# Estadística III - 3009137, semestre 01 de 2022

Equipo de trabajo No. 2 Serie No. 3 Curso: Ma – Ju, Horario: 10.00 am – 12.00 m

## **ANÁLISIS DE SERIES DE TIEMPO. AJUSTE DE TENDENCIA Y ESTACIONALIDAD**

Sofía Cuartas García[[1]](#footnote-1), Simón Cuartas Rendón[[2]](#footnote-2) y Deivid Zhang Figueroa[[3]](#footnote-3)

*Fecha de entrega: 25 – 06 – 2022*

1. **ANÁLISIS DESCRIPTIVO Y TEST HEGY DE RAÍCES UNITARIAS**

En este trabajo se van a plantear modelos SARIMA para ajustar la serie temporal del ***índice de ventas del sector manufacturero colombiano en pesos nominales*** que publica el Departamento Administrativo Nacional de Estadística (DANE) cada mes, con el objetivo de poder realizar pronósticos para este índice de cara al futuro, y a su vez, se van a contrastar estos modelos con modelos ARMA y modelos globales y locales que han sido estudiados anteriormente. Para ello, es importante tener presente que la serie temporal de índice de ventas en pesos nominales del sector manufacturero colombiano con la que se trabajará va desde enero de 2001 hasta noviembre de 2021, lo cual implica que se tienen observaciones para esta serie de tiempo (teniendo presente que aquellos datos asociados a los meses de marzo de 2020 en adelante fueron imputados para ignorar los efectos que tuvo la pandemia por la *COVID-19* en este índice, entre otros indicadores económicos), si bien para el ajuste de los todos los modelos fueron empleadas los valores del índice de ventas en pesos nominales para los primeros , dejando las últimas observaciones como parte del periodo *ex post* para verificar la calidad del pronóstico mediante la estrategia de validación cruzada.

Mencionado lo anterior, se comienza revisando las características de la serie de tiempo a trabajar en la ***figura 1*** considerando las observaciones de esta.

|  |  |
| --- | --- |
| Chart  Description automatically generated**(a)** | Chart, line chart  Description automatically generated**(b)** |

***Figura 1.*** Serie temporal de ***índice de ventas del sector manufacturero colombiano en pesos nominales*** desde enero de 2001 a noviembre de 2021. **(a)** Serie con el índice de ventas en pesos nominales en escala original. **(b).** Serie temporal con el índice de ventas en pesos nominales en escala logarítmica.

En la ***figura 1*** ***(a)*** se encuentra el índice de ventas del sector manufacturero colombiano en pesos nominales presentado en escala original. En esta se puede observar que la serie muestra tendencia creciente, ya que el índice bajo estudio tiende a aumentar con el tiempo; asimismo, se puede evidenciar cómo este índice va presentando una mayor varianza conforme pasa el tiempo, lo cual señala que esta serie es multiplicativa; de igual manera se aprecia que la tendencia es global en tanto es posible emplear una función suave del tiempo para modelarla, y además, se puede decir que la tendencia es determinística, puesto que su evolución se puede calificar como predecible. Por otro lado, es clara la presencia de la componente estacional en esta serie, debido a que existe un comportamiento repetitivo en cada año de la serie, y se puede decir que esta es aproximadamente exacta con meses, pues en general se observa que el índice de ventas en pesos nominales inicia con un mínimo en los meses de enero, y luego muestra un incremento rápido hasta alcanzar sus máximos anuales en los meses de diciembre, para luego mostrar una reducción brusca al comenzar enero del siguiente año y comenzar así un nuevo periodo anual, y por lo anterior es también razonable decir que la estacionalidad es determinística. No obstante, vale la pena señalar que en algunos años este comportamiento estacional no tiene la misma apariencia en comparación con la mayoría de los años, según como se acabó de describir, esto gracias a la presencia de ciclos en la serie, como ocurre por ejemplo entre los años 2008 a 2011, lo cual implica que una modelación global de la estacionalidad pueda tener inconvenientes en contraste con un modelo local que puede tener mejores resultados en este respecto. Para finalizar esta descripción inicial, se debe destacar que en la ***figura 1 (b)*** se aprecia como la variabilidad alrededor de la tendencia media de largo plazo se estabiliza al hacer una transformación logarítmica en la escala del índice de ventas del sector manufacturero colombiano en pesos nominales.

