza al 95%.

Figura 3.15. Pronóstico de potencia activa mensual mediante el modelo SARIMA$(2,1,1)(3,0,0)_{12}$.

3.3.4 ANÁLISIS DE RESULTADOS DEL PRONÓSTICO MENSUAL DE POTENCIA ACTIVA

El número de datos de entrenamiento de modelos mensuales son mayores que los utilizados para los modelos anuales. Los datos mensuales presentan mayores variaciones, lo que en un modelo anual no se puede visualizar debido a que solo presenta valores máximos o totales de todo el año.

La Tabla 3.14 muestra los resultados obtenidos al emplear el criterio AIC y BIC para seleccionar el modelo que presente un mayor ajuste con la serie mensual de potencia activa, el modelo SARIMA$(2,1,1)(3,0,0)_{12}$ es seleccionado por el criterio de Akaike (AIC) y Bayesiano (BIC) obteniendo los iguales resultados.

Tabla 3.14. Criterio de selección para serie mensual de potencia activa

Datos	**MAPE [%]**	
	AIC	**BIC**
	SARIMA$(2,1,1)(3,0,0)_{12}$.	SARIMA$(2,1,1)(3,0,0)_{12}$.
Entrenamiento (históricos)	1,08	1,08
Evaluación (reales)	1,42	1,42

En la Figura 3.16 se muestra en azul los valores de entrenamiento para el modelo mensual de potencia activa, en rojo los valores obtenidos o ajustados con el modelo SARIMA$(2,1,1)(3,0,0)_{12}$.

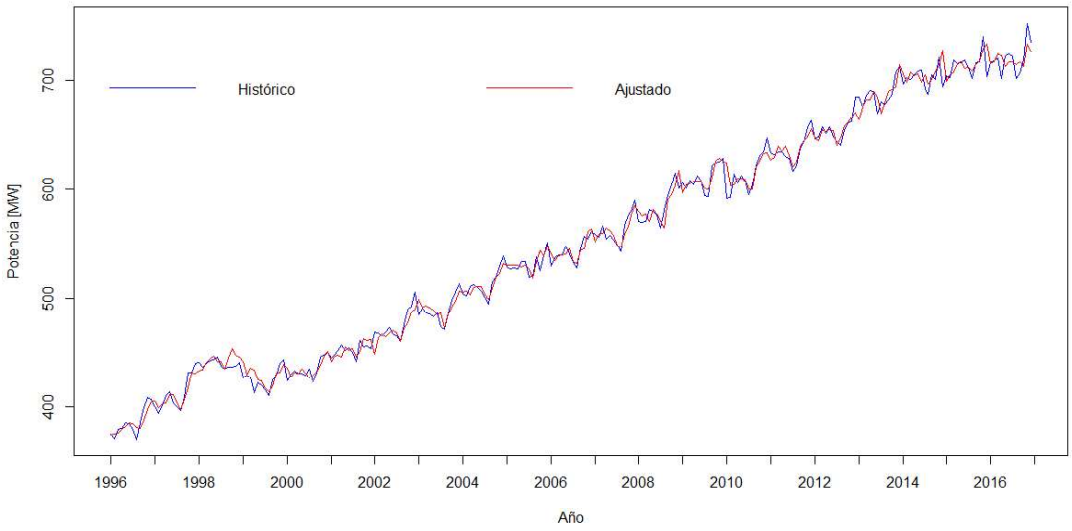

Figura 3.16. Comparativa entre datos de entrenamiento y valores ajustados del modelo SARIMA$(2,1,1)(3,0,0)_{12}$ para potencia activa.

La Tabla 3.15 contiene el pronóstico obtenido con el modelo SARIMA$(2,1,1)(3,0,0)_{12}$ y a partir de ello junto con los valores reales de potencia activa de enero de 2017 a diciembre de 2017 se calcula el MAPE. El pronóstico presentado por la Empresa Eléctrica Quito obtiene un valor de error menor al presentado con el modelo SARIMA, el año 2017 presentó un crecimiento mucho menor al que se venía obteniendo los años anteriores, es decir es un año que presenta un cambio significativo en la tendencia de demanda de potencia activa de la EEQ.

Tabla 3.15. Evaluación de potencia activa mensual para el año 2017 mediante MAPE

Fecha	**Potencia activa real [MW]**	**Potencia activa pronosticada en R [MW]**	**Potencia activa pronosticada por la EEQ [MW]**
Enero	733,94	726,04	733,33
Febrero	734,85	721,25	727,01
Marzo	739,85	734,73	736,26
Abril	734,15	719,00	729,75
Mayo	742,17	729,00	739,70
Junio	744,02	733,24	738,17
Julio	738,46	725,79	732,07
Agosto	728,97	708,49	728,50
Septiembre	737,89	709,60	738,17
Octubre	741,38	717,60	730,86
Noviembre	761,84	764,06	752,15
Diciembre	742,67	750,78	746,32
Máximo	761,84	764,06	752,15
MAPE (%)		1,821%	0,625%

En la Tabla 3.16 se muestra la comparativa entre datos reales, pronosticados en R y pronosticados por la EEQ para el año 2018. Este año recupera la tendencia que venía teniendo la serie de potencia activa mensual, por este motivo el modelo construido en lenguaje R obtiene un menor error que el obtenido en el año 2017 y en comparación con la metodología empleada por la EEQ.

Tabla 3.16. Evaluación de potencia activa mensual para el año 2018 mediante MAPE

Fecha	Potencia activa real [MW]	Potencia activa pronosticada en R [MW]	Potencia activa pronosticada por la EEQ [MW]
Enero	748,02	750,93	754,06
Febrero	748,57	747,10	745,41
Marzo	754,32	754,88	757,10
Abril	747,67	746,29	750,11
Mayo	755,90	741,25	760,51
Máximo	755,90	754,88	760,51
MAPE (%)		0,556%	0,507%

Luego de la evaluación del modelo construido en lenguaje R y el pronóstico obtenido por la metodología actual de la EEQ, se realiza el pronóstico de los siguientes siete meses mostrados en la Tabla 3.17.

Tabla 3.17. Pronóstico del modelo mensual de potencia activa

Fecha	Potencia activa pronosticada en R [MW]	Potencia activa real [MW]
Junio-18	729,432	722,15
Julio-18	735,329	739,98
Agosto-18	723,046	714,31
Septiembre-18	726,912	723,48
Octubre-18	735,070	731,64

Noviembre-18	737,997	728,92
Diciembre-18	740,738	765,47

3.4 FACTOR DE CARGA MENSUAL

La Tabla 3.18 muestra el valor de factor de carga mensual obtenido a partir de los resultados de energía y potencia activa. Los resultados se encuentran dentro de los valores obtenidos por la EEQ.

Tabla 3.18. Factor de carga mensual para el año 2017

Fecha	Factor de carga real (%)	Factor de carga mensual resultados R (%)	Factor de carga mensual Resultados EEQ (%)
Enero	66,1	63,6	65,5
Febrero	66,4	65,7	67,4
Marzo	67,1	63,7	65,8
Abril	66,5	64,5	66,6
Mayo	66,5	63,7	67,6
Junio	66,4	64,4	66,5
Julio	65,6	64,5	67,0
Agosto	66,9	64,8	66,6
Septiembre	68,2	65,6	67,4
Octubre	68,0	65,8	68,4
Noviembre	63,9	64,2	66,4
Diciembre	63,9	64,4	63,8
MAPE (%)		2,779	1,21

La Tabla 3.19 muestra el factor de carga obtenido para el año 2018, a través del pronóstico de potencia activa y energía, el MAPE obtenido mediante la metodología de la EEQ muestra un menor error, sin embargo el MAPE obtenido del factor de carga mediante el resultado de los pronósticos realizados en R se mantiene dentro del 15% permitido.

Tabla 3.19. factor de carga mensual para el año 2018

Fecha	**Factor de carga real (%)**	**Factor de carga mensual resultados R (%)**	**Factor de carga mensual resultados EEQ (%)**
Enero	65,2	64,0	65,8
Febrero	65,3	66,0	70,3
Marzo	65,9	64,2	65,9
Abril	66,5	65,2	66,9
Mayo	67,6	64,3	68,1
MAPE (%)		2,5	1,97

4. PRONÓSTICOS A LARGO PLAZO

En este capítulo se realiza el procedimiento para la obtención del pronóstico a largo plazo de los datos anuales de energía y potencia proporcionados por la EEQ que previamente fueron sometidos a un proceso de minería de datos que ha sido detallado en la sección 2.2.2.

Los resultados a largo plazo son de las series anuales y tienen un horizonte de 10 años para cumplimiento de la Regulación No. CONELEC 013/08. Los datos de aprendizaje son aquellos disponibles hasta el año 2016 y los datos para validación dependen de los datos provistos por la Empresa Eléctrica Quito. Al realizar la selección de los modelos anuales mediante el criterio de AIC y BIC, se obtuvo los mismos modelos como se detalla más adelante.

4.1 ENERGÍA ANUAL

4.1.1 PREPARACIÓN DE LOS DATOS DE ENERGÍA ANUAL

Los principios señalados en la sección 3.2.1 son empleados para la examinación y preparación de los datos de energía. Se considera el ordenamiento de los datos y la inspección de estos en búsqueda de datos atípicos que pudieran afectar al modelamiento.

4.1.2 MODELAMIENTO DE DATOS ANUALES DE ENERGÍA

El proceso de modelamiento empleado para datos anuales es el mismo mencionado en la sección 3.2.2

Análisis gráfico de datos

La serie de datos históricos anuales de energía está comprendida desde el año 1975 hasta el año 2017, de los cuales se emplearon como datos de entrenamiento desde el año 1975 hasta el año 2016 y para validación se dispuso únicamente del valor del año 2017. La Figura 4.1 muestra la serie de tiempo de energía anual a ser empleada para realizar su pronóstico.

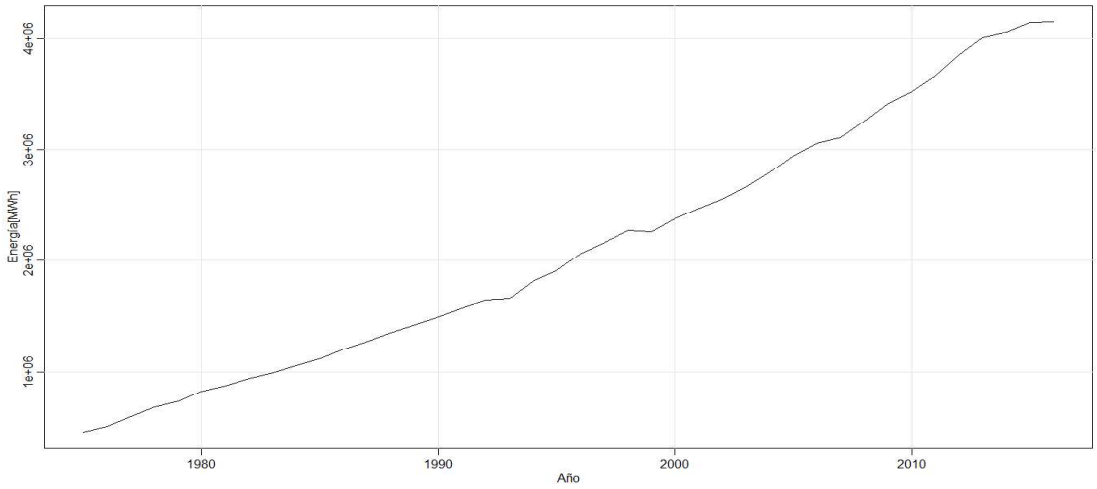

Figura 4.1. Datos de entrenamiento de energía anual.

De la serie de datos históricos o datos de entrenamiento puede ser extraída su tendencia como muestra la Figura 4.2.

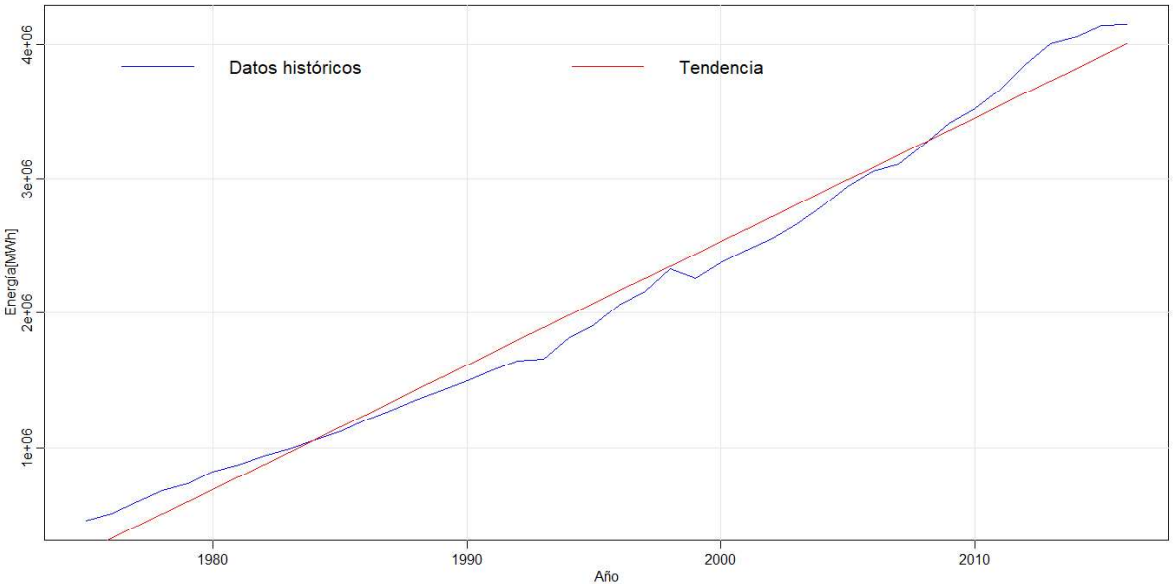

Figura 4.2. Datos de entrenamiento de energía anual y su tendencia.

La Figura 4.2 muestra la existencia de una tendencia significativa, por este motivo la serie es no estacionaria. Para comprobar lo expuesto por la Figura 4.2 se emplea la prueba de Dickey Fuller aumentada obteniendo como resultado un valor de $p = 0{,}7696$, no se puede rechazar la hipótesis nula y se confirma que se trata de una serie no estacionaria.

Análisis de la tendencia

Las series que presentan tendencia o series no estacionarias requieren de un proceso de diferenciación como se explicó en la sección 2.4.6. Se seleccionará únicamente la que obtenga un valor menor valor de *p* en la prueba de Dickey Fuller.

Debido a la tendencia expresada en la Figura 4.2, se propone una diferenciación regular de orden uno o dos, en datos anuales no se presenta tendencia estacional debido a que se cuenta únicamente con un dato de todo el año. Al aplicar estas diferenciaciones a los datos de entrenamiento se obtiene la información organizada en la Tabla 4.1.

Tabla 4.1. Prueba de Dickey Fuller aumentada para datos de energía anual.

Diferenciación	Modelo	Valor de *p*
Regular	$(p, 1, q)$	$0{,}2974$
Regular	$(p, 2, q)$	$0{,}01713$
Regular	$(p, 3, q)$	$0{,}1771$

La Tabla 4.1 muestra los valores de *p* obtenidos en la prueba Dickey Fuller, los cuales fueron obtenidos por la diferencia regular de orden 1, 2 y 3 y presentan una gran diferencia en los valores de *p* obtenidos respectivamente. La remoción de la tendencia se logra de forma adecuada al ser de segundo orden debido a que obtiene el menor valor de *p* en la prueba Dickey Fuller.

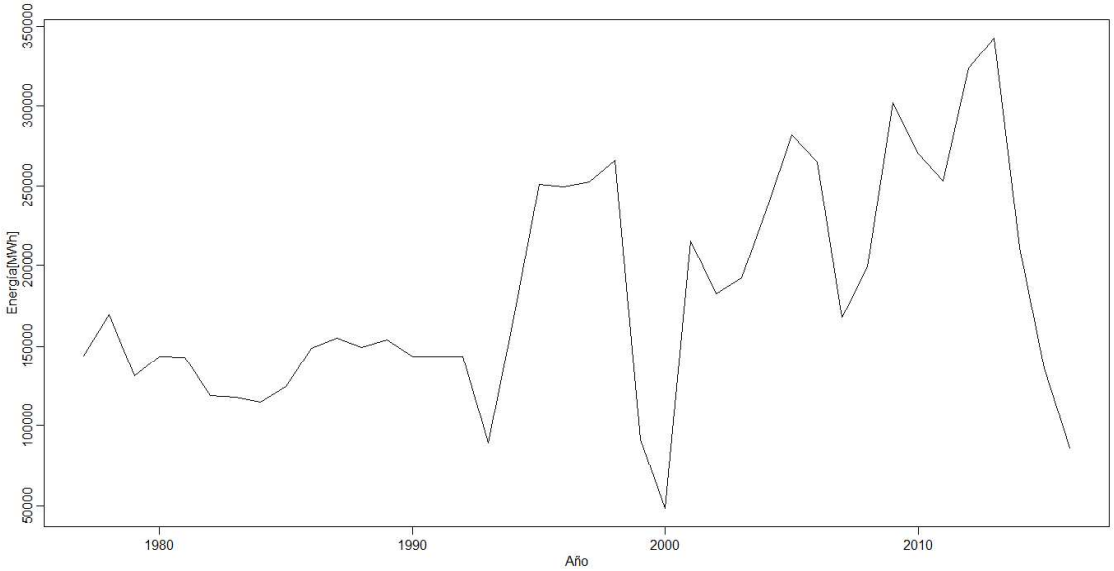

Figura 4.3. Diferenciación estacional de los datos de energía anual.

La Figura 4.3 muestra la serie de energía anual una vez que ha sido diferenciada, de este modo sus valores no presentan una tendencia.

Análisis ACF y PACF

El siguiente paso consiste en analizar el orden del modelo que se encuentra en construcción. Se grafica los valores de ACF y PACF de la serie ya diferenciada en la Figura 4.3.

