FUNDACIÓN UNIVERSITARIA LOS LIBERTADORES
FACULTAD DE INGENIERÍA Y CIENCIAS BÁSICAS

ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

MODELO SARIMA PARA EL PRONÓSTICO DEL NIVEL DEL RÍO MAGDALENA A LA ALTURA DEL MUNICIPIO DE BARRANCABERMEJA

ANDERSON FABIÁN SUESCÚN DÍAZ

INGENIERO QUÍMICO

Bogotá
2019

FUNDACIÓN UNIVERSITARIA LOS LIBERTADORES
FACULTAD DE INGENIERÍA Y CIENCIAS BÁSICAS
ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

ANDERSON FABIÁN SUESCÚN DÍAZ

MODELO SARIMA PARA EL PRONÓSTICO DEL NIVEL DEL RÍO MAGDALENA A LA ALTURA DEL MUNICIPIO DE BARRANCABERMEJA

ORIENTADOR: PROF. SÉBASTIEN LOZANO FORERO

Área de Concentración: Series de Tiempo

Trabajo presentado al programa de Especialización en Estadística Aplicada, de la Facultad de Ingeniería y Ciencias Básicas de la Fundación Universitaria los Libertadores para optar al título de **Especialista en Estadística Aplicada.**

Notas de aceptación

Firma del presidente del jurado

Firma del jurado

Firma del jurado

Bogotá D.C., Agosto de 2019

Las directivas de la Fundación Universitaria Los Libertadores, los jurados calificadores y el cuerpo docente no son responsables por los criterios e ideas expuestas en el presente documento. Estos corresponden únicamente a los autores y a los resultados de su trabajo.

Discipline, sooner or later, will overcome intelligence.
~ Japanese Proverb.

Agradecimientos

Con grato cariño, este nuevo logro en mi formación profesional es dedicado a mi madre Claudia Patricia, mi padre Juan, mi hermano Camilo y mi nona Gladys. Al profesor Sebastién Lozano, por su sencillez y constante apoyo en todo el proyecto. A la Fundación Universitaria Los Libertadores por la integral y excelente asistencia en mi formación.

Índice general

Índice de figuras

Índice de tablas

Resumen

El río Magdalena, siendo la principal ruta hídrica de Colombia, moviliza en promedio 1.4 millones de barriles de hidrocarburos al mes entre combustóleo, crudo y nafta, utilizando buquetanques de carga que abastecen desde la refinería de Barrancabermeja de Ecopetrol S.A. La logística de venta y transporte del combustóleo, asociada al medio de transporte ya sea por vía fluvial o terrestre, es netamente dependiente de los datos de navegabilidad, siendo los niveles de río a diferentes alturas de municipios rivereños como Barrancabermeja, los principales parámetros para escoger la logística de evacuación de este producto, siendo la opción fluvial la de mayor preferencia asociado a menores costos operativos. En el presente trabajo, se plantea a partir de un modelo en series de tiempo SARIMA , una herramienta de pronóstico no mayor a 3 semanas, que permita con un grado de certeza confiable pronosticar los niveles del río Magdalena a la altura del municipio de Barrancabermeja. Del presente trabajo, se obtuvo un modelo SARIMA $(2, 1; 2) \times (0; 1; 2)_{52}$, evaluado para las 12 primeras semanas del año 2019, donde la totalidad de los pronósticos se ubicaron en las bandas de confianza del 80%-95%, garantizando una fiabilidad adecuada para el modelo buscado.

CAPÍTULO 1

Introducción

Dentro del portafolio de productos de Ecopetrol S.A. (ECP), el combustóleo o Fuel Oil es un producto residual de los procesos de refinación del crudo, utilizado particularmente como combustible en hornos, secadores, rehervidores y como fuente de generación de potencia en barcazas y plantas generadoras de energía eléctrica. En la actualidad se producen 25 Kbpd de este producto que es evacuado por vía fluvial (60-70%) usando barcazas remolcadoras que se abastecen desde el muelle multimodal de la Refinería de Barrancabermeja (GRB); y terrestre (20-30%) utilizando camiones cisternas que abastecen desde los tanques e islas de llenado de la GRB, arribando aguas abajo al complejo portuario de la ciudad de Cartagena D.T.

La logística multimodal para la evacuación del combustóleo, esta priorizada hacia la evacuación fluvial, siendo esta opción la de menor costo operacional dado la imposibilidad estratégica impuesta por ECP para despachar vía poliducto. Bajo esta premisa, el estudio de las variables que controlan la navegabilidad del río Magdalena, como principal sistema multimodal para la evacuación del combustóleo, juega un papel de gran importancia en la logística de despacho, frecuencia de abastecimiento, tamaños de remolcadores, calado máximo, tiempos de carga y tasa de desocupación de los inventarios almacenados en GRB.