A continuación, es necesario observar la ***figura 2,*** para la que se considera la serie recortada con las primeras observaciones, esto es, excluyendo el periodo *ex post,* en la cual se presentan lo gráficos la serie recortada, su primera diferencia regular, su primera diferencia estacional y de la serie diferenciada por tendencia y estacionalidad (en todos los casos empleando la transformación logarítmica por tener la varianza estabilizada), así como de los gráficos de las funciones de autocorrelación asociadas.

|  |  |  |  |
| --- | --- | --- | --- |
| **Chart, line chart  Description automatically generated(a)** | **Chart  Description automatically generated(c)** | **Chart  Description automatically generated with medium confidence(e)** | **Chart  Description automatically generated(g)** |
| **Chart, histogram  Description automatically generated(b)** | **Chart, box and whisker chart  Description automatically generated(d)** | **Chart, histogram  Description automatically generated(f)** | **Chart  Description automatically generated(h)** |

***Figura 2.* (a).** Serie recortada, **(b)** ACF muestral de la serie recortada; **(c)** primera diferencia regular de la serie recortada, **(d)** ACF de la primera diferencia regular de la serie recortada; **(e) p**rimera diferencia estacional serie recortada, **(f)** ACF muestral de la primera diferencia estacional de la serie recortada; **(g)** primera diferencia regular y estacional de la serie recortada, **(h)** ACF muestral de la serie recortada diferenciada por tendencia y estacionalidad. Nótese que los rezagos asociados a la estacionalidad han sido delineados con una línea punteada azul oscura vertical.

Se comienza entonces discutiendo la serie temporal transformada mediante logaritmo natural y sin diferenciar por tendencia ni estacionalidad, graficada en la ***figura 2 (a).*** En esta salta a la vista inicialmente que la media no es constante gracias a la componente estructural de la tendencia, que hace que la media sea creciente, pero también gracias a la estacionalidad, que provoca que disminuya al comenzar cada año respecto al último periodo del año anterior y aumente hasta finalizar el año. No obstante, parece que la varianza es constante, lo cual se debe a la transformación logarítmica. Luego, estudiando la función de autocorrelación muestral (en adelante ***ACF***) asociada a esta serie e ilustrada en la ***figura 2 (b),*** definida en este caso como , se observa para primeros seis rezagos que la función de autocorrelación toma valores muy próximos a uno, lo que hace factible la existencia de raíz unitaria regular en esta serie, lo cual hace calcular la primera diferencia regular de la serie temporal, y de igual manera, se observa que los valores de la función de autocorrelación disminuyen muy lentamente, lo cual implica que la parte regular no es un proceso ergódico; después, al analizar la parte estacional en , se evidencia una cola positiva que decae de forma lenta, lo cual indica que también sea muy probable la existencia de una raíz unitaria estacional, por lo que resultará también conveniente hallar la primera diferencia estacional para el logaritmo del índice de ventas en pesos nominales del sector manufacturero, y como también se observa que estos valores se acercan a cero de forma lenta, se concluye que la parte estacional no es ergódica. Con todo lo anterior, se concluye que este primer proceso **no es estacionario en covarianza.**

Después, tomando la primera diferencia del logaritmo del índice de interés cuya gráfica se tiene en la ***figura 2 (c)***, se puede evidenciar que el nivel se ha estabilizado pero la media no es constante ya que persiste el patrón periódico en la serie, reflejado en los picos que se observan a lo largo de la serie y que ocurren justo en el comienzo y final de cada año (nótense los picos que se ciernen sobre las líneas punteadas verticales de los años 2005, 2010 y 2015); la varianza por su parte parece ser constante. Luego, pasando a la ACF de esta serie, definida como y retratada en la ***figura 2 (d)***, y comenzando con la parte regular, se puede evidenciar que la ACF, para los primeros seis rezagos, toma valores que son cercanos a cero, por lo que podría ser un proceso ergódico; ahora, respecto a la parte estacional, se puede notar que la ACF toma sus valores más altos en y con disminuciones progresivas muy pequeñas, lo que es señal de que el proceso, en su parte estacional, no es ergódico. Así, se concluye que el proceso diferenciado por tendencia **no es estacionario en sentido débil.** No obstante, es importante notar que basta con diferenciar una vez la serie por tendencia, ya que la tendencia es estabilizada y la parte regular del proceso es ergódico.