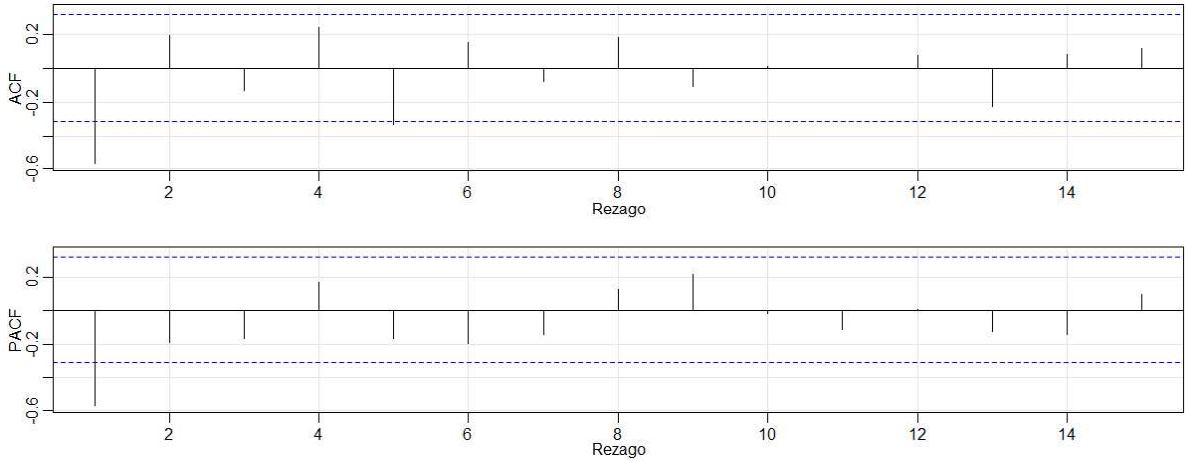

Figura 4.4. Valores de ACF y PACF para serie de energía anual diferenciada.

El correlograma de la Figura 4.4 es obtenido luego de remover la tendencia de la serie anual de energía. Los valores de la función de autocorrelación parcial decrecen y ningún rezago presenta un valor significante, mientras que el correlograma con los valores de autocorrelación muestra que el rezago quinto sobrepasa el intervalo de confianza. Se propone un modelo de media móvil MA(q) donde q es igual a cinco; sin embargo, se realizan modelos con rezagos menores considerando el principio de Parsimonia.

Tabla 4.2. Observaciones de los valores de ACF y PACF de la serie diferenciada.

	ARMA (0,5)
ACF	Los valores son insignificantes a partir del rezago 5
PACF	Decrece

La Tabla 4.2 resume los resultados obtenidos de la observación de los correlogramas de autocorrelación y autocorrelación parcial con los cuales se proponen los siguientes modelos para su posterior evaluación.

Tabla 4.3. Modelos propuestos para datos anuales de energía

Modelos	AIC	RMSE	MAE	MAPE (%)
ARIMA$(0,2,1)$	982,77	47710,75	32683,57	1,5377
ARIMA$(0,2,2)$	984,7	47675,48	32772.74	1,5381
ARIMA$(0,2,3)$	985,82	47131,84	32541,69	1,5275
ARIMA$(0,2,4)$	987,71	46855,19	32485,33	1,5009
ARIMA$(0,2,5)$	987,87	45144,51	30205,36	1,4534

La Tabla 4.3 muestra varios indicadores de los modelos planteados, se distingue que los modelos presentan un mejor ajuste al incrementar el valor de *p*. Sin embargo, el valor de MAPE es similar para todos los modelos y no justifica un modelo de orden mayor. En concordancia con el principio de Ockam o Parsimonia se toma el modelo ARIMA$(0,2,1)$.

Análisis de residuos

Una vez definido el modelo a emplearse, se debe realizar distintas pruebas para comprobar que éste capture la mayor cantidad de información relevante de los datos de entrenamiento.

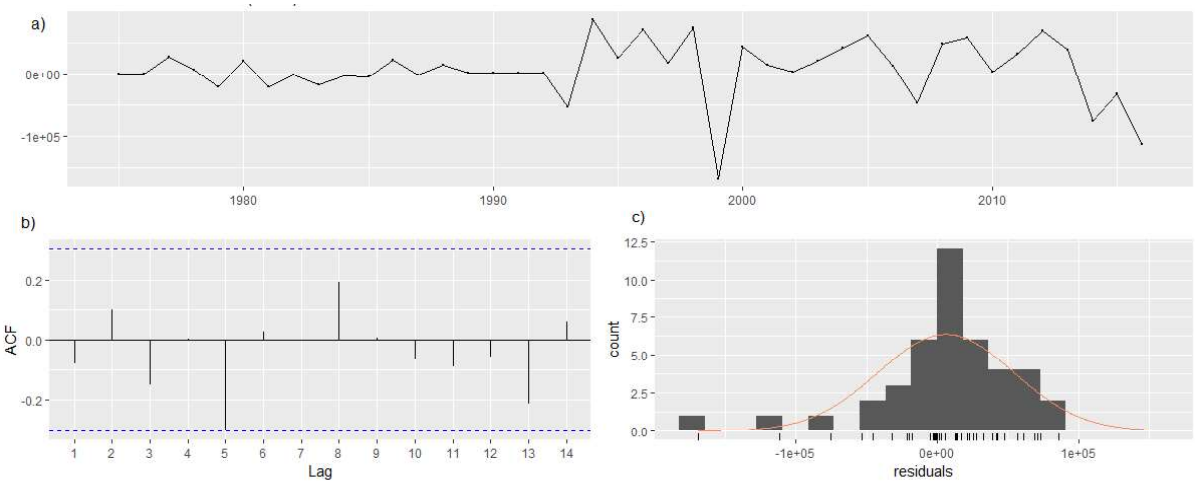

Figura 4.5. Valores de ACF y PACF para el modelo ARIMA$(0,2,1)$ de energía anual diferenciada.

La Figura 4.5 a) muestra la media obtenida de los residuos del modelo ARIMA$(0,2,1)$, manteniéndose alrededor de cero de acuerdo con lo mencionado en la sección 2.5.3. La Figura 4.5 b) es el correlograma de los valores de autocorrelación obtenida de los residuos del modelo, los valores se acercan a los obtenidos por un proceso de ruido blanco. Finalmente, la Figura 4.5 c) muestra que los residuos del modelo siguen una distribución

normal. Cada una de las distintas graficas de la Figura 4.5 aporta a la conclusión de que el modelo ARIMA$(0,2,1)$, ha capturado la mayor cantidad de información presentada en la serie de tiempo de demanda de energía anual.

Sin embargo, se presenta dos pruebas más para desechar la hipótesis de correlación entre residuos del modelo. Los resultados obtenidos son escritos en la Tabla 4.4.

Tabla 4.4. Pruebas a residuos para el modelo ARIMA$(0,2,1)$ de energía anual.

Prueba	**Valor de *p***
Ljung-Box	0,4647
Dickey Fuller aumentada	1,554e-6
Durbin Watson **(Valor de *d*)**	1,9804

El modelo planteado en la prueba de Ljung-Box obtiene un valor de *p* mayor a 0.05 mostrando la independencia de los residuos [45] y en la prueba de Dickey Fuller aumentada muestra un valor mínimo de *p*. Finalmente la prueba de Durbin Watson obtiene un valor de *d* muy cercano a dos, lo que permite concluir que la correlación entre residuos es mínima.

4.1.3 PRONÓSTICO

Luego de haber construido el modelo mediante la observación y aplicación de pruebas se realiza el pronóstico del modelo ARIMA$(0,2,1)$, obtenido con los datos anuales de energía en MWh. Los resultados se muestran en la Tabla 4.5.

Tabla 4.5. Pronóstico de energía anual, mediante el modelo ARIMA$(0,2,1)$.

Año	**Energía [MWh]**
2017	4’246.029,270
2018	4’345.339,678
2019	4’444.650,087
2020	4’543.960,495
2021	4’643.270,904
2022	4’742.581,312

2023	4’841.891,720
2024	4’941.202,129
2025	5’040.512,537
2026	5’139.822,946

La Figura 4.6 reúne los datos históricos de energía anual que sirvieron como datos de entrenamiento (negro), los valores pronosticados desde el año 2017 al año 2026 se encuentran en color rojo y el nivel de confianza para el intervalo de predicción se dispuso al 95%.

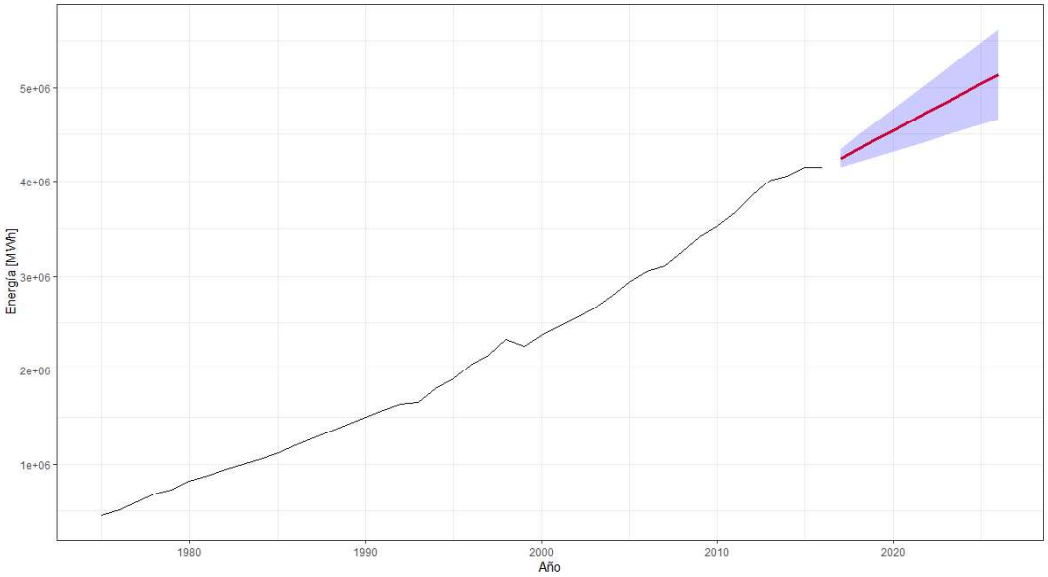

Figura 4.6. Pronóstico anual de energía.

4.1.4 ANÁLISIS DE RESULTADOS DEL PRONÓSTICO ANUAL DE ENERGÍA

La Tabla 4.6 muestra los valores de MAPE con respecto al año de evaluación, el modelo construido en RStudio presenta un menor error y con esto un mejor ajuste que el obtenido por el modelo propuesto por la EEQ; sin embargo, los dos resultados se mantienen en el rango aceptable de ±15%.

Tabla 4.6. Pronóstico anual de energía y MAPE

Año	Energía real [MWh]	Energía pronosticada en RStudio [MWh]	Energía pronosticada por EEQ S.A [MWh]

2017	4’228.000,296	4’246.029,270	4’292.607,778
MAPE (%)		0,426	1,528

La Tabla 4.7 muestra los valores de pronóstico obtenidos mediante la construcción del modelo ARIMA en lenguaje R y los resultados obtenidos con la metodología actual de la EEQ. Inicialmente los valores se mantienen bastante cercanos, pero con el paso de los años el pronóstico de la EEQ es mayor al pronóstico de realizado en lenguaje R.

Tabla 4.7. Pronóstico anual de energía obtenido en R y por la EEQ

Año	**Energía pronosticada en R [MWh]**	**Energía pronosticada por la EEQ [MWh]**
2017	4’246.029,270	4’292.607,778
2018	4’345.339,68	4’439.405,85
2019	4’444.650,087	4’579.139,741
2020	4’543.960,495	4’721.182,000
2021	4’643.270,904	4’865.565,777
2022	4’742.581,312	5’012.327,940
2023	4’841.891,720	5’161.508,085
2024	4’941.202,129	5’313.148,528
2025	5’040.512,537	5’467.294,313
2026	5’139.822,946	5’623.993,204

La Figura 4.7 muestra una comparativa grafica entre los datos históricos de energía anual de la EEQ de color rojo y los valores obtenidos mediante el modelo ARIMA$(0,2,1)$ que se encuentran de color azul.

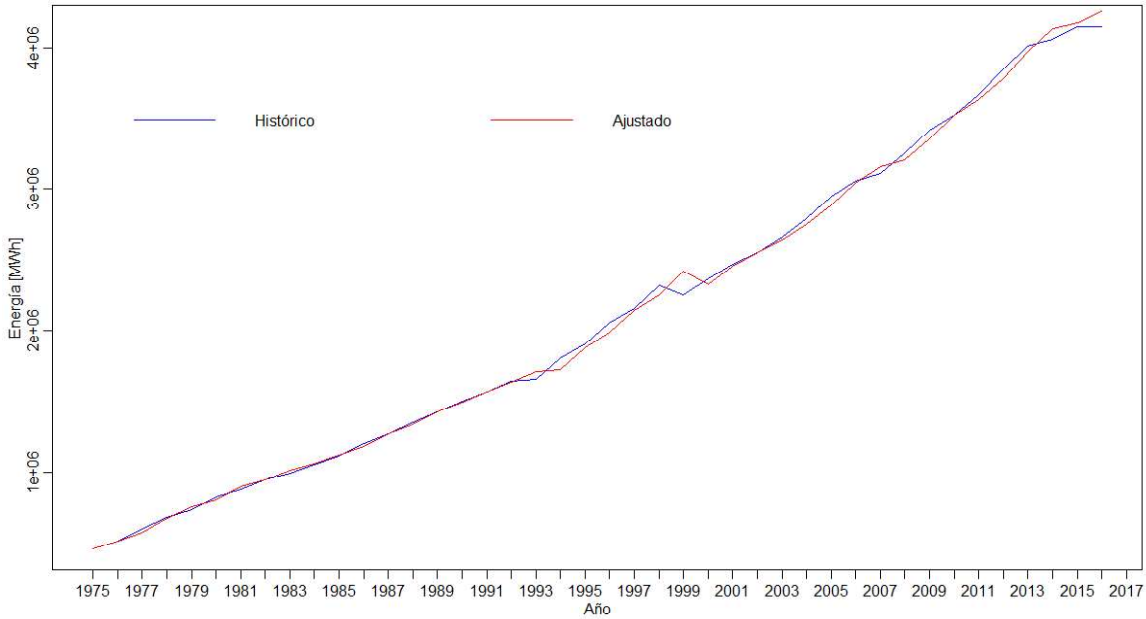

Figura 4.7. Datos históricos y ajustados de energía anual de la EEQ, modelados ARIMA$(0,2,1)$.

Se puede observar que los valores futuros dependen de los valores y errores incluso de dos periodos atrás, esto como muestra la Ecuación 4.1.

$$y_t = e_t - 0{,}83145e_{t-1} + 2y_{t-1} - y_{t-2} + \mu \tag{4.1}$$

4.2 POTENCIA ACTIVA ANUAL

4.2.1 PREPARACIÓN DE LOS DATOS DE POTENCIA ACTIVA ANUAL

Los principios señalados en la sección 3.3.1 son empleados para la examinación y preparación de los datos de potencia activa. Se considera el ordenamiento de los datos y la inspección de estos en búsqueda de datos atípicos que pudieran afectar al modelamiento.

4.2.2 MODELAMIENTO DE DATOS ANUALES DE POTENCIA ACTIVA

El proceso de modelamiento empleado para datos anuales es el mencionado en la sección 3.2.2.

Análisis grafico de datos

La serie de datos históricos anuales de potencia activa está comprendida desde el año 1970 hasta el año 2016. La Figura 4.8 muestra esta serie de tiempo.

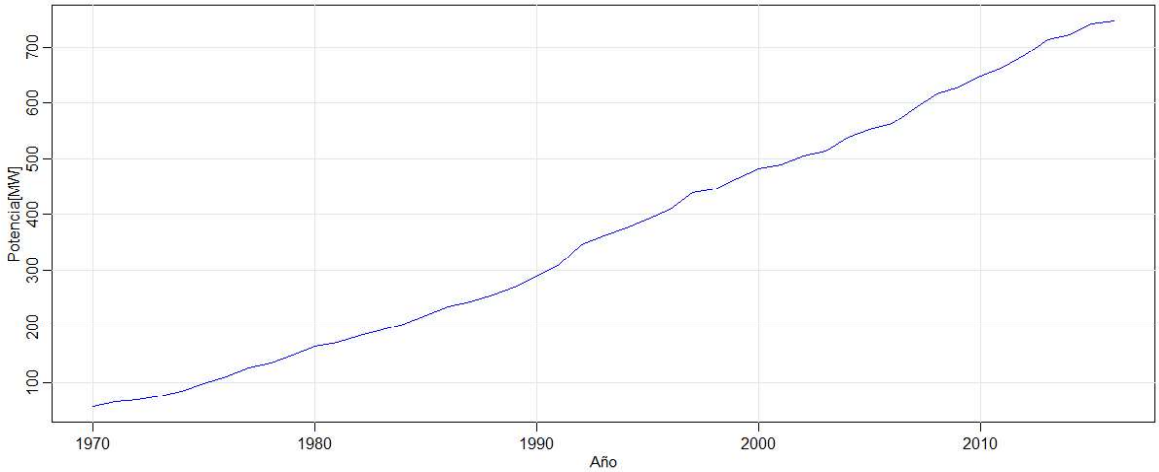

Figura 4.8. Datos de entrenamiento de potencia activa anual.

De la serie de datos históricos o datos de entrenamiento de potencia activa puede ser extraída su tendencia como muestra la Figura 4.9.