El río Magdalena, es el río más importante de Colombia, atravesando la región Andina, que es el centro del desarrollo de Colombia. Fluye de sur a norte con una longitud de alrededor de 1536 kilómetros y un caudal medio de 7100 m^3/s. El área de la cuenca es de alrededor de 257000 kilómetros cuadrados, lo que corresponde al 22.8% de la superficie total de

Colombia. La precipitación anual se estima en un 2000 mm con una variación en el interior de la cuenca de 800 mm a 5000 mm en algunas zonas.

El presente trabajo, propone una herramienta de pronóstico en series de tiempo, basada en un modelo SARIMA (2,1,1) x (0,1,2) para la predicción con un grado de certeza confiable a periodos no mayores a 3 semanas, el nivel del rio Magdalena a la altura del municipio de Barrancabermeja. El modelo está construido a partir de los históricos semanales del nivel del rio desde Enero de 2013 a Diciembre de 2018, recopilados por Cormagdalena. Este tipo de modelos de predicción son ampliamente utilizados para planeación estimada de volúmenes evacuados por el río, asociada a la relación directa que existe entre la profundidad del rio y el calado permitido para la carga de barcazas.

1.1. Objetivos

Establecido el alcance de este trabajo enfocado al modelo de pronóstico, se plantea dentro del proyecto los siguientes objetivos:

1.1.1. Objetivo General

- Contribuir a la planeación de la logística multimodal para la evacuación de combustóleo, a partir de un modelo en series de tiempo.

1.1.2. Objetivos Específicos

- Consolidar los reportes diarios del nivel del río Magdalena desde Enero de 2013 a Diciembre de 2019, en un histórico semanal representativo utilizando medidas de tendencia central.

- Obtener un modelo confiable de predicción en series de tiempo para el pronóstico no mayor a 3 semanas del nivel del río Magdalena a la altura del municipio de Barrancabermeja.

1.2. Organización del trabajo

El presente trabajo se organiza de la siguiente forma:

En el capítulo 2, se presentan los fundamentos teóricos para la construcción del modelo en series de tiempo, como son la metodología Box-Jenkins, modelos SARIMA y los criterios de información.

En el capítulo 3, se presenta la metodología para consolidar la base de datos diaria obtenida de Cormagdalena desde Enero de 2013 a Diciembre de 2018, en una nuevo histórico semanal utilizando la mediana como medida representativa para construir el modelo de pronóstico. También se explican los criterios de evaluación del modelo y su pronóstico.

En el capítulo 4, se presenta la consolidación de los datos del nuevo histórico semanal construido, basado en soportes visuales, graficas de tendencia, funciones de autocorrelación, autocorrelación parcial, pruebas de raíz unitaria y gráficas descriptivas para determinar la normalidad, homogeneidad, estacionariedad y estacionalidad de los datos.

En el capítulo 4.1, se presentan los resultados del ajuste de dos modelos univariados SARIMA, y la selección del modelo escogido a partir de los criterios de información de Akaike (AIC), Bayesiano (BIC) y Akaike modificado (AIC_c).

En el capítulo 4.2, se presentan los resultados de verificación de supuestos, pruebas de normalidad y autocorrelación serial de los residuos aplicados al modelo SARIMA escogido.

En el capítulo 4.3, se muestran los resultados del pronóstico de las 12 primeras semanas del año 2019 con su respectivo análisis de desviación respecto al reporte real emitido por Cormagdalena, para evaluar la efectividad del modelo.

Finalmente, en el capítulo 5 se enuncian las conclusiones de este trabajo y el planteamiento de las iniciativas de mejora integrando variables exógenas al modelo (SARIMAX).

1.3. Soporte computacional

El análisis estadístico y parte de las gráficas mostradas en el presente documento fueron realizadas en el ambiente de programación libre R en su versión 3.5.2 para la plataforma Windows, con ayuda del editor RStudio en su versión 1.1.463. Este lenguaje de programación fue creado por Ross Ihaka y Robert Gentleman en la Universidad de Auckland (Ihaka, 1996).

La elaboración de este documento, tablas y graficas restantes fueron hechas con el procesador de textos Microsoft Word y las hojas de cálculo de Microsoft Excel, que hacen parte de la suite ofimática de Microsoft Office.

CAPÍTULO 2

Marco Metodológico

El estudio presentado en este trabajo pretende proponer un modelo de predicción en series de tiempo para pronosticar el nivel del río Magdalena a la altura del municipio de Barrancabermeja. Ninguna relación de este tipo esta referenciada en la literatura consultada para este trabajo.

La base de datos histórica del nivel del rio Magdalena se obtuvo a través del portal web de Corporación Autónoma Regional del Río Grande de la Magdalena - CORMAGDALENA, consultada en abril de 2019. Los datos históricos corresponden a lo reportado entre enero de 2013 y diciembre de 2018.

Utilizando la metodología Box-Jenkins explicada en la Especialización en los módulos de aprendizaje: Series de Tiempo I y II, se determinó ajustar un modelo SARIMA (Auto Regresivo Integrado de Medias Móviles Estacional), debido al alcance planteado de pronóstico a corto plazo y la simplicidad que ofrece un modelo univariado.