A continuación, se va a analizar el proceso diferenciado por estacionalidad, graficado en la ***figura 2 (e)***, en la cual se puede ver que ya no hay un patrón periódico exacto, si bien el nivel no es estable, siendo claro entonces que la media no es constante; de la misma manera, se evidencia que la variación del proceso diferenciado por estacionalidad no es igual para todos los periodos (evidenciado, por ejemplo, en el primer quinquenio con una alta variabilidad en contraste con el último), por lo que no se cumple el supuesto de homocedasticidad; pero, de todos modos es importante señalar que ya no se observa un patrón periódico en la serie. Así pues, avanzando con la ACF muestral, presentada en la ***figura 2 (f)*** y definida en este caso como , resalta para la parte regular con los primeros cinco rezagos que estos toman valores que no son cercanos a uno y que presentan un decaimiento adecuado a cero, por lo que se concluye que la parte regular es ergódica y, además, parece que su patrón es tipo cola sinusoidal exponencial. Luego, en cuanto a la parte regular, se observa que siempre toma valores que pueden ser considerados estadísticamente iguales a cero, por lo que se puede tomar igualmente como ergódico. Con todo esto se concluye que la serie diferenciada por primera vez por estacionalidad **no es estacionaria en covarianza,** aunque se debe señalar que es suficiente con diferenciarla una vez por estacionalidad en tanto esto hace desaparecer las componentes periódicas exactas o aproximadamente exactas y la parte estacional es ergódica, eliminando entonces la raíz unitaria estacional.

Por último, al revisar el proceso asociado a la serie temporal del índice de ventas del sector manufacturero colombiano en pesos nominales tomando tanto la primera diferencia regular como la primera diferencia estacional, y cuya gráfica se tiene en la ***figura 2 (g)***, se observa que el nivel ha sido estabilizado y que ya no se aprecian patrones periódicos exactos o casi exactos (nótese que ya no hay picos separados entre sí por la misma cantidad de periodos), y además, se puede juzgar al proceso como homocedástico en tanto la varianza es aproximadamente igual a lo largo del proceso. Después, analizando la ***figura 2 (h)*** que grafica la función de autocorrelación muestral para este proceso, que se define como , se tiene para la parte regular que para el primer y el cuarto rezago la función de autocorrelación es estadísticamente diferente de cero, aunque se debe resaltar que para el primer rezago toma un valor que no es próximo a uno y este es el mayor de todos entre los rezagos analizados para la parte regular , siendo finalmente la función de autocorrelación muestral muy próxima a cero, por lo que se puede asumir que la parte regular es ergódica; dicho esto, y analizando la parte estacional, se observa que para la función de autocorrelación muestral es estadísticamente significativa al tomar valores que superan los límites de Bartlett, pero este no es el caso para , donde no solo no se superan tales límites sino que se tiene un valor muy próximo a cero, de tal suerte que se puede considerar que el proceso en su parte estacional es ergódico, y cumpliéndose la ergodicidad para la parte estacional y regular, entonces se concluye que el proceso es ergódico, lo cual muestra además que ya no hay evidencia de existencia de raíces unitarias regulares y estacionales, por lo que no es necesario diferenciar más por tendencia o estacionalidad, por lo que los modelos SARIMA que van ser planteados considerarán . En definitiva, cumpliéndose los supuestos de media constante, homocedasticidad y ergodicidad, se concluye que este proceso **sí es estacionario en sentido débil.**

Finalmente, se tiene la ***tabla 1*** que resume todas las conclusiones argumentadas anteriormente.