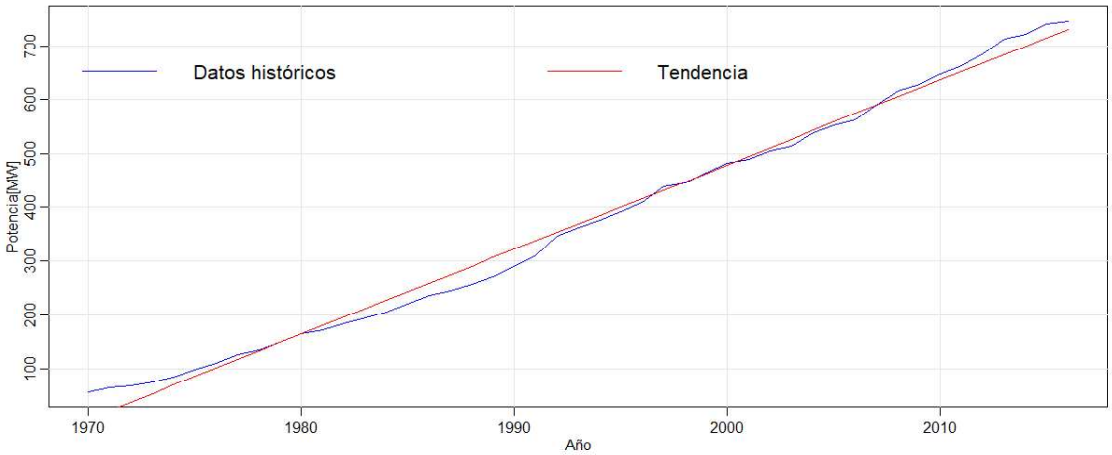

Figura 4.9. Datos de entrenamiento de potencia activa anual y su tendencia.

Se puede notar que existe una tendencia, por este motivo la serie es no estacionaria. Para comprobar lo expuesto por la Figura 4.9 se emplea la prueba de Dickey Fuller aumentada obteniendo como resultado un valor de $p = 0{,}483$, no se puede rechazar la hipótesis nula y se confirma que se trata de una serie no estacionaria.

Análisis de la tendencia

Debido a que la serie presenta una tendencia requiere de un proceso de diferenciación como se explicó en la sección 2.4.6. Se selecciona únicamente la diferenciación que

obtenga un valor menor de *p* en la prueba de Dickey Fuller. Se propone una diferenciación regular de orden uno o dos, en datos anuales no se presenta tendencia estacional debido a que se cuenta únicamente con un dato de todo el año. Al aplicar estas diferenciaciones a los datos de entrenamiento se obtiene la información organizada en la Tabla 4.8.

Tabla 4.8. Prueba de Dickey Fuller aumentada para datos de potencia activa anual.

Diferenciación	Modelo	Valor de *p*
Regular	$(p, 1, q)$	0,08421
Regular	$(p, 2, q)$	0,01
Regular	$(p, 3, q)$	0,5804

La Tabla 4.8 muestra los valores de *p* obtenidos en la prueba Dickey Fuller, el valor mínimo de *p* es obtenido por una diferenciación de orden d igual a dos.

La Figura 4.10 muestra la serie de potencia activa anual luego de haber sido diferenciada dos veces para remover su tendencia, como se puede observar sus valores fluctúan alrededor de 0.

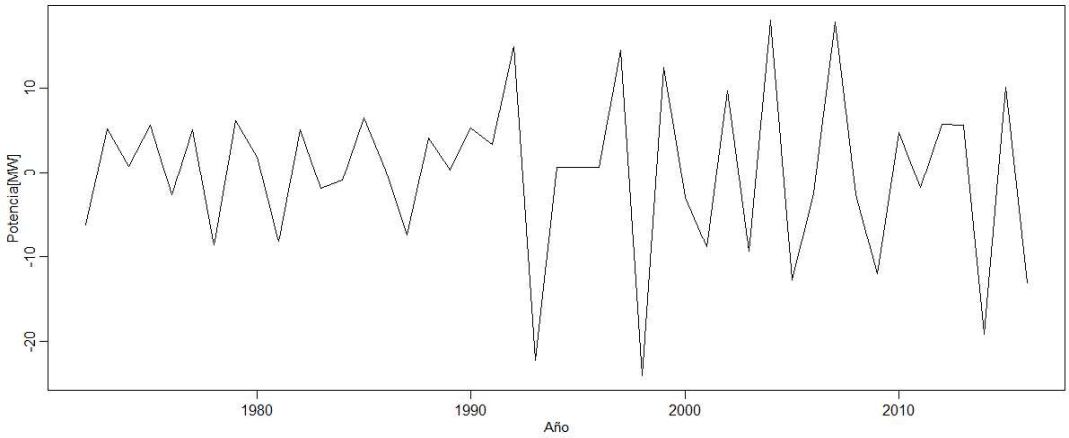

Figura 4.10. Diferenciación estacional de los datos de potencia activa anual.

Análisis ACF y PACF

El siguiente paso consiste en analizar el orden del modelo que se encuentra en construcción. Se grafica los valores de ACF y PACF de la serie ya diferenciada mostrada en la Figura 4.11.

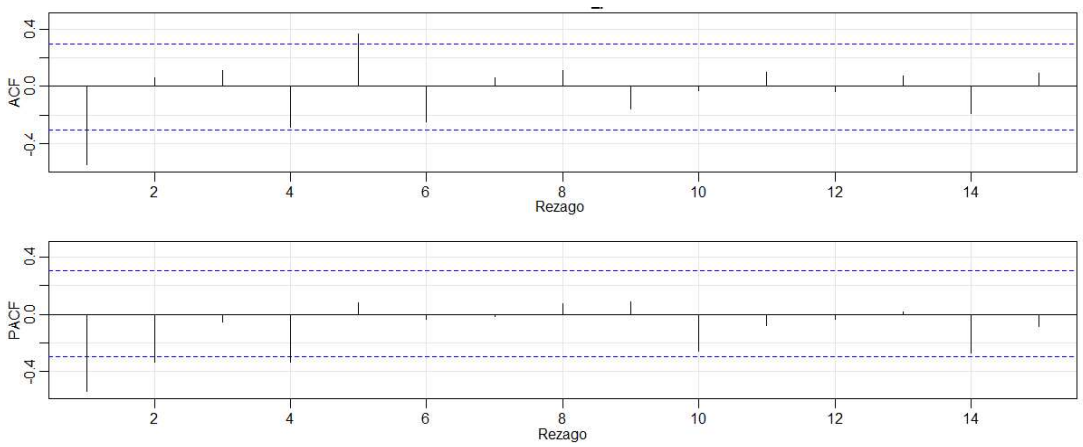

Figura 4.11. Valores de ACF y PACF para serie de potencia activa anual diferenciada.

El correlograma de la Figura 4.11 es obtenido luego de remover la tendencia de la serie anual de potencia activa. Los valores de la función de autocorrelación parcial decrecen y ningún rezago presenta algún valor significante, mientras que el correlograma con los valores de autocorrelación muestra que el rezago quinto sobrepasa el intervalo de confianza. Se propone un modelo de media móvil MA(q) donde q es igual a cinco; sin embargo, se realizan modelos con rezagos menores considerando el principio de Parsimonia.

Tabla 4.9. Observaciones de los valores de ACF y PACF de la serie diferenciada.

	ARMA(0,5)
ACF	Los valores son insignificantes a partir del rezago 5.
PACF	Decrece

La Tabla 4.9 muestra los resultados obtenidos de la observación de los correlogramas de autocorrelación y autocorrelación parcial con los cuales se proponen los siguientes modelos para su posterior evaluación.

Tabla 4.10. Modelos propuestos para datos anuales de potencia activa

Modelos	**AIC**	**RMSE**	**MAE**	**MAPE (%)**
ARIMA$(0,2,1)$	982,77	47710,75	32683,57	1,537728
ARIMA$(0,2,2)$	984,7	47675,48	32772.74	1,538121

ARIMA$(0,2,3)$	985,82	47131,84	32541,69	1,527531
ARIMA$(0,2,4)$	987,71	46855,19	32485,33	1,500887
ARIMA$(0,2,5)$	987,87	45144,51	30205,36	1,453375

La Tabla 4.10 muestra varios indicadores de los modelos planteados, se distingue que los modelos presentan un mejor ajuste al incrementar el valor de *q*. Sin embargo, el valor de MAPE es similar para todos los modelos y no justifica un modelo de orden mayor. En concordancia con el principio de Ockam o se toma el modelo ARIMA$(0,2,1)$.

Análisis de residuos

Una vez definido el modelo a emplearse, se debe realizar distintas pruebas a sus residuos para comprobar que éste capture la mayor cantidad de información relevante de los datos de entrenamiento. La Figura 4.12 a) muestra la media obtenida de los residuos del modelo ARIMA$(0,2,1)$, manteniéndose alrededor de cero de acuerdo con lo mencionado en la sección 2.5.3. La Figura 4.12 b) es el correlograma de los valores de autocorrelación obtenida de los residuos del modelo, los valores se acercan a los obtenidos por un proceso de ruido blanco. Finalmente, la Figura 4.12 c) muestra que los residuos del modelo siguen una distribución normal.

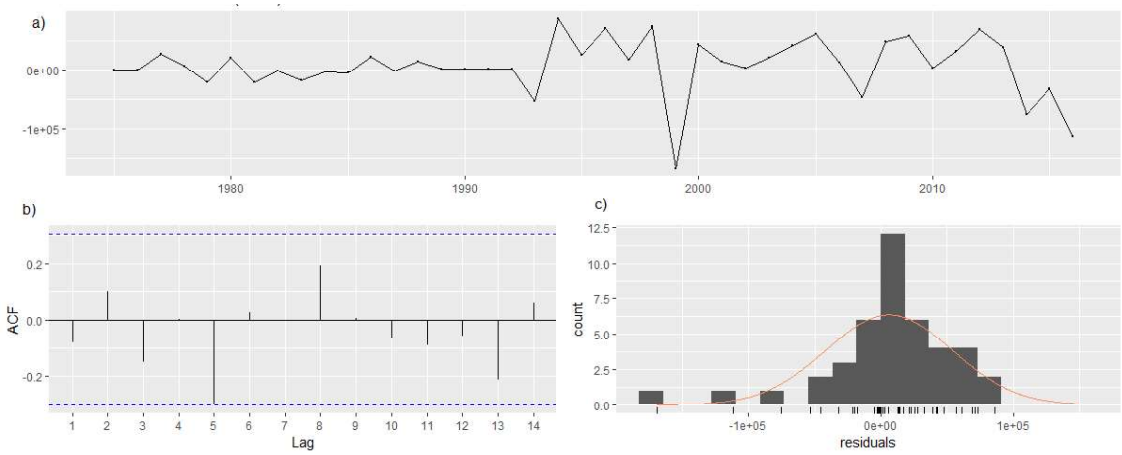

Figura 4.12. Valores de ACF y PACF para el modelo ARIMA$(0,2,1)$ de potencia activa anual diferenciada.

Cada una de las distintas graficas de la Figura 4.12 aporta a la conclusión de que el modelo ARIMA$(0,2,1)$, ha capturado la mayor cantidad de información presentada en la serie de tiempo de potencia activa anual.

Sin embargo, se realizan tres pruebas más para desechar la hipótesis de correlación entre residuos del modelo. Los resultados obtenidos son escritos en la Tabla 4.11.

Tabla 4.11. Pruebas a residuos del modelo ARIMA$(0,2,1)$ de potencia activa anual.

Prueba	**Valor de *p***
Ljung-Box	0,4647
Dickey Fuller aumentada	1,554e-6
Durbin Watson **(Valor de *d*)**	2.110

El modelo planteado en la prueba de Ljung-Box obtiene un valor de p mayor a 0.05 mostrando la independencia de los residuos [45]. En la prueba de Dickey Fuller aumentada muestra un valor mínimo de *p* y finalmente en la prueba de Durbin Watson se obtuvo un valor de *d* cercano a 2 lo que permite concluir que la correlación entre residuos es mínima.

4.2.3 PRONÓSTICO

Luego de haber construido el modelo mediante la observación y aplicación de pruebas se realiza el pronóstico del modelo ARIMA$(0,2,1)$ para potencia activa anual. Los resultados se muestran en la Tabla 4.12.

Tabla 4.12. Pronóstico de potencia activa anual, mediante el modelo ARIMA$(0,2,1)$.

Año	**Potencia activa [MW]**
2017	761,522
2018	777,654
2019	793,786
2020	809,918
2021	826,050
2022	842,182
2023	858,314
2024	874,446
2025	890,578

2026	906,710

Los valores para el año 2017 y 2018 de la Tabla 4.12 difieren de los valores máximos de potencia activa obtenidos en la Tabla 3.13, debido a que los datos de entrenamiento son diferentes, en el primer caso se emplean los valores totales de todo un año mientras que en el segundo caso se empleando 12 datos por año, esto permite que para cada caso existe un modelo diferente que toma distintas componentes.

La Figura 4.13 reúne los datos históricos de potencia activa anual que sirvieron como datos de entrenamiento (negro), los valores pronosticados desde 2017 hasta el año 2026 se encuentran en color rojo y el nivel de confianza para el intervalo de predicción se dispuso al 95%.

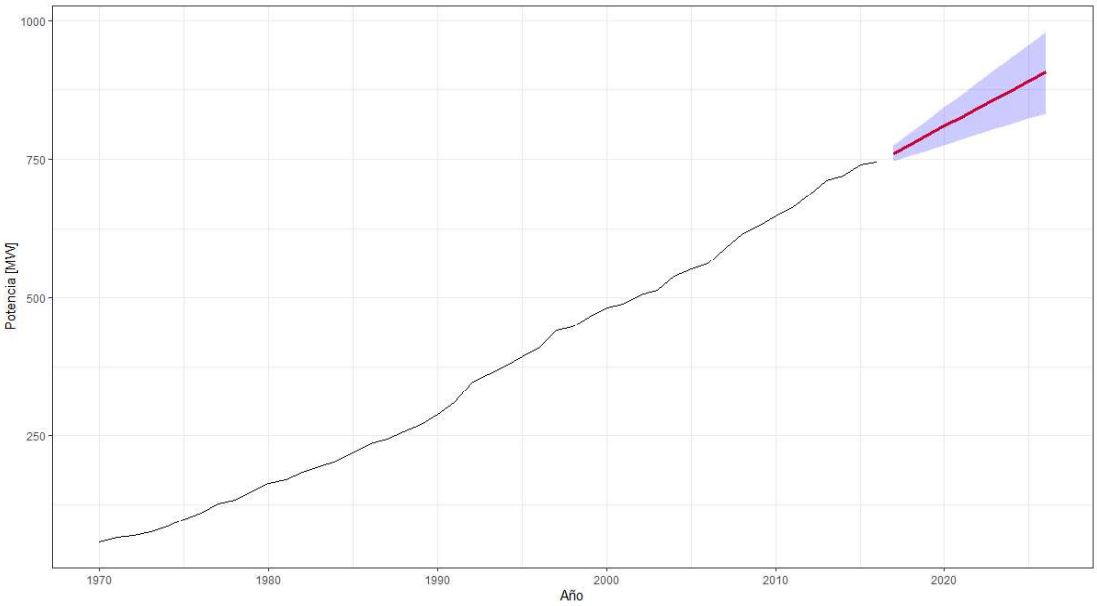

Figura 4.13. Pronóstico de potencia activa anual.

4.2.4 ANÁLISIS DE RESULTADOS DEL PRONÓSTICO DE POTENCIA ACTIVA ANUAL

El error porcentual absoluto medio mencionado en la sección 2.5.2 es el método más común de evaluación de pronósticos debido a su facilidad en la interpretación. El valor aceptable del MAPE se encuentra en un rango de ±15% [46]. El error porcentual absoluto medio calcula el ajuste del modelo con respecto a los datos reales ya sean datos de evaluación o datos de aprendizaje.

Tabla 4.13. Pronóstico anual de potencia activa y MAPE

Año	Potencia activa real [MW]	Potencia activa pronosticada en RStudio [MW]	Potencia activa pronosticada por EEQ [MW]
2017	750,780	761,522	752,145
MAPE (%)		1,431	0,182

En la Tabla 4.13 se muestra el valor del MAPE entre el valor real de potencia activa obtenido en el año 2017 y el valor de predicción obtenido en RStudio, siendo éste del 1.431% manteniéndose dentro del rango aceptable. También se muestra los valores obtenidos por la metodología actual empleada por la Empresa Eléctrica Quito siendo este valor de error menor al logrado por los modelos ARIMA.

La Tabla 4.14 muestra el pronóstico obtenido para los datos de potencia activa anual mediante el modelo ARIMA$(0,2,1)$, el pronóstico presentado por la EEQ se muestra con valores superiores al desarrollado en RStudio.

Tabla 4.14. Pronóstico anual de potencia activa

Año	Potencia activa pronosticada en R [MW]	Potencia activa pronosticada por la EEQ [MW]
2017	761,522	752,145
2018	777,654	772,333
2019	793,786	792,707
2020	809,918	813,268
2021	826,050	834,016
2022	842,182	854,950
2023	858,314	876,072
2024	874,446	897,380
2025	890,578	918,875
2026	906,710	940,556

La Figura 4.14 muestra una comparativa entre la gráfica obtenida con los valores de entrenamiento de potencia activa anual (rojo) y los obtenidos con el modelo elaborado en lenguaje R (azul).