CAPÍTULO 3

Marco Teórico

3.1. Series de Tiempo Univariadas

En esta sección, se presentan de forma muy general los modelos estadísticos tradicionales en series de tiempo univariadas y los métodos de evaluación a través de los criterios de información.

3.1.1. Metodología Box-Jenkins

El modelamiento que se introducirá en esta sección, consiste en ajustar modelos a la estructura general SARIMA $(p\ d,\ q)$ x $(P,\ D,\ Q)_s$ y alguna de sus variaciones. La estrategia para la construcción del modelo se basa en un ciclo iterativo y utiliza los propios datos para la elección de la estructura de este modelo. El modelado de Box-Jenkins, consta de tres etapas para la selección de un modelo: la identificación, estimación y revisión diagnóstica [1], siendo explicado a través de un diagrama de flujo, como el mostrado en la Figura 1:

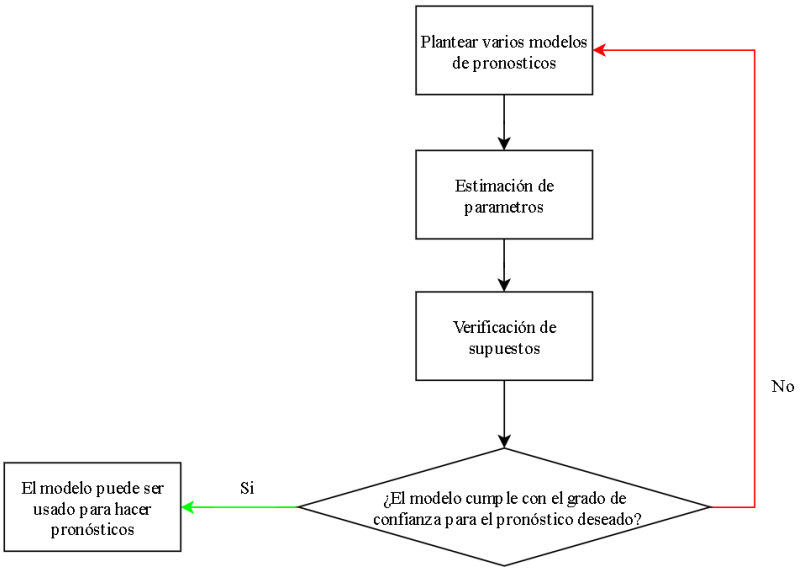

Figura 1. Diagrama de flujo: Metodología Box-Jenkins

La metodología es netamente iterativa, y no se limita a encontrar un único modelo, ya que siempre estará la posibilidad de realizar modificaciones en los parámetros del modelo para perfeccionar el ajuste. En el presente trabajo, se realizaron 729 ajustes de modelos, seleccionando 2 de ellos para evaluarlos y seleccionar el más apropiado

3.2. Modelo SARIMA

Los modelos SARIMA (Autoregresivo integrado de medias móviles estacional), son modelos que incluye una componente estacional, con periodo de estacionalidad *s*. La diferencia estacional de primer orden es la diferencia entre la observación actual y la correspondiente observación del periodo pasado, y es calculada como $z_t = y_t - y_{t-s}$. Para series semanales el periodo de estacionalidad es $s=52$. El modelo es generalmente definido bajo los parámetros $(p, d, q) \times (P, D, Q)_s$, donde (p,d,q) son los parámetros de la componente estacionaria y (P, D, Q), los parámetros de la componente estacional.

3.3. Criterios de Información

Conceptualmente, los datos observados contienen información que se puede expresar de una manera compacta a través de un modelo analítico. El objetivo ideal de la selección de modelos es conseguir una traslación perfecta, uno a uno, de manera que no se pierda información durante el proceso de generación del modelo. Este objetivo es imposible debido a que el conjunto de datos siempre está constituido por un número finito de elementos, que contienen una cantidad limitada de información. Teniendo en cuenta esta dificultad, el objetivo real es obtener el modelo que mejor se ajuste a los datos, esto es, el modelo que

pierda la menor cantidad de información posible. En series de tiempo, la evaluación de los modelos ajustados se hizo de acuerdo a los criterios de información de Akaike (AIC), Bayesiano (BIC) y Akaike modificado (AIC_c).

3.3.1. Criterio de Información de Akaike (AIC)

El criterio de información de Akaike proporciona un método simple y objetivo que selecciona el modelo más adecuado para caracterizar los datos experimentales [2]. Este criterio, que se enmarca en el campo de la teoría de la información, se define como:

$$AIC = -2\,log\left(\mathcal{L}(\hat{\theta})\right) + 2K \quad (1)$$

Donde log($\mathcal{L}$ ($\hat{\theta}$)) es el logaritmo de la máxima verosimilitud, que permite determinar los valores de los parámetros libres de un modelo estadístico, y *K* es el número de parámetros libres del modelo. Esta expresión proporciona una estimación de la distancia entre el modelo y el mecanismo que realmente genera los datos observados, que es desconocido y en algunos casos imposibles de caracterizar. Como la estimación se hace en función de los datos experimentales, esta distancia es siempre relativa y dependiente del conjunto de datos experimentales. Por tanto, un valor individual de AIC no es interpretable por sí solo, y los valores AIC sólo tienen sentido cuando se realizan comparaciones utilizando los mismos datos experimentales [2].

3.3.2. Criterio de información de Akaike modificado (AIC_c)

Cuando el número de parámetros *(K)* es muy elevado en relación con el tamaño de la muestra (n) los resultados que proporciona AIC pueden no ser satisfactorios. En estos casos se utiliza una aproximación de segundo orden [2]:

$$AIC_C = AIC + \frac{2K(K+1)}{n-K-1} \quad (2)$$

Cuando el cociente *n/K* es suficientemente grande, ambos valores (AIC y AICc) son muy similares.

3.3.3. Criterio de información bayesiano (BIC)

El BIC es calculado para los diferentes modelos como una función de la bondad de ajuste del *log Lik*, el número de parámetros ajustados *(K)* y el número total de datos *(n)*. El modelo se basa en parte en la función de probabilidad. Para mejorar la inconsistencia del criterio AIC, Akaike 1978 y Schwarz 1978 presentaron un criterio de selección de modelos desde la perspectiva bayesiana. Schwarz estableció que la solución de bayes consiste en seleccionar

el modelo con una alta probabilidad *a posteriori*. La función está dada por la maximización del logaritmo de la máxima verosimilitud denotado como (ln L) y K es el número de parámetros de la función de probabilidad (parámetros en el modelo) [3]. El criterio de información bayesiana (BIC) se define como:

$$BIC = 2\,ln(L) + K\,ln(n) \quad (3)$$

3.4. Evaluación del pronóstico

Es importante evaluar la precisión del pronóstico utilizando pronósticos genuinos. En consecuencia, el tamaño de los residuos no es una indicación confiable de cuán grandes son los errores de pronóstico reales. La precisión de los pronósticos solo se puede determinar considerando qué tan bien se desempeña un modelo en datos nuevos que no se usaron al ajustar el modelo.

Al elegir los modelos, es una práctica común separar los datos disponibles en dos partes, datos de entrenamiento y de prueba, donde los datos de entrenamiento se usan para estimar cualquier parámetro de un método de pronóstico y los datos de prueba se usan para evaluar su precisión. Debido a que los datos de prueba no se usan para determinar los pronósticos, debe proporcionar una indicación confiable de qué tan bien es probable que el modelo pronostique sobre nuevos datos.

El tamaño del conjunto de pruebas suele ser de aproximadamente el 20% del total de la muestra, aunque este valor depende de la duración de la muestra y de la distancia que desee pronosticar. El conjunto de pruebas idealmente debería ser al menos tan grande como el horizonte de pronóstico máximo requerido. Los siguientes puntos deben tenerse en cuenta:

- Un modelo que se ajuste bien a los datos de entrenamiento no necesariamente se pronosticará bien.
- Siempre se puede obtener un ajuste perfecto utilizando un modelo con suficientes parámetros.
- Sobreponer un modelo a los datos es tan malo como no identificar un patrón sistemático en los datos.

3.3.1. **Errores de Pronóstico**

Un "error" de pronóstico es la diferencia entre un valor observado y su pronóstico. Aquí, "error" no significa un error, significa la parte impredecible de una observación. Se puede escribir como:

$$e_{T+h} = y_{T+h} - \hat{y}_{T+h|T} \quad (4)$$

Donde los datos de entrenamiento son dado por $\{y_1, \ldots, y_T\}$ y los datos de prueba están dados por $\{y_{T+1}, y_{T+2}, \ldots\}$

Tenga en cuenta que los errores de pronóstico son diferentes de los residuos de dos maneras. Primero, los residuos se calculan en el conjunto de entrenamiento, mientras que los errores de pronóstico se calculan en el conjunto de prueba. En segundo lugar, los residuos se basan en pronósticos de un paso, mientras que los errores de pronóstico pueden involucrar pronósticos de varios pasos.

Podemos medir la precisión del pronóstico al resumir los errores del pronóstico de diferentes maneras.

Errores dependientes de la escala

Los errores de pronóstico están en la misma escala que los datos. Las medidas de precisión que se basan solo en e_t son, por lo tanto, dependientes de la escala y no se pueden usar para hacer comparaciones entre series que involucran diferentes unidades.