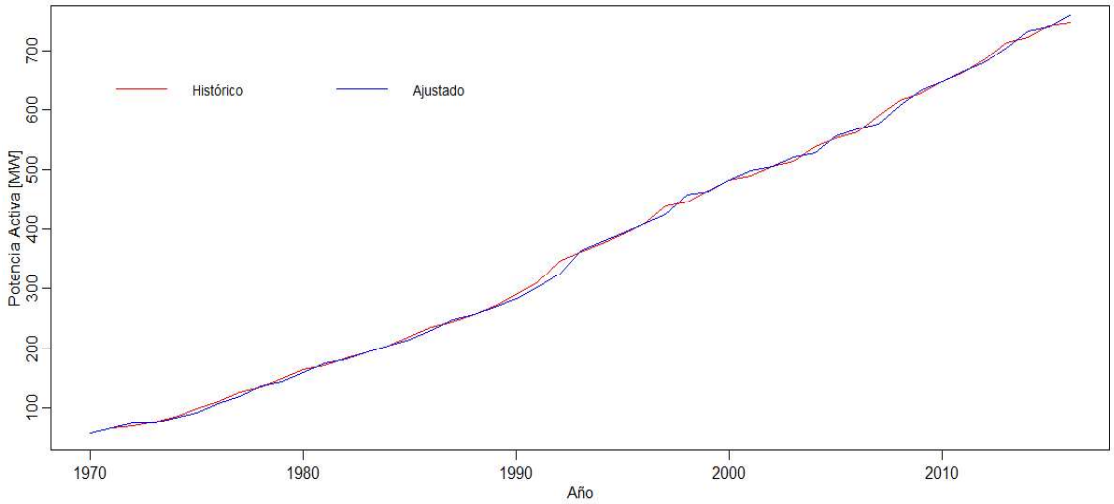

Figura 4.14. Datos históricos y ajustados de potencia activa anual para la EEQ, modelados ARIMA$(0,2,1)$

La Ecuación 4.2 fue construida a partir de la Ecuación 2.32, y se puede verificar que existen dos diferenciaciones debido a que cuenta con un término y_{t-2}, esto se debe a que la tendencia fue cuadrática como se mostró en la Figura 2.14 de la sección 2.4.6.

$$y_t = e_t - 0{,}9034745e_{t-1} + 2y_{t-1} - y_{t-2} + \mu \tag{4.2}$$

4.3 POTENCIA REACTIVA ANUAL

4.3.1 PREPARACIÓN DE LOS DATOS DE POTENCIA REACTIVA ANUAL

La serie de datos de potencia reactiva tiene un comportamiento donde su patrón no puede ser fácilmente identificado debido a la aleatoriedad de sus datos. Se considera el ordenamiento de los datos y la inspección de estos en búsqueda de datos atípicos que pudieran afectar al modelamiento.

4.3.2 MODELAMIENTO DE DATOS ANUALES DE POTENCIA REACTIVA

El proceso de modelamiento empleado para datos anuales es el mismo mencionado en la sección 3.2.2.

Análisis grafico de datos

La serie de datos históricos anuales de potencia reactiva está comprendida desde el año 1993 hasta el año 2017, de los cuales se emplearon como datos de entrenamiento desde

el año 1993 hasta el año 2016 y para validación se dispuso únicamente del valor del año 2017. La Figura 4.15 muestra la serie de tiempo potencia reactiva a ser empleada para realizar su pronóstico.

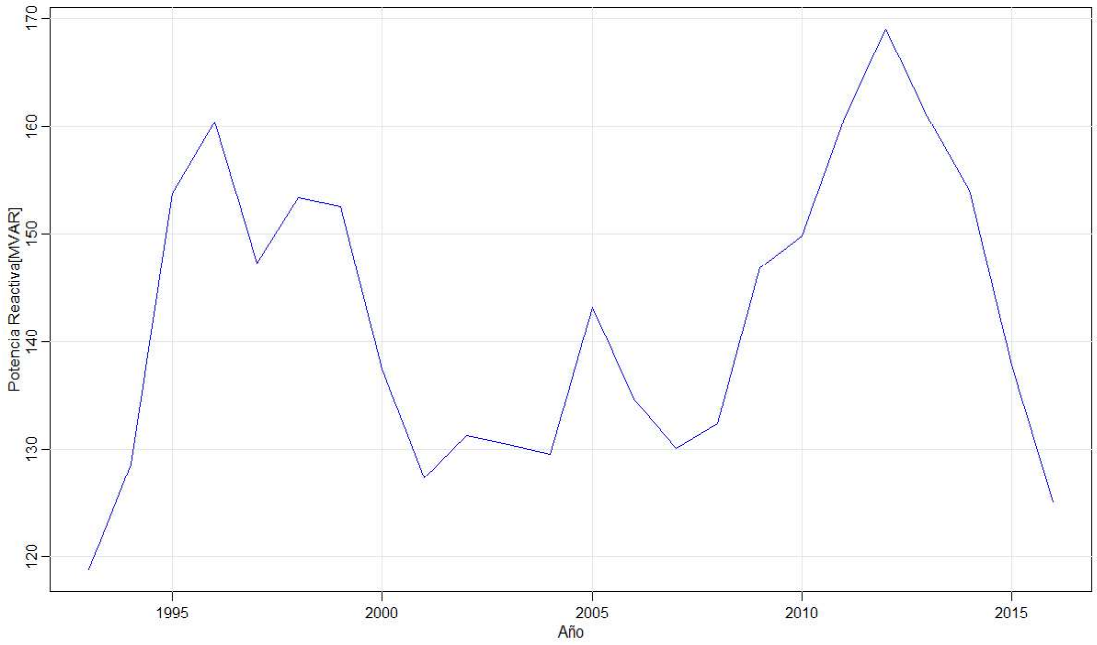

Figura 4.15. Datos de entrenamiento de potencia reactiva anual.

De la serie de datos históricos o datos de entrenamiento puede ser extraída su tendencia como muestra la Figura 4.16.

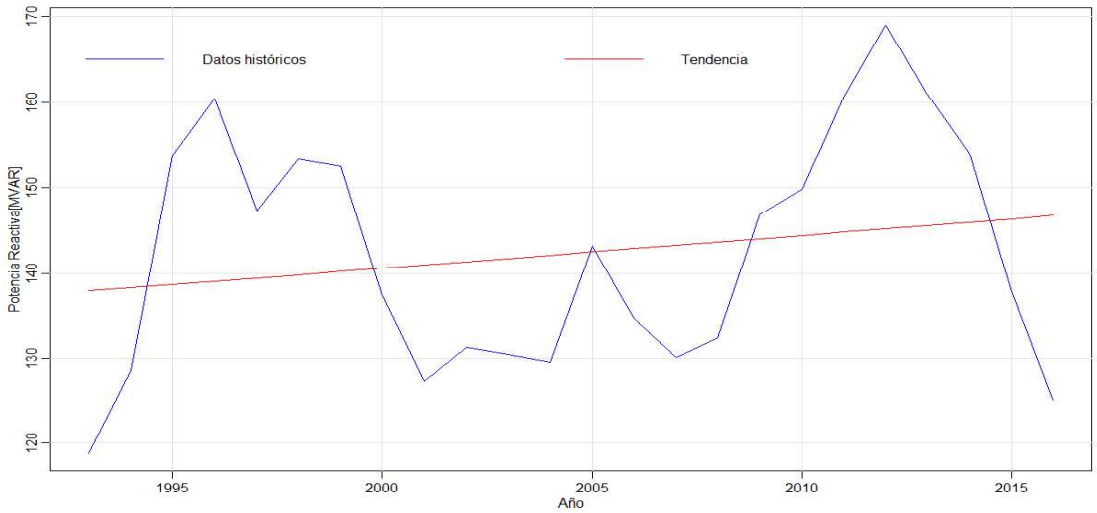

Figura 4.16. Datos de entrenamiento de potencia reactiva anual y su tendencia.

La tendencia presentada por la serie de potencia reactiva anual no es significativa por este motivo la serie es estacionaria. Para comprobar lo expuesto por la Figura 4.16 se emplea la prueba de Dickey Fuller aumentada obteniendo como resultado un valor de $p = 0{,}5686$, y se confirma que se trata de una serie estacionaria.

Análisis de la tendencia

La serie de potencia reactiva no presenta una tendencia significativa por lo que no requiere de un proceso de diferenciación como se explicó en la sección 2.4.6.

Análisis ACF y PACF

En este paso se analiza el orden del modelo que se encuentra en construcción. Se grafican los valores de ACF y PACF de la serie estacionaria en la Figura 4.17.

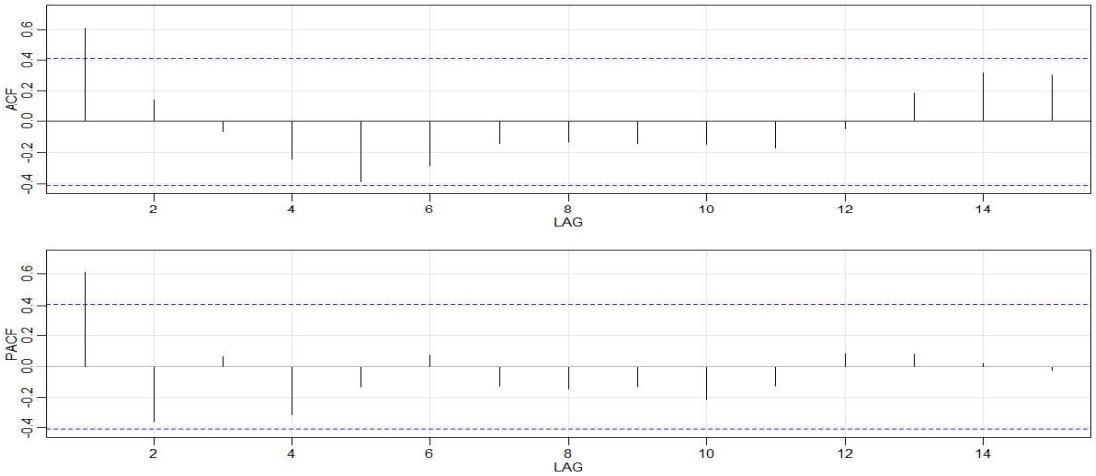

Figura 4.17. Valores de ACF y PACF para serie de potencia reactiva anual estacionaria.

Los valores de la función de autocorrelación decrecen y ningún rezago presenta un valor significante, mientras que el correlograma con los valores de autocorrelación parcial muestra que luego del segundo rezago no existen rezagos que sobrepasan el intervalo de confianza. Se propone un modelo de media móvil AR(p) donde p es igual a dos.

Tabla 4.15. Observaciones de los valores de ACF y PACF de la serie estacionaria.

	ARMA (0,2)
ACF	Decrece
PACF	Los valores son insignificantes a partir del rezago 2

La Tabla 4.15 resume los resultados obtenidos de la observación de los correlogramas de autocorrelación y autocorrelación parcial con los cuales se propone el modelo ARIMA$(2,0,0)$.

Tabla 4.16. Modelo propuesto para datos anuales de potencia reactiva

Modelo	AIC	RMSE	MAE	MAPE (%)
ARIMA$(2,0,0)$	180,03	8,4766	6,9748	4,9137

La Tabla 4.16 muestra varios indicadores del modelo planteado, el valor MAPE se mantiene por debajo del 15% aceptable [46].

Análisis de residuos

Una vez definido el modelo a emplearse, se debe realizar distintas pruebas para comprobar que éste capture la mayor cantidad de información relevante de los datos de entrenamiento.

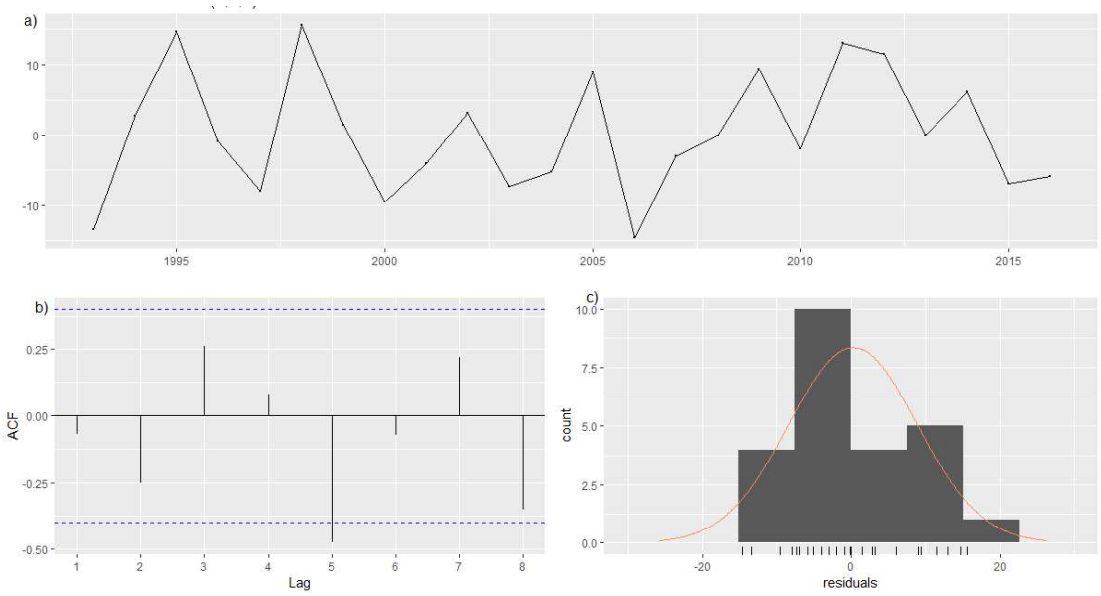

Figura 4.18. Prueba a los residuos para el modelo ARIMA$(2,0,0)$ de potencia reactiva.

La Figura 4.18 a) muestra la media obtenida de los residuos del modelo ARIMA$(2,0,0)$, que se mantiene alrededor de cero de acuerdo con lo mencionado en la sección 2.5.3. La Figura 4.18 b) es el correlograma de los valores de autocorrelación obtenida de los residuos del modelo, los valores se acercan a los obtenidos por un proceso de ruido blanco. Finalmente, la Figura 4.18 c) muestra que los residuos del modelo presentan una distribución normal.

Cada una de las distintas graficas de la Figura 4.18 aporta a la conclusión de que el modelo ARIMA$(2,0,0)$, ha capturado la mayor cantidad de información presentada en la serie de tiempo de demanda de potencia reactiva.

Sin embargo, se presenta dos pruebas más para desechar la hipótesis de correlación entre residuos del modelo. Los resultados obtenidos son escritos en la Tabla 4.17.

Tabla 4.17. Pruebas a residuos para el modelo ARIMA$(2,0,0)$ de potencia reactiva.

Prueba	**Valor de *p***
Ljung-Box	0,4161
Dickey Fuller aumentada	0,00032
Durbin Watson **(Valor de *d*)**	2,0031

El modelo planteado en la prueba de Ljung-Box obtiene un valor de *p* mayor a 0.05 mostrando la independencia de los residuos [45] y en la prueba de Dickey Fuller aumentada muestra un valor mínimo de *p*. Finalmente la prueba de Durbin Watson obtiene un valor de *d* muy cercano a dos, lo que permite concluir que la correlación entre residuos es mínima.

4.3.3 PRONÓSTICO

Luego de haber construido el modelo mediante la observación y aplicación de pruebas se realiza el pronóstico del modelo ARIMA$(2,0,0)$, obtenido con los datos anuales de potencia reactiva. Los resultados se muestran en la Tabla 4.18.

Tabla 4.18. Pronóstico de potencia reactiva, mediante el modelo ARIMA$(2,0,0)$.

Año	**Potencia reactiva [MVAR]**
2017	125,511
2018	132,626
2019	140,057
2020	144,365
2021	145,145
2022	143,749

2023	141,840
2024	140,510
2025	140,069
2026	140,285

La serie de potencia reactiva no presenta un patrón definido y el conjunto de datos de entrenamiento es pequeño por lo que el uso de modelos ARIMA no es recomendado.

La Figura 4.19 reúne los datos históricos de potencia reactiva que sirvieron como datos de entrenamiento (negro), los valores pronosticados desde el año 2017 al año 2026 se encuentran en color azul y el nivel de confianza para el intervalo de predicción se dispuso al 95%.

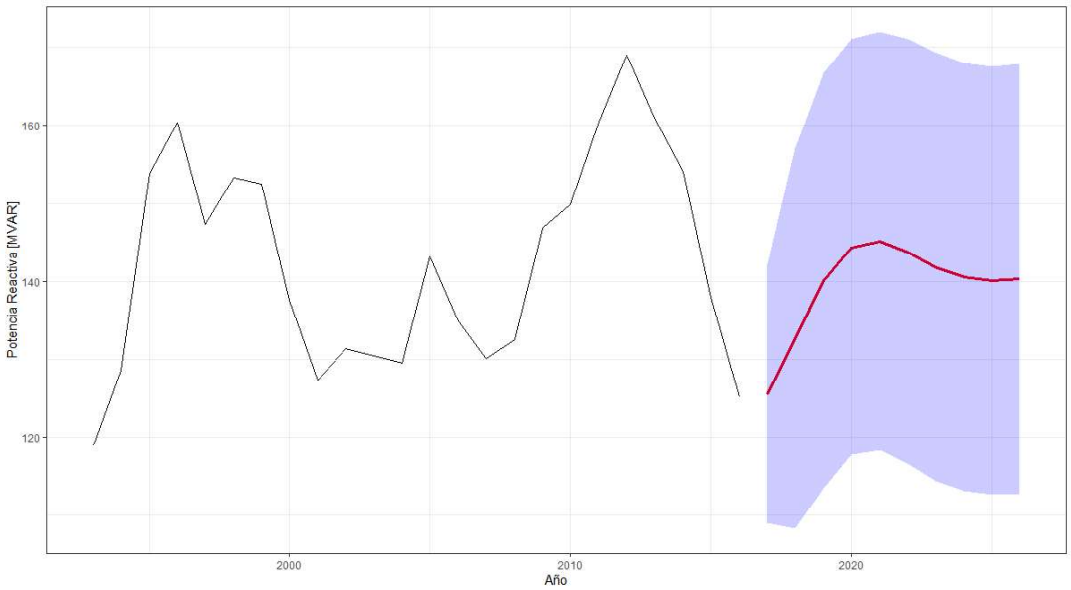

Figura 4.19. Pronóstico anual de potencia reactiva.

4.3.4 ANÁLISIS DE RESULTADOS DEL PRONÓSTICO ANUAL DE POTENCIA REACTIVA

La Tabla 4.19 muestra los valores de pronóstico obtenidos mediante la construcción del modelo ARIMA en lenguaje R y los resultados obtenidos con la metodología actual de la EEQ. Los valores se mantienen bastante cercanos, pero con el paso de los años el pronóstico de la EEQ es mayor al pronóstico de realizado en lenguaje R.

Tabla 4.19. Pronóstico anual de potencia reactiva obtenido en R y por la EEQ

Año	Potencia reactiva en RStudio [MVAR]	Potencia reactiva EEQ [MVAR]
2017	125,511	131,145
2018	132,626	136,793
2019	140,057	137,516
2020	144,365	140,816
2021	145,145	144,196
2022	143,749	147,657
2023	141,840	151,200
2024	140,510	154,829
2025	140,069	158,545
2026	140,285	162,350

La Figura 4.20 muestra una comparativa grafica entre los datos históricos de potencia reactiva anual de la EEQ de color rojo y los valores obtenidos mediante el modelo ARIMA$(2,0,0)$ que se encuentran de color azul.