Las dos medidas dependientes de la escala más utilizadas se basan en los errores medio absolutos (MAE) o los errores cuadráticos medios (RMSE):

$$\text{MAE} = \text{media}(|e_t|) \quad (5)$$
$$\text{RMSE} = \sqrt{(|e_t^2|)} \quad (6)$$

Cuando se comparan los métodos de pronóstico aplicados a una sola serie de tiempo, o a varias series de tiempo con las mismas unidades, el MAE es popular porque es fácil de entender y calcular. Un método de pronóstico que minimiza el MAE conducirá a pronósticos de la mediana, mientras que minimizar el RMSE conducirá a pronósticos de la media. En consecuencia, el RMSE también se usa ampliamente, a pesar de ser más difícil de interpretar.

Porcentajes de Error

El porcentaje de error viene dado por: $p_t=100e_t/y_t$. Los errores de porcentaje tienen la ventaja de estar libres de unidades, por lo que se usan con frecuencia para comparar los rendimientos de pronóstico entre conjuntos de datos. La medida más utilizada es el error medio porcentual absoluto (MAPE):

$$\text{MAPE} = \text{media}\,(|p_t|) \quad (7)$$

Las medidas basadas en errores de porcentaje tienen la desventaja de ser infinitas o indefinidas si $y_t=0$ para algún t en el periodo de interés, y teniendo valores extremos cuando y_t está cerca de cero.

Error escalado

Los errores escalados fueron propuestos por Hyndman & Koehler [4]como una alternativa al uso de errores porcentuales al comparar la precisión de los pronósticos en series con diferentes unidades. Propusieron escalar los errores basándose en el MAE de entrenamiento a partir de un método de pronóstico simple.

Para una serie de tiempo no estacional, una manera útil de definir un error escalado usa pronósticos sencillos:

$$q_j = \frac{e_j}{\frac{1}{T-1}\sum_{t=m+1}^{T}|y_t - y_{t-m}|} \quad (8)$$

Definiéndose así el error medio absoluto escalado (MASE) como:

$$\text{MASE} = \text{media}\,(|q_j|)$$

CAPÍTULO 4

Resultados

4.1. Consolidación de datos

El primer acercamiento al histórico de datos de nivel del río Magdalena a la altura del municipio de Barrancabermeja, consistió en la simplificación de los 2190 datos comprendidos entre el 1 de Enero de 2013 y el 31 de Diciembre de 2018, en una base de datos semanal de 312 datos, donde el dato representantes son calculados como la mediana de los datos de nivel en periodos de 7 días. En la Figura 2, se muestra gráficamente el resumen de esta simplificación.

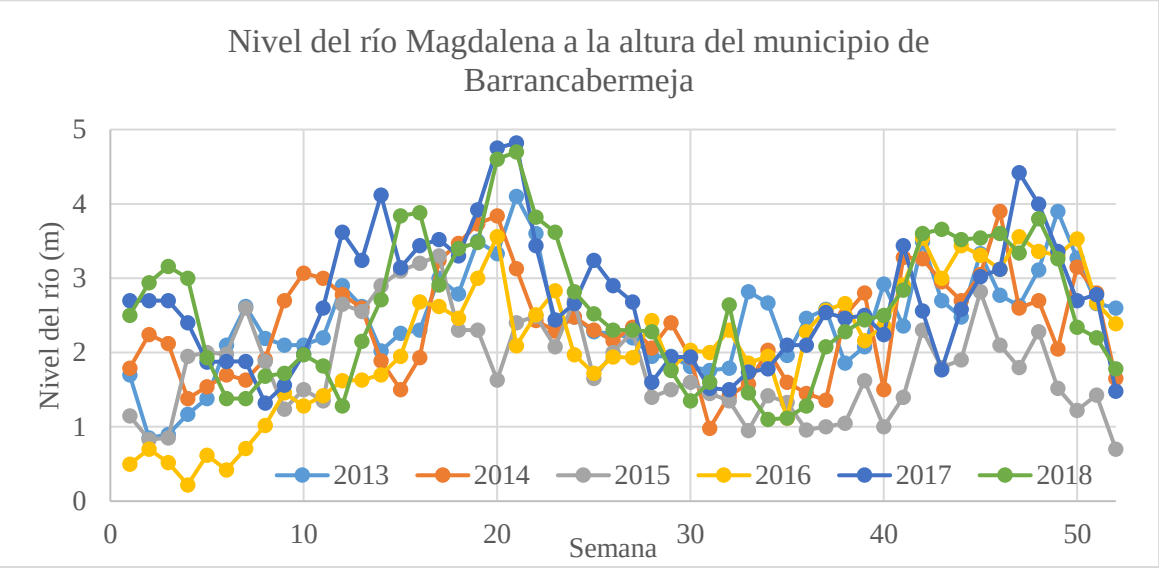

Figura 2. Nivel semanal del río Magdalena 2013 – 2018

El análisis de los datos, empieza con la identificación del comportamiento de la serie de tiempo. En un primer acercamiento, se observa que el comportamiento de la serie de tiempo es periódico, presentado periodos de bajo nivel de rio para los meses de Febrero y Marzo (Semana 5 a 12), seguido de incrementos de nivel en el mes de Abril (Semana 20) y decrecimientos hasta inicios de Noviembre (Semana 40), donde el rio nuevamente recupera su nivel hasta la mitad del mes de Diciembre (Semana 50). Finalmente las últimas dos semanas del año el comportamiento es de reducción de nivel debido a las altas temperaturas registradas en los municipios rivereños. Se destaca el comportamiento heterogéneo para las el mes de Enero de cada año (Semana 1 a 4). De acuerdo a las anteriores características, se establece como premisa que el comportamiento del nivel del rio Magdalena es estacional y estacionario.

Descriptivamente, el nivel del río Magdalena se mantiene en 2.34±0.85 m, presentándose datos atípicos de bajo nivel en Febrero 2016 y alto nivel en Mayo 2017 y2018, debido a los fenómenos de El Niño y La Niña. En el Anexo 1, se encuentra consolidado la compilación de datos semanales desde el año 2013-2018, utilizados para la construcción del modelo del presente trabajo.

4.2. Identificación y diagnóstico del modelo

Como herramienta de identificación del modelo, se construyen los autocorrelogramas y autocorrelogramas parciales muestrales de los datos. La Figura 3 presenta los datos con dos diferencias: en su componente estacional y en su componente no estacional. La prueba de Dickey-Fuller aumentada aporta un p-valor menor del 0.01 para estos datos. Por tanto procederemos con los datos con las diferenciaciones ya mencionadas (d = 1, D = 1).

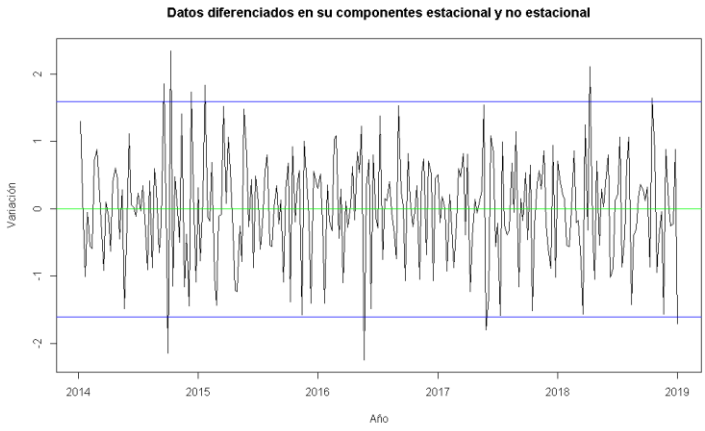

Figura 3. Representación gráfica de la serie de nivel con diferenciación en su componente estacional y no estacional

Para encontrar los modelos candidatos y evaluar su ajuste con el histórico, se programaron en código 729 modelos, programando todas las posibles combinaciones de los parámetros estacionarios *(p,d,q)* y estacionales *(P, D, Q)*, iniciando con el ajuste1: SARIMA (0,0,0) x $(0, 0, 0)_{52}$ y terminando con el ajuste729: SARIMA (2,2,2) x $(2, 2, 2)_{52}$. Elaborados los modelos, se realizaron las pruebas de autocorrelación serial, pruebas de normalidad en los residuos, comportamiento visual de los residuos, y jerarquizando los modelos a partir de los criterios de información: AIC, BIC y AIC_c, se concluye que el mejor modelo para describir y pronosticar el nivel del río Magdalena a la altura del municipio de Barrancabermeja es un modelo SARIMA (2, 1; 2) x $(0; 1; 2)_{52}$. Los parámetros del modelo seleccionado se dan en la Tabla 1. La varianza estimada del modelo es 0.6031 y el BIC calculado es 479.6.

Tabla 1. Parámetros del modelo SARIMA (2, 1, 2) x $(0, 1; 2)_{52}$

Parámetros	ar1	ar2	ma1	ma2	sma1	sma2
Estimación	1.4911	-0.5745	-1.8925	0.9135	-0.8203	0.2805
Error Estándar	0.0659	0.0628	0.0411	0.0388	0.0755	0.1012

En la Figura 4, se muestra el ajuste del modelo SARIMA (2,1, 2) x $(0, 1, 2)_{52}$ con respecto al histórico semanal de 2013 a 2018. Se destaca el óptimo ajuste del modelo para el año 2013, traslapando totalmente el ajuste sobre el histórico.