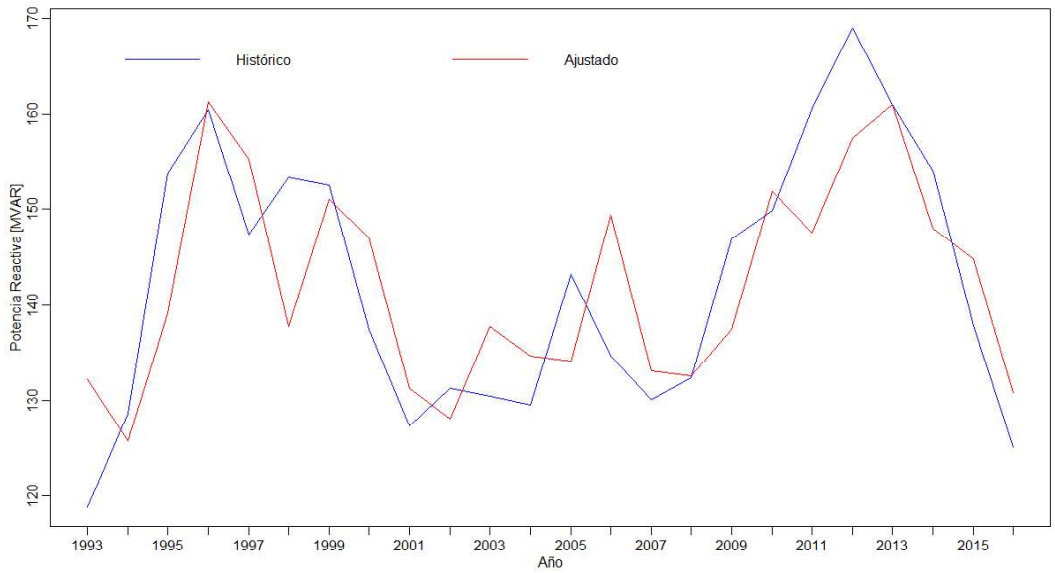

Figura 4.20. Datos históricos y ajustados de potencia reactiva anual de la EEQ, modelados ARIMA$(2,0,0)$

Mediante la Ecuación 4.3 se puede escribir el modelo ARIMA$(2,0,0)$.

$$y_t = c + 1{,}0767y_{t-1} - 0{,}5191y_{t-2} + \varepsilon_t$$

El modelo matemático muestra que el valor pronosticado depende únicamente del valor de los dos últimos errores.

4.4 FACTOR DE POTENCIA

Al realizar el pronóstico de la potencia activa y reactiva anual fácilmente se puede calcular el factor de potencia como se menciona en la sección 2.1 y comprobar que no supere los límites establecidos por la Regulación No. CONELEC 004/01, donde se estipula que el valor de factor de potencia debe tener un mínimo 0,92 [47].

Los resultados obtenidos en RStudio son exportados en un archivo de texto para su manejo. De esta manera los resultados son colocados en tablas para su posterior análisis, como se muestra en la Tabla 4.20. El cálculo del factor de potencia se obtiene mediante la Ecuación 2.4.

Tabla 4.20. Pronóstico de potencia activa, reactiva y cálculo de factor de potencia

Año	Potencia activa [MW]	Potencia reactiva [MVAR]	Factor de potencia
2017	761,522	125,511	0,9862
2018	777,654	132,626	0,9852
2019	793,786	140,058	0,9841
2020	809,918	144,365	0,9838
2021	826,050	145,145	0,9843
2022	842,182	143,749	0,9852
2023	858,314	141,841	0,9861

2024	874,446	140,511	0,9869
2025	890,578	140,070	0,9875
2026	906,710	140,285	0,9879

La Tabla 4.20 muestra los valores de factor de potencia calculados a partir de los resultados del pronóstico de potencia activa anual con el modelo ARIMA$(0,2,1)$ y potencia reactiva anual con el modelo ARIMA$(2,0,0)$. Estos valores de factor de potencia se encuentran acorde a la regulación anteriormente mencionada.

Tabla 4.21. Factor de potencia obtenido en R y por la EEQ.

Año	Factor de potencia pronosticado en R	Factor de potencia pronosticado por la EEQ
2017	0,9862	0,9851
2018	0,9852	0,9847
2019	0,9841	0,9853
2020	0,9838	0,9853
2021	0,9843	0,9853
2022	0,9852	0,9854
2023	0,9861	0,9854
2024	0,9869	0,9854
2025	0,9875	0,9854
2026	0,9879	0,9854

La Tabla 4.21 muestra los valores de factor de potencia obtenidos mediante la construcción y pronóstico del modelo en lenguaje R y la metodología empleada por la EEQ, los dos resultados muestran valores que cumplen con la Regulación No. CONELEC 004/01, dando desde un punto de vista eléctrico validez a sus resultados.

4.5 PRONÓSTICO ANUAL DE POTENCIA ACTIVA Y ENERGÍA A TRAVÉS DEL PRONÓSTICO MENSUAL

Los valores de potencia activa anual también pueden ser obtenidos mediante el pronóstico mensual de potencia activa seleccionando el valor de potencia máxima de ese año. Mientras que los valores de energía anual se obtienen mediante la suma de consumo mensual de todo un año. Los resultados obtenidos se muestran en la Tabla 4.22.

Tabla 4.22. Pronóstico de potencia activa y energía a partir del pronóstico mensual

Año	Potencia activa [MW]	Energía [MWh]
2017	761,844	4’185.613,644
2018	776,097	4’302.483,904
2019	789,892	4’421.821,371
2020	804,259	4’533.185,285
2021	819,475	4’640.188,065
2022	835,017	4’744.817,510
2023	850,983	4’848.155,381
2024	867,333	4’950.790,377
2025	883,975	5’053.042,867
2026	900,889	5’155.087,197

Estos resultados difieren de los obtenidos mediante la serie anual de potencia activa debido a que las dos series son diferentes y sus componentes (estacionalidad, tendencia y componente irregular) también son diferentes. Además, los modelos que se construyeron están en función de los datos de entrenamiento y si estos son diferentes, el modelo construido y su pronóstico por ende será diferente.

4.6 FACTOR DE CARGA

De igual manera mediante el pronóstico anual de la potencia activa y la energía consumida, se calcula los valores de factor de carga y para su validación se calcula su error porcentual absoluto medio.

Tabla 4.23. Pronóstico de potencia activa, energía y cálculo de factor de carga

Año	Potencia activa [MW]	Energía [MWh]	Factor de carga [%]
2017	761,522	4'246.029,270	63,649
2018	777,654	4'345.339,678	63,787
2019	793,786	4'444.650,087	63,919
2020	809,918	4'543.960,495	64,046
2021	826,050	4'643.270,904	64,167
2022	842,182	4'742.581,312	64,284
2023	858,314	4'841.891,720	64,397
2024	874,446	4'941.202,129	64,505
2025	890,578	5'040.512,537	64,609
2026	906,710	5'139.822,946	64,711

En la Tabla 4.23 se reúnen los resultados obtenidos del pronóstico anual de potencia activa y energía total consumida por la carga de la Empresa Eléctrica Quito a excepción del Metro de Quito debido a que este trabajo se basa en datos históricos.

Tabla 4.24. Pronóstico anual de factor de carga y MAPE

Año	Factor de carga real [%]	Factor de carga calculado [%]	factor de carga pronosticado por la EEQ [%]
2017	64,29	63,65	65,15
MAPE [%]		0,99	2,87

La Tabla 4.24 contiene la evaluación del factor de carga obtenido mediante el criterio del error porcentual absoluto medio. El dato de prueba es del año 2017 con el que se consigue un MAPE de 0,99% mediante los modelos ARIMA implementados el lenguaje R, mientras que la metodología actual de la Empresa Eléctrica Quito obtiene un error porcentual absoluto medio mayor de 2,87%.

La Tabla 4.25 muestra los resultados obtenidos de factor de carga a partir de los modelos de potencia activa y energía desarrollados mediante los modelos ARIMA y la metodología de la EEQ. Los valores mediante los modelos desarrollados en lengua R son menores al valor calculado a partir del pronóstico desarrollado por la EEQ.

Tabla 4.25. Pronóstico anual de factor de carga obtenido en R y por la EEQ

Año	Factor de carga pronosticado en R [%]	Factor de carga pronosticado por la EEQ [%]
2017	63,649	65,150
2018	63,787	65,617
2019	63,919	65,943
2020	64,046	66,269
2021	64,167	66,597
2022	64,284	66,926
2023	64,397	67,256
2024	64,505	67,588
2025	64,609	67,922
2026	64,711	68,258

4.7 ANÁLISIS DE METODOLOGÍAS DE PRONÓSTICO

A continuación, se realiza un breve análisis comparativo entre la metodología empleada por la EEQ y la expuesta en este trabajo.

- La Empresa Eléctrica Quito emplea modelos de regresión polinomial para sus pronósticos, esta metodología presenta una desventaja en los modelos mensuales, debido a que requiere 12 modelos diferentes, es decir un modelo por cada mes y la metodología expuesta en este trabajo emplea un único modelo para una serie de tiempo.
- La preparación de los datos históricos en cualquier modelo es ineludible, sin embargo, los modelos ARIMA requieren de una mayor minuciosidad debido a que dependerá de esto el detectar o no una estacionalidad.
- Los modelos ARIMA y SARIMA al modelar series de tiempo que no presenten un patrón optan por realizar el pronóstico con el último valor real, es decir que el pronóstico se mantendrá constante e igual al último valor real, esto ocurre en el pronóstico anual de potencia reactiva anual.
- Los modelos ARIMA dependen únicamente de sus datos de aprendizaje y por eso la etapa de minería de datos se convierte en un requisito. El plantear un modelo mediante el análisis grafico de las correlaciones requiere de experticia y cada año debe ser evaluado nuevamente.

4.8 ANÁLISIS COMPARATIVO TENDENCIAL

4.8.1 AGENCIA DE REGULACIÓN Y CONTROL DE ELECTRICIDAD

La Agencia de Regulación y Control de Electricidad (ARCONEL) es el organismo regulador y de monitoreo para la operación y el desarrollo del sector eléctrico en Ecuador. ARCONEL es la entidad encargada de emitir regulaciones para las compañías eléctricas, realizar estudios y análisis técnicos, promover iniciativas de protección ambiental en el sector e imponer sanciones de suspensión [48].

Energía

El área de concesión de la Empresa Eléctrica Quito presenta una gran carga para el Sistema Nacional Interconectado (SNI), esto se ve reflejado en el comportamiento tendencial entre el requerimiento anual total de energía pronosticado para todo Ecuador por ARCONEL en su plan maestro de electricidad 2016-2025 y el pronóstico de energía para la EEQ obtenido mediante el modelo ARIMA $(0,2,1)$ desarrollado en RStudio mostrado en la Figura 4.21.

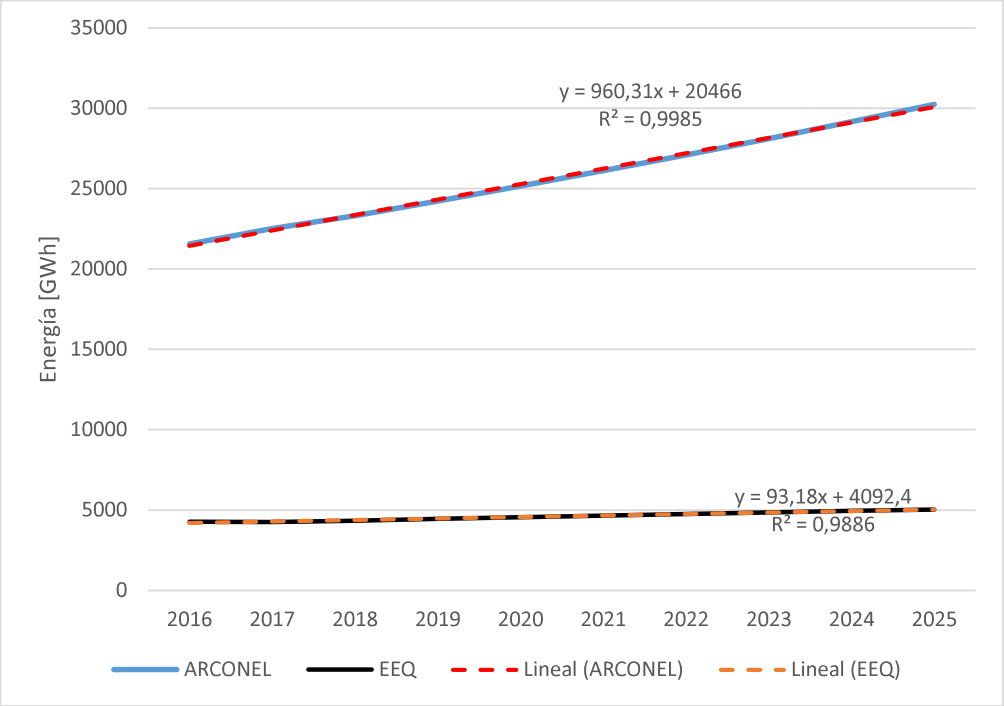

Figura 4.21. Tendencia del pronóstico de energía mediante el modelo ARIMA $(0,2,1)$ para la EEQ y el pronóstico presentado por ARCONEL.

El análisis tendencial permite determinar si una serie es estacionaria o estacional, debido a la fuerte tendencia marcada en la Figura 4.21. se trata de una serie no estacionaria.

Tabla 4.26. Coeficiente de Pearson para pronóstico de energía mediante el modelo ARIMA $(0,2,1)$ para la EEQ y el pronóstico presentado por ARCONEL.

	ARCONEL	**EEQ**
Coeficiente de Pearson	0,99	0,99

La Tabla 4.26 muestra que los coeficientes de Pearson obtenidos mediante la Ecuación 2.10 del pronóstico presentado por ARCONEL para Ecuador en su Plan Maestro de Electricidad y el pronóstico obtenido mediante los modelos ARIMA $(0,2,1)$ para la EEQ presentan una perfecta correlación positiva, dando por resultado una tendencia creciente.

La Figura 4.22 muestra el pronóstico obtenido en RStudio y el pronóstico obtenido por ARCONEL para energía de la Empresa Eléctrica Quito. Las dos series mantienen una tendencia creciente, sin embargo, el pronóstico desarrollado por ARCONEL obtiene valores superiores al desarrollado en RStudio.

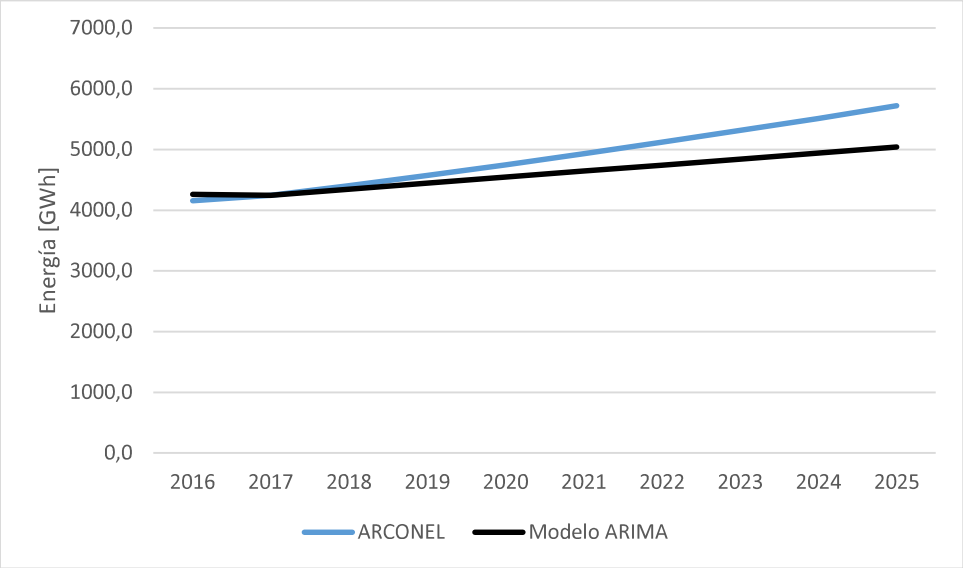

Figura 4.22. Pronóstico de energía para la EEQ mediante el modelo ARIMA $(0,2,1)$ y metodología de ARCONEL.

La Tabla 4.27 muestra que el pronóstico obtenido mediante el modelo ARIMA$(0,2,1)$ desarrollado en RStudio y la metodología del ARCONEL para la EEQ tienen una relación lineal perfecta y positiva. Los valores obtenidos por ARCONEL hasta el año 2018 son menores en un 2% al pronóstico obtenido mediante el modelo ARIMA y a partir de este año los valores del pronóstico obtenido por ARCONEL son mayores al obtenido por la EEQ llegando con una diferencia del 13% al año 2025, sin embargo, la tendencia es creciente para las dos series.

Tabla 4.27. Energía pronosticada para la EEQ mediante el modelo ARIMA $(0,2,1)$ y la metodología de ARCONEL.

	Energía pronosticada [GWh]	
Año	**Modelo ARIMA$(0,2,1)$**	**Metodología ARCONEL**
2016	4259,1	4154
2017	4246,0	4244
2018	4345,3	4399
2019	4444,7	4571
2020	4544,0	4748

2021	4643,3	4930
2022	4742,6	5118
2023	4841,9	5311
2024	4941,2	5510
2025	5040,5	5716
Pearson		0,99

Potencia Activa

La tendencia del pronóstico de potencia activa presentado por el ARCONEL para el SNI es creciente al igual que la tendencia que presenta el pronóstico de potencia activa para la EEQ obtenido mediante el modelo ARIMA$(0,2,1)$ como muestra la Figura 4.23.