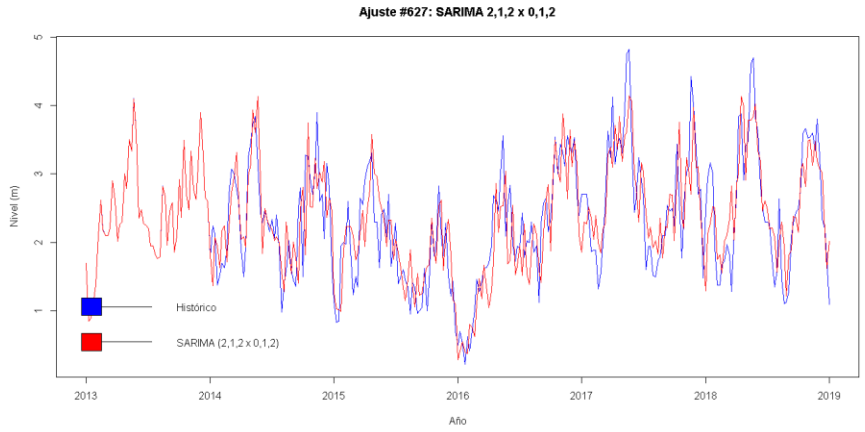

Figura 4. Validación del histórico frente al modelo ajustado SARIMA (2,1, 2) x $(0, 1, 2)_{52}$

Según los ACF y PACF residuales mostrados en las Figura 5 y Figura 6, no se destaca algún problema de autocorrelación serial para el alcance de pronóstico planteado en este trabajo.

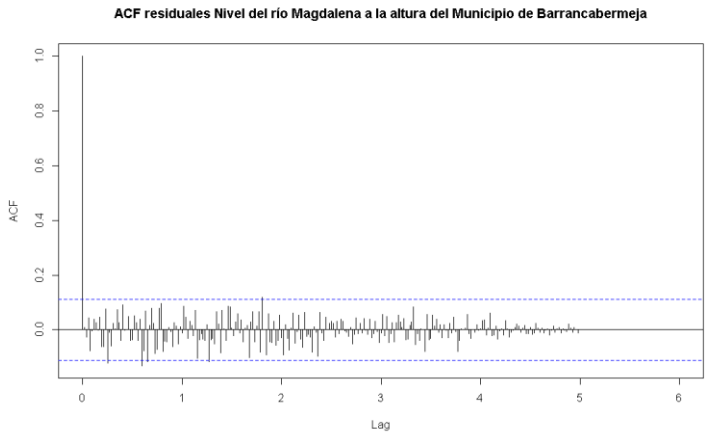

Figura 5. ACF Residuales del modelo ajustado SARIMA (2, 1; 2) x (0; 1; 2)$_{52}$

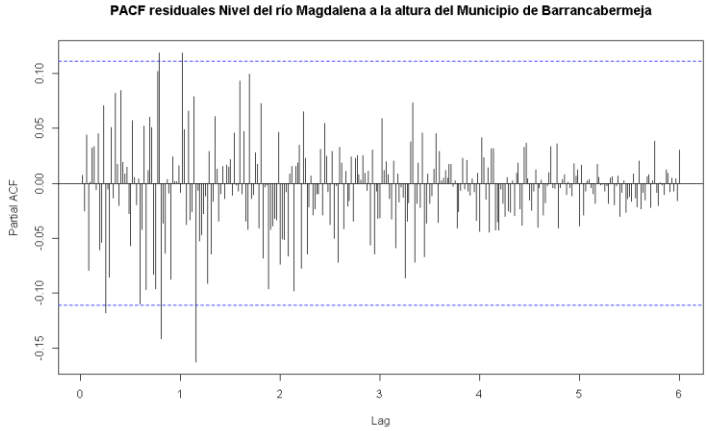

Figura 6. PACF Residuales del modelo ajustado SARIMA (2, 1; 2) x (0; 1; 2)$_{52}$

La distribución de los residuos mostrada en la Figura 7, muestra una distribución de residuos de -0.001764±0.479. El p-valor de la prueba de Ljung-Box es de 0.8961, sugiriendo que con una significancia estadística del 5%, no existe evidencia para rechazar la hipótesis de no correlación serial en los residuos. Finalmente, la normalidad de los residuos se configura como un ruido blanco no gaussiano pues el p-valor en la prueba de Jarque-Bera es de $5.36x10^{-3}$, sugiriendo que no existe evidencia estadísticamente significativa sobre la normalidad de los residuos.

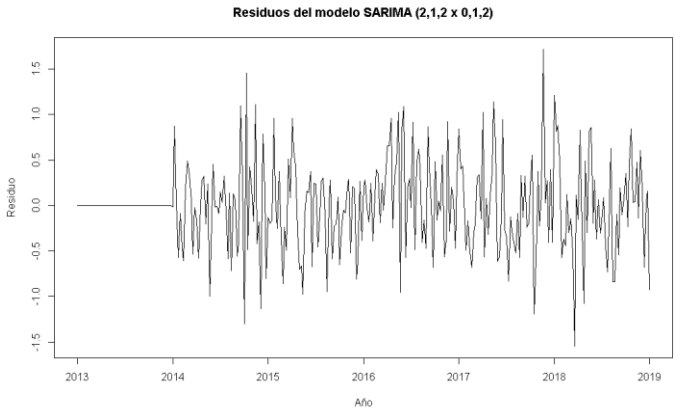

Figura 7. Residuales del modelo SARIMA (2, 1; 2) x (0; 1; 2)$_{52}$

4.3. Validación del modelo

La validación del modelo, consistió en el pronóstico del nivel del río durante las primeras 12 semanas del año 2019, comprendidas entre el 1 de Enero de 2019 y el 25 de Marzo de 2019. Estos datos fueron simplificados de la misma manera, utilizando la mediana de los 7 días, como el dato representativo de cada semana. En la Figura 8 se muestra un comparativo visual entre la predicción del modelo y los datos reales reportados. Se destaca que la totalidad de los pronósticos realizados, se encuentran en las bandas de confianza del 80% y 95%.