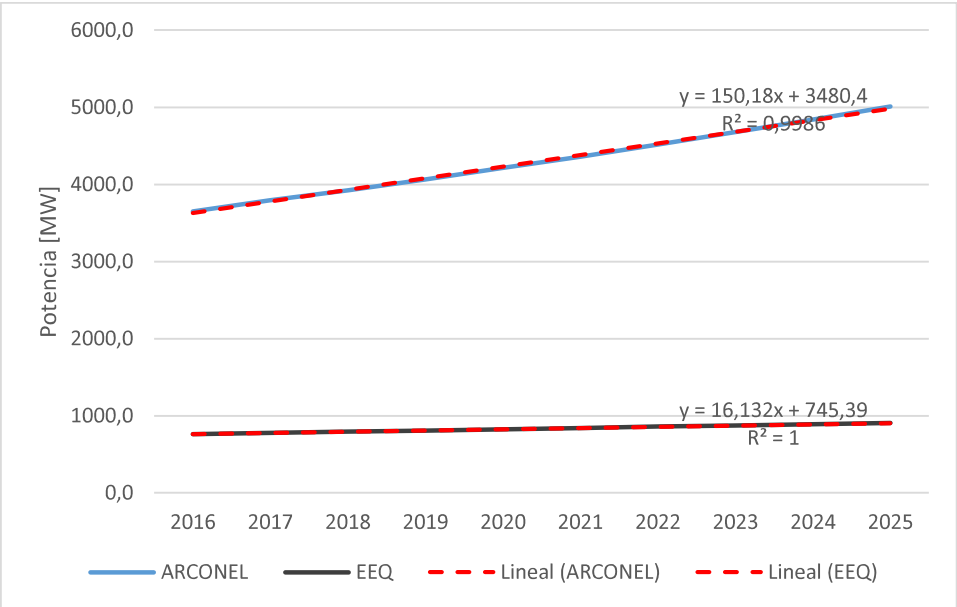

Figura 4.23. Tendencia del pronóstico de potencia activa mediante el modelo ARIMA $(0,2,1)$ para la EEQ y el pronóstico presentado por ARCONEL.

La Tabla 4.28 muestra que el coeficiente de Pearson obtenido por pronóstico presentado por el ARCONEL en su Plan Maestro de Electricidad y el pronóstico obtenido mediante el modelo ARIMA$(0,2,1)$ da como resultado una perfecta correlación positiva.

Tabla 4.28. Coeficiente de Pearson para el pronóstico de potencia activa mediante el modelo ARIMA $(0,2,1)$ para la EEQ y el pronóstico presentado por ARCONEL.

	ARCONEL	**EEQ**
Coeficiente de Pearson	0,99	0,99

La Figura 4.24 muestra la gráfica de los pronósticos obtenidos en RStudio y por ARCONEL para la serie de potencia activa anual de la Empresa Eléctrica Quito. Las dos series mantienen una tendencia creciente, sin embargo, a partir del año 2020 el pronóstico presentado por ARCONEL supera a los valores obtenidos mediante el modelo ARIMA$(0,2,1)$.

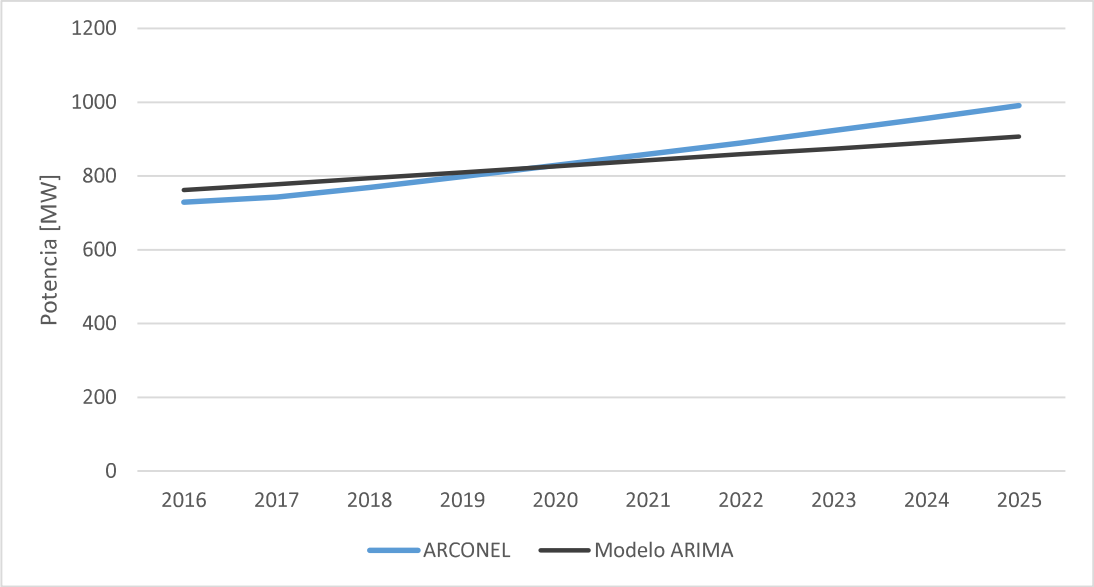

Figura 4.24. Pronóstico de potencia activa para la EEQ mediante el modelo ARIMA $(0,2,1)$ y metodología de ARCONEL.

La Tabla 4.29 muestra que el pronóstico obtenido mediante el modelo ARIMA$(0,2,1)$ y la metodología del ARCONEL para la EEQ tienen una relación lineal perfecta y positiva. Los valores obtenidos por ARCONEL como se muestran en la tabla desde el año 2020 superan a los valores del pronóstico obtenido mediante el modelo ARIMA en un 2% llegando con una diferencia del 9% al año 2025, sin embargo, el comportamiento de las dos series es una tendencia creciente.

Tabla 4.29. Potencia activa pronosticada mediante modelos ARIMA y la metodología de ARCONEL

	Potencia activa pronosticada [MW]	
Año	**Modelo ARIMA$(0,2,1)$**	**Metodología ARCONEL**
2016	761,522	729
2017	777,654	743
2018	793,786	769
2019	809,918	798
2020	826,05	828
2021	842,182	859
2022	858,314	890
2023	874,446	923
2024	890,578	956
2025	906,71	991
Pearson		0,99

4.8.2 CENTRO NACIONAL DE CONTROL DE ENERGÍA

El Centro Nacional de Control de Energía (CENACE) se constituye como el Operador Nacional de Electricidad en enero de 2015, el cual es un órgano técnico estratégico adscrito al Ministerio de Electricidad y Energía Renovable, que actúa como operador técnico del Sistema Nacional Interconectado y administrador comercial de las transacciones de bloques energéticos, responsable del abastecimiento continuo de energía eléctrica al mínimo costo posible, preservando la eficiencia global del sector [49].

Energía

La Empresa Eléctrica Quito junto con CNEL Guayaquil representan las cargas más fuertes para el Sistema Nacional Interconectado (SNI) de tal manera que el sistema es condicionado a su crecimiento, esto se ve reflejado en el comportamiento tendencial entre el requerimiento mensual total de energía pronosticado para todo Ecuador por CENACE en su plan maestro Bianual Octubre 2016 a Septiembre 2018 y el pronóstico de energía

mensual para la EEQ obtenido con el modelo SARIMA$(2,0,1)(0,1,2)_{12}$desarrollado en RStudio mostrado en Figura 4.25.

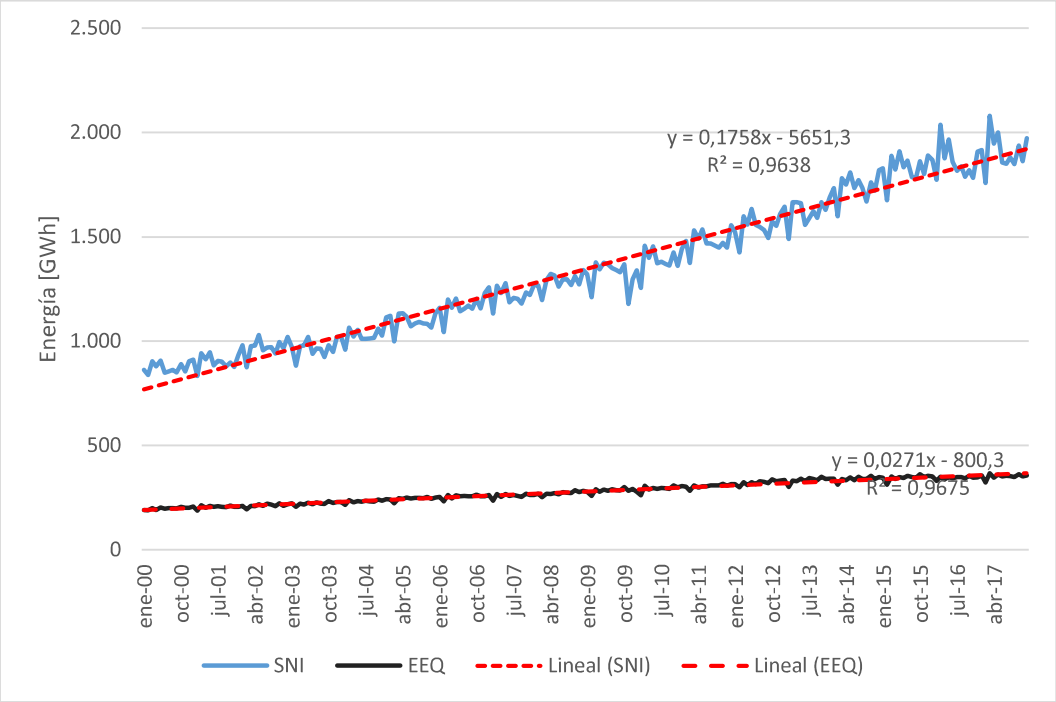

Figura 4.25. Comportamiento de la tendencia de energía de la EEQ y CENACE

La Figura 4.25 muestra el comportamiento histórico que ha tenido el despacho de energía realizado por CENACE y el exigido por la carga de la EEQ desde enero de 2000 a diciembre de 2017, los dos se encuentran al alza, sin embargo, es necesario corroborar esta información mediante el coeficiente de Pearson.

Tabla 4.30. Coeficiente de Pearson para datos históricos de energía de CENACE y EEQ.

	ARCONEL	**EEQ**
Coeficiente de Pearson	0,98	0,98

Los coeficientes de Pearson mostrados en la Tabla 4.30 al ser valores cercanos a 1 y positivos, dan como resultados una perfecta correlación positiva, indicando la existencia de una tendencia creciente.

La Figura 4.26 muestra la gráfica de los pronósticos obtenidos mediante el modelo SARIMA$(2,0,1)(0,1,2)_{12}$ y por CENACE para la Empresa Eléctrica Quito. Las dos series mantienen un comportamiento similar, sin embargo, el pronóstico desarrollado por

CENACE obtiene valores superiores al desarrollado en RStudio a excepción del mes de diciembre por lo que es necesario mediante el MAPE discernir el pronóstico que tenga un mejor ajuste a los datos reales de energía, esto se realizó mediante la Tabla 4.31.

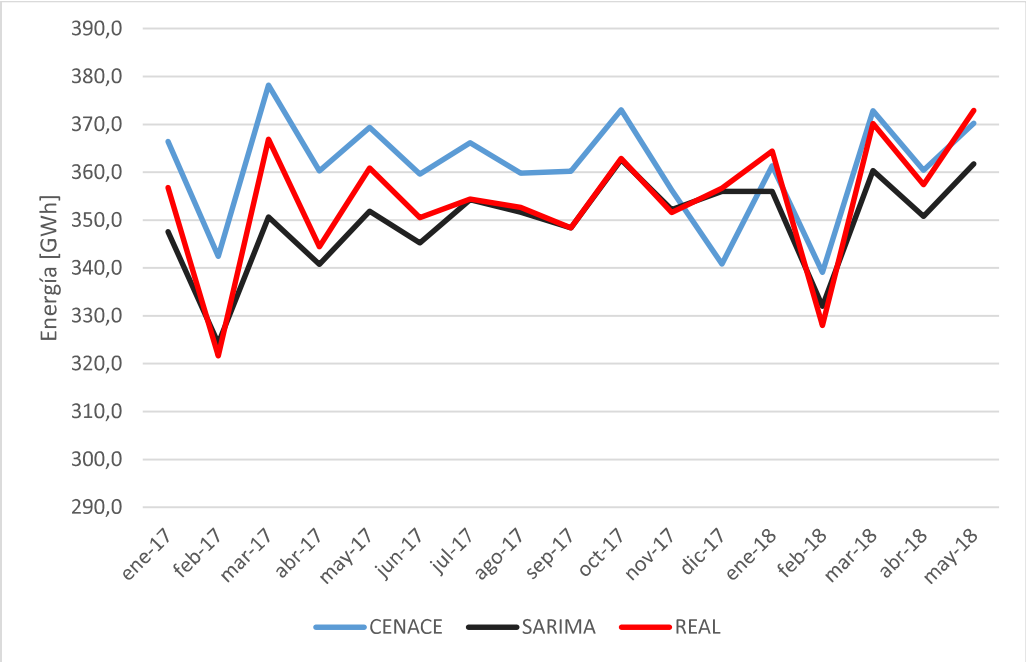

Figura 4.26. Pronóstico de energía para la EEQ desarrollado en RStudio y el presentado por CENACE.

La Tabla 4.31 muestra los valores del pronóstico de energía obtenidos mediante la metodología de CENACE y los obtenidos mediante el modelo SARIMA$(2,0,1)(0,1,2)_{12}$ implementado en RStudio. El MAPE obtenido mediante el modelo construido en RStudio es menor al obtenido con los resultados del CENACE, de esta manera se muestra que el modelo SARIMA$(2,0,1)(0,1,2)_{12}$ ha obtenido resultados más cercanos a los datos reales.

Tabla 4.31. Energía pronosticada mediante modelos SARIMA y la metodología de CENACE

		Energía pronosticada [GWh]	
Fecha	**Real**	**CENACE**	**SARIMA$(2,0,1)(0,1,2)_{12}$**
ene-17	356,8	366,4	347,5
feb-17	321,7	342,5	324,3
mar-17	366,9	378,2	350,6

abr-17	344,5	360,3	340,8
may-17	360,9	369,4	351,9
jun-17	350,5	359,7	345,2
jul-17	354,4	366,2	354,3
ago-17	352,7	359,8	351,7
sep-17	348,4	360,2	348,4
oct-17	362,9	373,1	362,7
nov-17	351,6	356,3	352,2
dic-17	356,7	340,8	356,0
ene-18	364,4	361,4	356,0
feb-18	328,0	339,1	332,0
mar-18	370,2	372,8	360,4
abr-18	357,4	360,5	350,7
may-18	372,9	370,2	361,8
MAPE (%)		2,68%	1,45%

Potencia

La Figura 4.27 muestra el comportamiento tendencial de los datos históricos desde enero de 2003 a diciembre de 2017 de potencia activa del CENACE y de la EEQ, la tendencia presentada de forma gráfica es creciente, sin embargo, se debe analizar el coeficiente de Pearson para cada serie histórica.

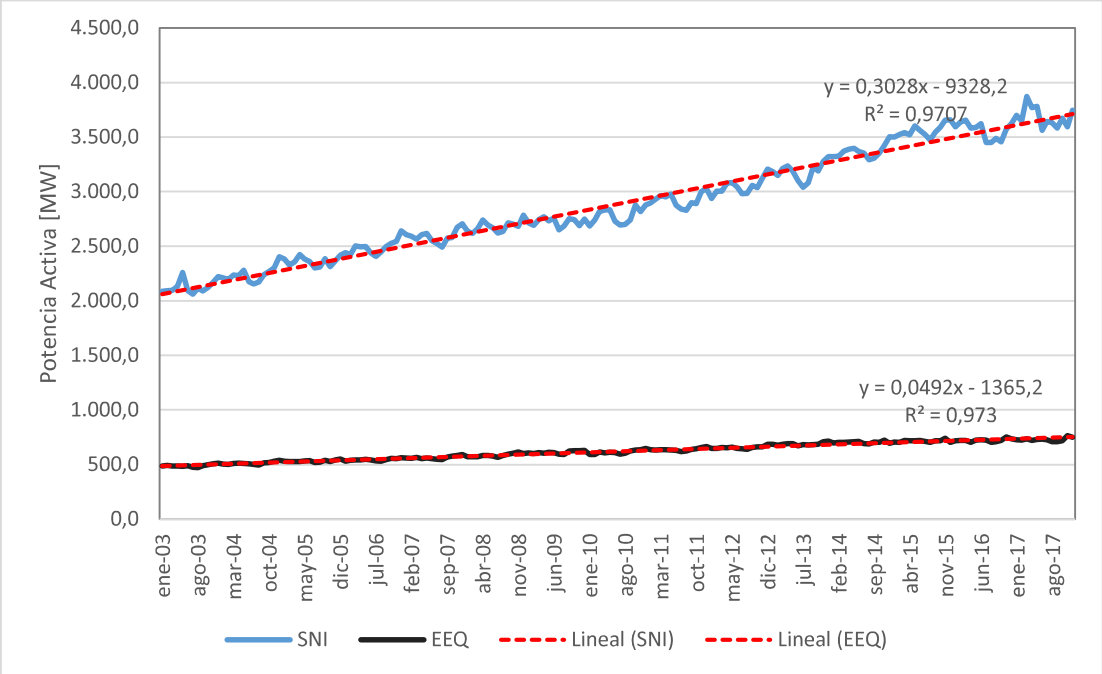

Figura 4.27. Comportamiento de la tendencia de potencia activa de la EEQ y CENACE

Los coeficientes de Pearson mostrados en la Tabla 4.32 fueron obtenidos a partir de los datos históricos de potencia activa del CENACE y EEQ, al ser valores cercanos a 1 y positivos, muestran que las dos series tienen una perfecta correlación positiva, indicando la existencia de una tendencia creciente.

Tabla 4.32. Coeficiente de Pearson para datos históricos de potencia activa de CENACE y EEQ.

	ARCONEL	**EEQ**
Coeficiente de Pearson	0,98	0,98

La Figura 4.28 muestra la gráfica de los pronósticos de potencia activa obtenidos mediante el modelo SARIMA$(2,1,1)(3,0,0)_{12}$ implementado en RStudio y por CENACE para la Empresa Eléctrica Quito. Las dos series mantienen un comportamiento similar, sin embargo, el pronóstico desarrollado por CENACE obtiene valores inferiores al modelo SARIMA$(2,1,1)(3,0,0)_{12}$. Se requiere del análisis mediante el MAPE para discernir el modelo que tenga un mejor ajuste al comportamiento de la EEQ.

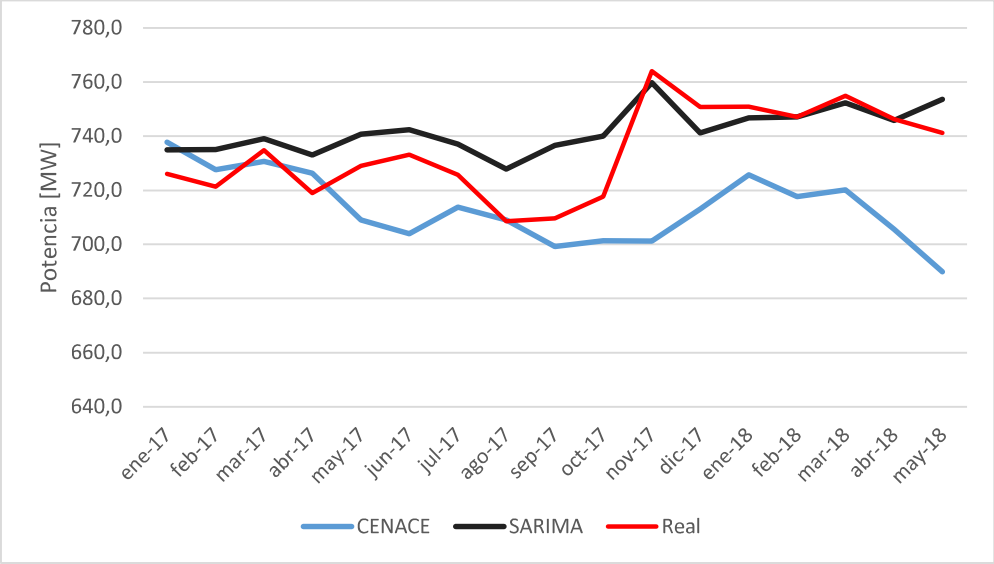

Figura 4.28. Pronóstico de potencia activa para la EEQ desarrollado en RStudio y el presentado por CENACE.

La Tabla 4.33 muestra los valores del pronóstico de potencia activa obtenidos mediante la metodología de CENACE y los obtenidos mediante el modelo SARIMA$(2,1,1)(3,0,0)_{12}$ implementado en RStudio. El MAPE obtenido mediante el modelo construido en RStudio es menor al obtenido con los resultados del CENACE, de esta manera se muestra que el modelo SARIMA$(2,1,1)(3,0,0)_{12}$ ha obtenido resultados más cercanos a los datos reales de potencia activa.

Tabla 4.33. Potencia activa pronosticada mediante modelos SARIMA y la metodología de CENACE

		Potencia pronosticada [MW]	
Fecha	**Real**	**CENACE**	**SARIMA$(2,1,1)(3,0,0)_{12}$**
ene-17	726,04	737,8	735,01
feb-17	721,25	727,6	735,05
mar-17	734,73	730,6	739,11
abr-17	719,00	726,3	733,13
may-17	729,00	709,0	740,76
jun-17	733,24	704,0	742,48

jul-17	725,79	713,7	737,12
ago-17	708,49	709,1	727,85
sep-17	709,60	699,3	736,49
oct-17	717,60	701,4	739,93
nov-17	764,06	701,3	759,84
dic-17	750,78	713,0	741,24
ene-18	750,93	725,7	746,78
feb-18	747,10	717,6	747,08
mar-18	754,88	720,2	752,32
abr-18	746,29	705,6	745,74
may-18	741,25	689,9	753,70
MAPE (%)		3,16	1,43

5. CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

- La minería de datos permite llevar una organización adecuada del procedimiento a seguir para obtener conocimiento a partir de bases de datos históricos.
- La metodología propuesta instituye un conjunto de pasos secuenciales que tienen como objetivo construir un modelo del comportamiento de la demanda eléctrica de la Empresa Eléctrica Quito. Se empleó los valores reales del 2017 y 2018 para mostrar la eficiencia de los modelos a través de los niveles de error (MAPE) que fueron menores al 2,7%. De esta manera los modelos obtenidos se establecen como modelos adecuados para la previsión de demanda.
- Los residuos de los modelos seleccionados al ser analizados mostraron que no existen correlaciones entre ellos, su media tiende a cero y presentan una distribución normal, lo que implica que el pronóstico contiene toda la información relevante construida a partir de los datos de entrenamiento.
- Los modelos ARIMA y SARIMA como se pudo observar en este trabajo, se constituyen como una adecuada herramienta de pronóstico para las empresas distribuidoras de energía, debido a que estos modelos reúnen características propias del consumo de energía, como lo son el crecimiento de la demanda año tras año (tendencia) y la repetición de patrones en ciertas estaciones del año (estacionalidad).
- Los modelos anuales presentan mayor simplicidad debido a que son valores máximos o absolutos de todo un año, lo que provoca que año tras año vayan aumentado y esto se traduzca en una tendencia lineal la cual es fácilmente modelada por los modelos ARIMA, esto se ve afianzado con el error porcentual absoluto medio (MAPE) que fue menor en comparación con los modelos conseguidos por la metodología actual de la Empresa Eléctrica Quito.
- Los modelos mensuales presentan mayor complejidad debido a que reúnen características estacionales y no estacionales, las cuales son modeladas a través de los modelos SARIMA$(p, d, q)(P, D, Q)_m$, al ser contrastados con la metodología actual empleada por la Empresa Eléctrica Quito presentan un MAPE menor y presentan la ventaja de utilizar un único modelo para toda la serie de datos.

- El modelo de energía mensual se presenta como una opción fuerte para los estudios realizados por la EEQ debido al pronóstico cercano que ha obtenido los años 2017 y 2018. Cabe recalcar que los modelos anuales han obtenido resultados que permiten corroborar su validez a través de variables eléctricas como lo son el factor de potencia y factor de carga manteniéndose dentro de los rangos adecuados.

- El análisis de tendencias concluye que la EEQ sigue el comportamiento del Sistema Nacional Interconectado, esto se basa en el resultado obtenido del coeficiente de Pearson donde la correlación del SNI es perfecta y positiva.

- El RMSE obtenido mediante el modelo SARIMA$(2,0,1)(0,1,2)_{12}$ para energía mensual fue contrastado con el RMSE del modelo seleccionado por el toolbox *Regression Learner* de Matlab, dando como resultado una desviación menor del modelo SARIMA.

5.2 RECOMENDACIONES

- El proceso manual de modelamiento depende de la experticia del predictor o usuario, mientras que la selección del modelo mediante el paquete *forecast* de R lo realiza a través de varias pruebas matemáticas, lo que genera un grado de confianza. Se recomienda realizar los dos procesos y contrastarlos.

- Mediante los diversos análisis realizados a los modelos y a los pronósticos resultados de ellos, se pone a consideración este trabajo como una herramienta que cumple todos los objetivos planteados, y que fácilmente puede ser implementada en el sector energético.

- El lenguaje R a través de su consola RStudio es un instrumento gratuito y práctico para cálculos estadísticos que provee de varias herramientas que permiten el análisis de la base de datos a utilizar. Este entorno ofrece diferentes paquetes con funciones que permiten realizar la minería de datos de forma íntegra, iniciando desde el proceso de preparación de datos hasta la implementación.

- El error porcentual absoluto medio (MAPE) fue el índice de error empleado para la evaluación de los modelos construidos en lenguaje R debido a su gran aplicación en diferentes casos de estudio a nivel mundial, sin embargo, este índice no es el único que puede ser empleado para la evaluación de resultados por ejemplo se puede emplear el RMSE, MAE, etc. Se invita a desarrollar un trabajo de

investigación donde se determine el índice que se acople de mejor manera a la situación del país.

- Se invita al lector a realizar la selección del modelo mediante los criterios explicados en la sección 2.6.2 y comprobarlos con la función *auto.arima* de esta manera se adquiere mayor experticia en el proceso de selección.

6. REFERENCIAS BIBLIOGRÁFICAS

[1] E. A. Feinberg and D. Genethliou, “Load Forecasting BT - Applied Mathematics for Restructured Electric Power Systems: Optimization, Control, and Computational Intelligence,” J. H. Chow, F. F. Wu, and J. Momoh, Eds. Boston, MA: Springer US, 2005, pp. 269–285.

[2] J. F. Tabares Muñoz, C. A. Velásquez Galvis, and M. Valencia Cárdenas, “Comparison of Statistical Forecasting Techniques for Electrical Energy Demand,” *Rev. Ing. Ind.*, vol. 13, no. 1, pp. 19–31, 2014.

[3] J. Casazza and F. Delea, *Understanding Electric Power Systems: An Overview of Technology, the Marketplace, and Government Regulation*. 2010.

[4] G. E. Harper, *Fundamentos de electricidad*, no. v. 3. Mexico: Grupo Noriega, 1994.

[5] C. L. Wadhwa, *Electrical Power Systems*, First edit. Tunbridge Wells: New Academic Science Limited, 2012.

[6] P. G. Antonio and O. J. Jesús, *CIRCUITOS ELÉCTRICOS. VOL. II*, Primera. Madrid: UNED, 2014.

[7] H. Beaty, H. Mahrous, and A. Seidman, *Handbook of electric power calculations*, Tercera ed. Chicago: McGraw Hill, 2001.

[8] T. Wildi, *Máquinas Eléctricas y Sistemas De Potencia*, Sexta edic. Mexico: Person Educación, 2007.

[9] V. M. Vélez M., R. A. Hincapíe I., and R. A. Gallego R., “Low voltage distribution system planning using diversified demand curves,” *Int. J. Electr. Power Energy Syst.*, vol. 61, pp. 691–700, 2014.

[10] S. Castaño, *Redes de Distribucion de Enetgia*, Tercera. Manizales: Univ. Nacional de Colombia, 2004.

[11] V. Venizelou, N. Philippou, M. Hadjipanayi, G. Makrides, V. Efthymiou, and

G. E. Georghiou, “Development of a novel time-of-use tariff algorithm for residential prosumer price-based demand side management,” *Energy*, vol. 142, pp. 633–646, 2018.

[12] C. C. Aggarwal, “An Introduction to Data Mining,” in *Data Mining: The Textbook*, Cham: Springer International Publishing, 2015, pp. 1–26.

[13] V. Kotu and B. Deshpande, *Predictive Analytics and Data Mining*. Waltham: Elsevier Inc, 2015.

[14] S. Džeroski, “Relational Data Mining BT - Data Mining and Knowledge Discovery Handbook,” O. Maimon and L. Rokach, Eds. Boston, MA: Springer US, 2010, pp. 887–911.

[15] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Elsevier Science, 2011.

[16] K. P. SOMAN, S. DIWAKAR, and V. AJAY, *DATA MINING: THEORY AND PRACTICE*. PHI Learning, 2006.

[17] F. Nava, *Procesamiento de series de tiempo*, 2nd ed. Mexico, 2013.

[18] J. Hanke and D. Wichern, *Pronósticos en los negocios*, 9a ed. Mexico: Pearson Educación, 2010.

[19] “Partial Correlation,” in *The Concise Encyclopedia of Statistics*, New York, NY: Springer New York, 2008, pp. 412–414.

[20] R. J. Hyndman and G. Athanasopoulos, *Forecasting : principles and practice*. 2015.

[21] P. Keat and Y. Philip, *Economía de Empresa*, Cuarta. Mexico: Pearson Educación, 2004.

[22] J. C. Lapponi, *Estatística usando Excel*. Elsevier Editora Lida., 2005.

[23] J. Dagnino, “Correlación,” *Rev. Chil. Anest.*, vol. 43, pp. 150–153, 2014.

[24] L. Marsh and D. R. Cormier, *Spline regression models*. Sage Publications, 2001.

[25] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to Statistical*

Learning, vol. 7, no. 10. 2000.

[26] W. Wei, *Time Series Analysis: Univariate and Multivariate Methods*, Second. Pearson Educación, 2006.

[27] M. Jesús, S. Naranjo, C. García-Martos Profesoras, and C. García-Martos, “Series Temporales Univariantes.”

[28] M. A. L. identificación del modelo a partir de la fac y la Facp, “Unidad 3: Modelos ARIMA,” *Curso de predicción económica y empresarial*, 2004. [Online]. Available: https://www.uam.es/docencia/predysim/prediccion_unidad4/4_3_doc3.pdf.

[29] H. Björn, “How you can use linear regression models to predict quadratic, root, and polynomial functions.” [Online]. Available: https://medium.freecodecamp.org/learn-how-to-improve-your-linear-models-8294bfa8a731. [Accessed: 20-Mar-2018].

[30] K. Methaprayoon, C. Yingvivatanapong, W. J. Lee, and J. R. Liao, “An integration of ANN wind power estimation into unit commitment considering the forecasting uncertainty,” *IEEE Trans. Ind. Appl.*, vol. 43, no. 6, pp. 1441–1448, 2007.

[31] E. H. Etuk, “Predicting Inflation Rates Of Nigeria Using A Seasonal Box-Jenkins Model,” vol. 1, no. 3, pp. 27–37, 2012.

[32] J. Jimenez Mares, K. Donado Mercado, and C. G. Quintero M., “A Methodology for Short-Term Load Forecasting,” *IEEE Lat. Am. Trans.*, vol. 15, no. 3, pp. 400–407, 2017.

[33] A. Azadeh, M. Saberi, and M. Anvari, “An integrated artificial neural network algorithm for performance assessment and optimization of decision making units,” *Expert Syst. Appl.*, vol. 37, no. 8, pp. 5688–5697, 2010.

[34] J. W. Cort and M. Kenji, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” *Clim. Res.*, vol. 30, no. 1, pp. 79–82, 2005.

[35] A. Azadeh, M. Saberi, S. F. Ghaderi, A. Gitiforouz, and V. Ebrahimipour,

"Improved estimation of electricity demand function by integration of fuzzy system and data mining approach," *Energy Convers. Manag.*, vol. 49, no. 8, pp. 2165–2177, 2008.

[36] J. D. Velasquez Henao, Y. Olaya Morales, and C. J. Franco Cardona, "Time Series Analysis and Forecasting in Energy Markets Using the R Language," *Dyna-Colombia*, vol. 78, no. 165, pp. 287–296, 2011.

[37] F. Fabozzi, S. Focardi, S. Rachev, and B. Arshanapalli, "Model Selection Criterion: AIC and BIC," in *The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications*, vol. 19, no. 1979, 2014, pp. 267–281.

[38] E. Erdogdu, "Electricity demand analysis using cointegration and ARIMA modelling: A case study of Turkey," *Energy Policy*, vol. 35, no. 2, pp. 1129–1146, 2007.

[39] J. M. Arranz and M. M. Zamora, "Análisis de Autocorrelación," 2015, p. 62.

[40] Empresa Eléctrica Quito, "Planificación institucional y técnica: misión, visión, Plan Estratégico, Plan Expansión - Empresa Eléctrica Quito," 2015. [Online]. Available: http://www.eeq.com.ec:8080/nosotros/planificacion;jsessionid=5E650E29C1C491B54463DF6F7A919DF5. [Accessed: 17-Dec-2018].

[41] F. Dobronski, *EL ECUADOR, LOS HECHOS MÁS IMPORTANTES DE SU HISTORIA, PARTE 1*, Segunda. Quito: Academia Nacional de Historia Militar del Ecuador, 2014.

[42] X. Cruz and F. Vera, "Desarrollo de un Modelo de Crecimiento en Ventas Para un Laboratorio Farmacèutico Dirigido a Clientes Importantes Aplicando un Modelo Estadìstico de Proyecciòn," Escuela Politecnica del Litoral, 2015.

[43] J. L. Beck and K. V. Yuen, "Bayesian Probabilistic Approach," *J. Eng. Mech.*, vol. 130. No. 2, no. February, pp. 192–203, 2004.

[44] D. J. Bartholomew, "No Title," *Oper. Res. Q.*, vol. 22, no. 2, pp. 199–201.

[45] D. Ömer Faruk, "A hybrid neural network and ARIMA model for water quality

time series prediction," *Eng. Appl. Artif. Intell.*, vol. 23, no. 4, pp. 586–594, 2010.

[46] R. Law and N. Au, "A neural network model to forecast Japanese demand for travel to Hong Kong," *Tour. Manag.*, vol. 20, no. 1, pp. 89–97, 1999.

[47] Agencia de Regulación y Control de Electricidad, *Regulación No.CONELEC 004/01*. 2001, pp. 1–25.

[48] ARCONEL, *RESOLUCIÓN No. ARCONEL 081/17*. Ecuador, 2017, p. 11.

[49] Corporación Nacional de Electricidad, "Quiénes Somos," 2014. [Online]. Available: http://www.cenace.org.ec/index.php?option=com_content&view=article&id=264&Itemid=53. [Accessed: 04-Nov-2018].

ANEXOS

ANEXO A. Datos de energía, potencia activa y reactiva

ANEXO B. Comparativa SARIMA y Machine Learning para energía mensual.