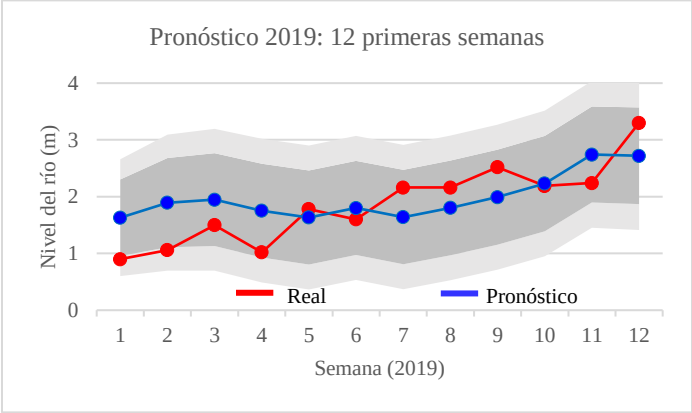

Figura 8. Pronóstico del nivel del río Magdalena 12 periodos de 2019 con bandas de confianza del 80% y 95%.

La validación también incluyo los errores de pronóstico asociado al histórico utilizado para la estimación de parámetros, mostrados en la Tabla 2.

Tabla 2. Errores de Pronóstico para el modelo SARIMA (2, 1; 2) x (0; 1; 2)$_{52}$

ME	RMSE	MAE	MPE	MAPE	MASE
-0.00176431	0.4786551	0.3464962	-4.08355	17.44325	0.8092434

Para el pronóstico realizado en los primeros 12 periodos, se observan zonas con un error residual mayor para las cuatro primeras semanas, en comparación con los pronósticos a partir de la quinta semana. El error promedio residual de este test de 12 semanas fue del 26%, destacando que a partir de la quinta semana, el error residual oscila entre 2% al 13%.

CAPÍTULO 5

Conclusiones y Recomendaciones

- El modelo SARIMA (2, 1; 2) x $(0; 1; 2)_{52}$, puede ser utilizado como un modelo de pronóstico simplificado del nivel del río Magdalena a la altura de Barrancabermeja con un error promedio del 26% entre los datos reales y los correspondientes pronósticos para las 12 semanas, destacando que a partir de la semana 5 de la evaluación realizada, el modelo pronosticó con una mayor precisión, reduciendo el error en el intervalo del 2% al 13%.

- La totalidad de datos pronosticados en la validación del modelo, se encuentran dentro de las bandas de confianza del 80% y 95% del modelo SARIMA (2, 1; 2) x $(0; 1; 2)_{52}$, demostrando que estas bandas de confianza pueden ser utilizadas como herramienta de decisión rápida para la planificación de volúmenes evacuados por medio fluvial.

- Para próximos estudios, se recomienda evaluar la incidencia de variables exógenas continuas como radiación solar (kW/m^2), temperatura promedio del río (°C) y precipitación (mm) a la altura de Barrancabermeja en el pronóstico del nivel del río. De validarse su significancia dentro del modelo de pronóstico, se propone construir un modelo mejorado SARIMAX (Autoregresivo integrado de medias móviles estacional con variables exógenas) para aumentar la robustez y mejore la calidad del pronóstico.

Referencias Bibliográficas

[1] A. Barreras Serrano, E. Sánchez López, F. Figueroa Saavedra, J. Á. Olivas Valdéz y C. Pérez Linares, «Uso de un modelo univariado de series de tiempo para la predicción, en el corto plazo, del comportamiento de la producción de carne de bovino en Baja California, México,» *Instituto de Investigaciones en Ciencias Veterinarias de la Universidad Autónoma de Baja California*, p. 9, 2014.

[2] D. R. Martínez, J. L. Albín, J. C. Cabaleiro, T. F. Pena, F. F. Rivera y V. Blanco, «El Criterio de Información de Akaike en la Obtención de Modelos Estadısticos de Rendimiento,» *Conference: XX Jornadas de Paralelismo.*, pp. 439-444, 2009.

[3] D. S. Calderón Rivera, C. F. Navarrete López y J. L. Díaz Arévalo, «Ajustes de distribuciones probabilísticas para la variable temperatura media multianual para el departamento de Boyacá (Colombia).,» *Ingeniería y Región*, nº 14, pp. 125-142, 2015.

[4] R. J. Hyndman y A. B. Koehler, «Another look at measures of forecast accuracy,» *International Journal of Forecasting*, nº 22, pp. 679-688, 2006.