ANEXO C. Datos de energía, potencia activa y reactiva tabulados (Cd)

ANEXO D. Scripts para pronóstico en R (Cd)

ANEXO E. Video instructivo (Cd)

ANEXO A. DATOS DE ENERGÍA, POTENCIA ACTIVA Y REACTIVA

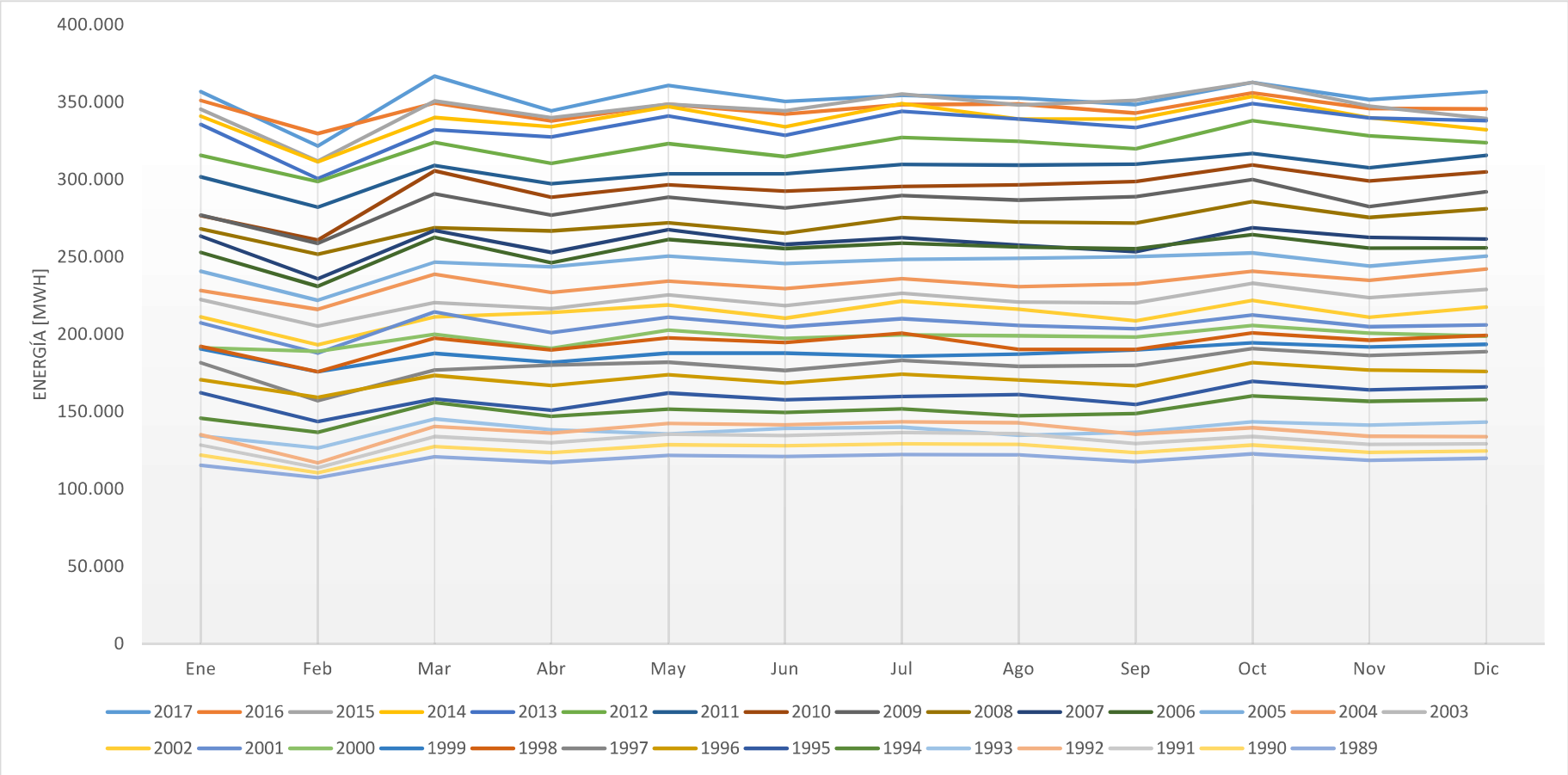

Figura A.1. Serie mensual de energía proporcionada por la EEQ

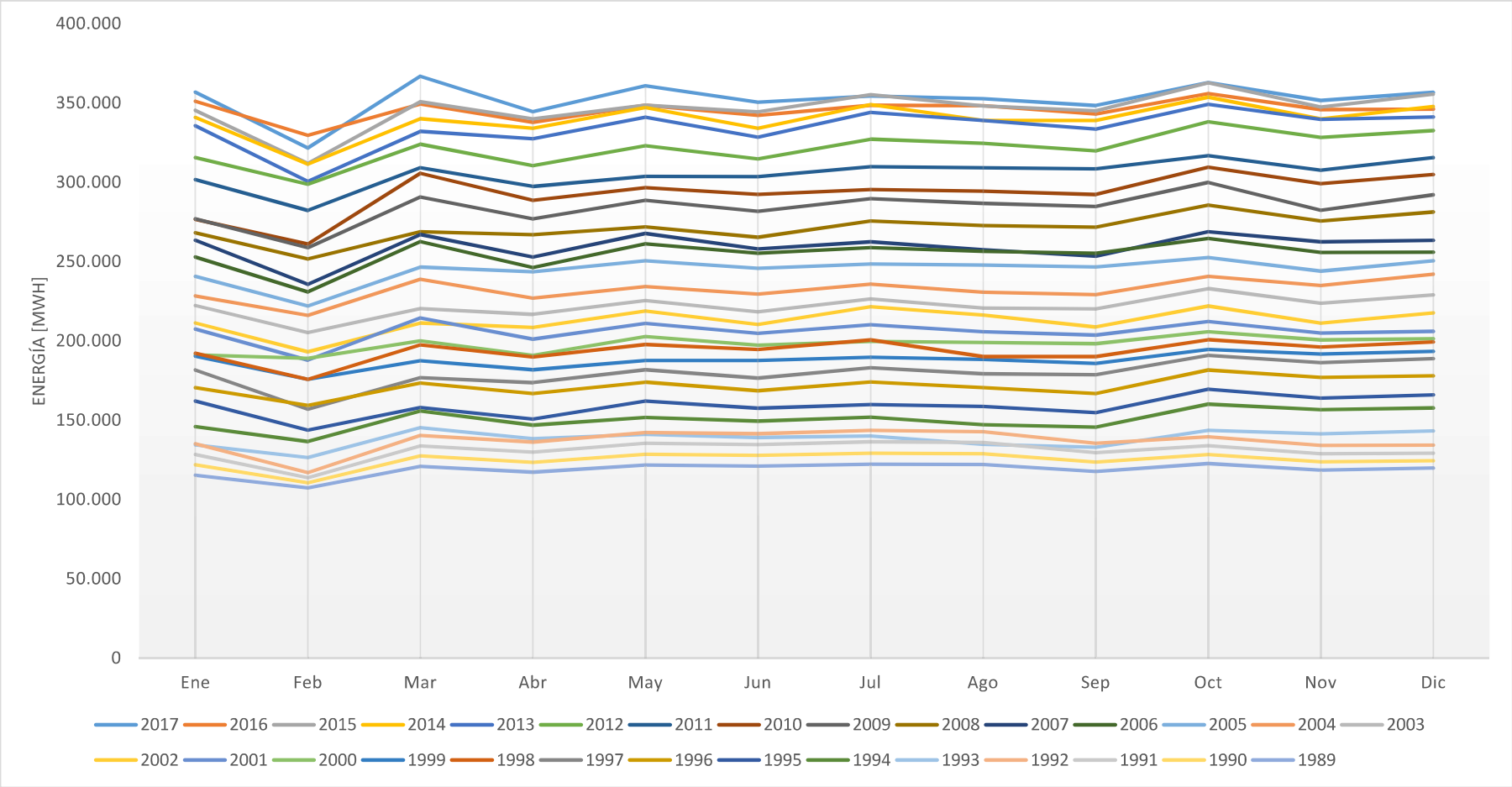

Figura A.2. Serie mensual de energía empleada para pronóstico

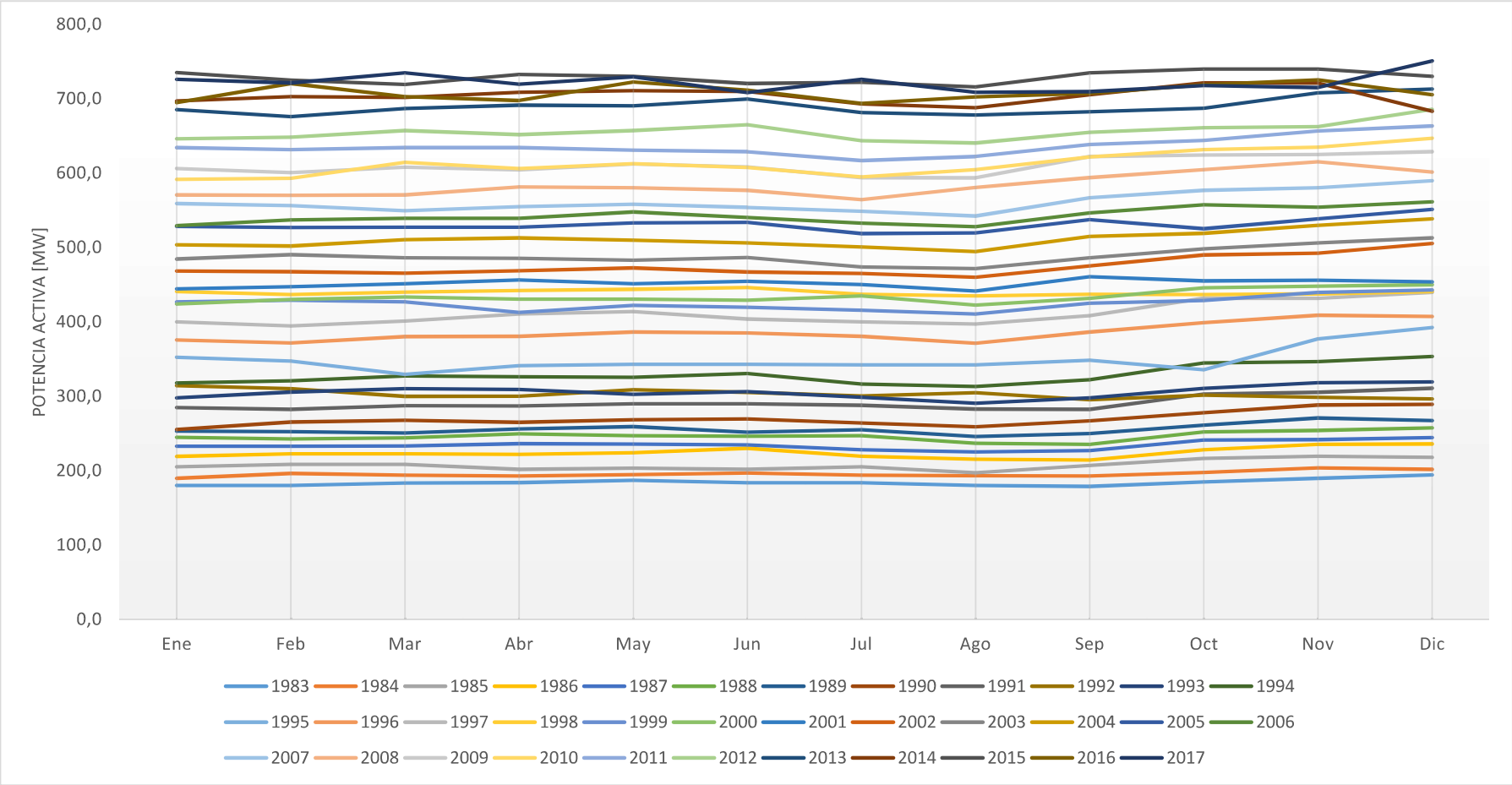

Figura A.3. Serie mensual de potencia activa proporcionada por la EEQ

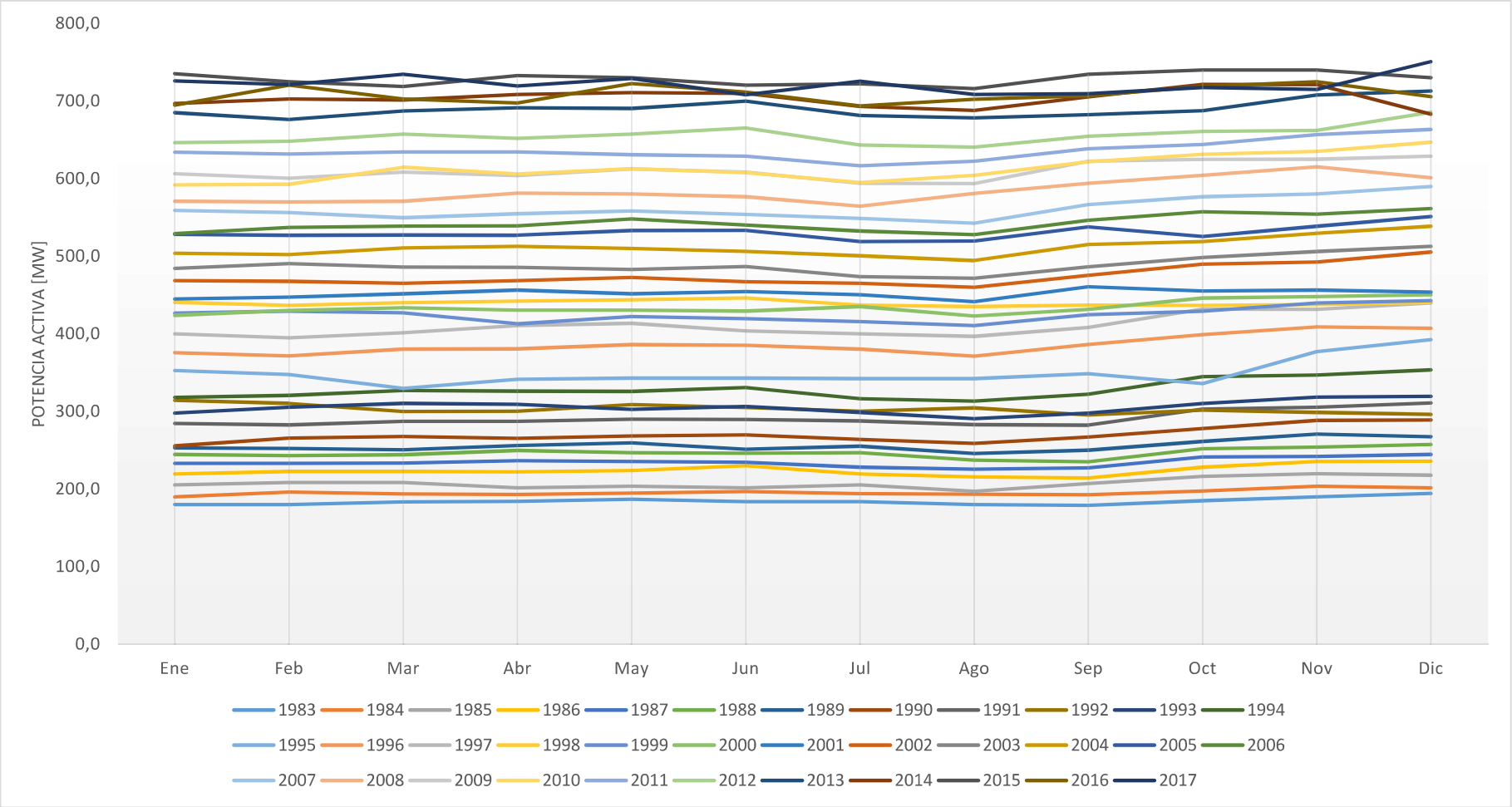

Figura A.4. Serie mensual de potencia activa empleada para pronóstico

ANEXO B. COMPARATIVA SARIMA Y MACHINE LEARNING PARA ENERGÍA MENSUAL.

Los pronósticos se basan en el modelo que mejor describa los datos reales, sin importar el grado de complejidad.

Mediante el toolbox *Regression Learner* de MATLAB se realizó la selección de un modelo regresivo para los datos de energía mensual desde enero de 1989 hasta diciembre de 2016. *Regression Learner* cuenta con diecinueve tipos de modelos los cuales fueron empleados, resaltando que con el modelo SARIMA propuesto se obtuvo un menor RMSE en comparación con el modelo de regresión de un proceso Gaussiano, tal como muestra la Tabla B.1.

Tabla B.1. Modelo de energía mensual a través de SARIMA y Machine Learning

Modelo	RMSE
Proceso de Regresión Gaussiano	10693
SARIMA$(2,0,1)(0,1,2)_{12}$	3634,5

La Figura B.1. muestra los datos de entrenamiento en color azul, mientras que los valores ajustados a través del modelo de regresión Gaussiana se encuentran en color amarillo. Para realizar la regresión los datos de entrenamiento se miran como una tendencia y se pierde su estacionalidad.

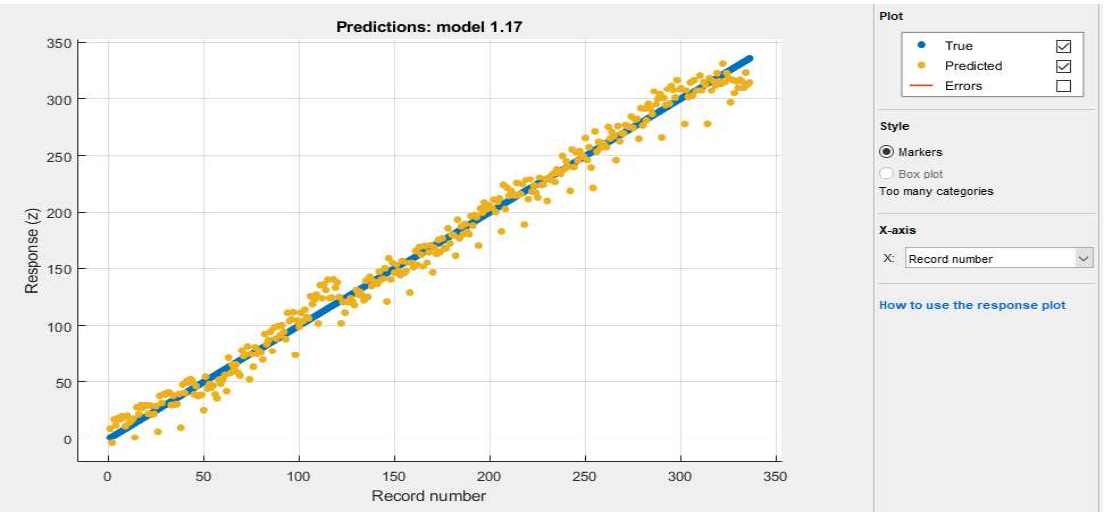

Figura B.1. Datos de energía mensual a través de Machine Learning.

ANEXO C. Datos de energía, potencia activa y reactiva tabulados (Cd)

ANEXO D. Scripts para pronóstico en R (Cd)

ANEXO E. Video instructivo (Cd)

ORDEN DE EMPASTADO