**Tabla 1.** Resumen de los análisis realizados sobre cada una de las cuatro series. Con ✔ se indica el cumplimiento del supuesto y con ✘ el no cumplimiento del supuesto.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Característica** | **Serie uno** | **Serie dos** | **Serie tres** | **Serie cuatro** |
|  |  |  |  |
| **Media constante** | ✘ | ✘ | ✘ | ✔ |
| **Homocedasticidad** | ✔ | ✔ | ✘ | ✔ |
| **Ergodicidad** | ✘ | ✘ | ✔ | ✔ |
| Ergodicidad en la parte regular | ✘ | ✔ | ✔ | ✔ |
| Ergodicidad en la parte estacional | ✘ | ✘ | ✔ | ✔ |
| **Estacionariedad en covarianza** | **✘** | **✘** | **✘** | **✔** |

Ahora, se va a realizar el ***test HEGY*** para esta serie, el cual es un test tipo *Dickey Fuller aumentado, ADF,* con la salvedad de que la serie temporal será representando como un proceso autorregresivo infinito que supone que este es invertible. Así, se tiene que un proceso ARMA tendrá raíz unitaria estacional de periodo doce si su polinomio AR tiene como raíces una o más raíces periódicas del polinomio , las cuales se expresan como , y de este modo, si se asume que el proceso dado por logaritmo del índice de ventas del sector manufacturero colombiano en pesos nominales es invertible, entonces tendría representación AR(∞), donde el polinomio autorregresivo será , de manera que la representación AR de la serie será dada por:

Y así, la representación del modelo de regresión lineal múltiple apropiado para el test ADF se muestra en la ***ecuación 1:***

**[1]**

.

Donde las variables resultan de aplicar filtros tales que los coeficientes de regresión estén asociadas a las raíces , y tales variables son las siguientes:

Así pues, el juego de hipótesis para las pruebas del test HEGY y sus interpretaciones son las presentadas en la segunda columna de la ***tabla 2.*** Para cada una de las siete pruebas presentadas en tal prueba se tiene que sus estadísticos de prueba bajo poseen la distribución de un proceso estocástico complejo. Con esto presente, se acude a ***R*** para poder realizar estas pruebas y los resultados son presentados en las últimas tres columnas de la ***tabla 2***, y como se puede observar, no se rechazó la hipótesis nula para el primer, tercer, quinto, sexto y séptimo test, lo cual indica que hay evidencia muestral suficiente para sugerir que el proceso asociado a la serie temporal de interés **presenta raíces unitarias regulares y estacionales,** a saber:cuatrimestral, semestral, con cinco ciclos al año y anual**,** con una significancia de , lo cual es un indicador de que no es estacionario y es necesario aplicarle filtro de diferencia mixto, como el resultado en la ***figura 2 (g)***, que como se analizó recientemente, al tomar la primera diferencia regular y la primera diferencia estacional, se llega a un proceso estacionario en covarianza.

***Tabla 2.*** Juego de hipótesis para las pruebas del test HEGY y sus resultados calculados con el paquete estadístico ***R.***  En la última columna se usa “✔” para indicar que se rechaza la hipótesis nula.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Test** | **Hipótesis** |  | **Frecuencia angular (*θ*)** | **Ciclos por año (*j*)** | **Estadístico de prueba calculado** | **Valor p ()** | **¿Rechazo?** |
| 1 |  | Hay raíz unitaria regular o no estacional |  | 0 | 2.9135488 | 0.10000000 |  |
| No hay raíz unitaria regular o no estacional |
| 2 |  | Hay raíz unitaria estacional bimensual |  | 6 | -2.8854438 | 0.01000000 | ✔ |
| No hay raíz unitaria estacional bimensual |
| 3 |  | Hay raíz unitaria estacional cuatrimestral |  | 3 | 0.8104567 | 0.10000000 |  |
| No hay raíz unitaria estacional cuatrimestral |
| 4 |  | Hay raíz unitaria estacional trimestral |  | 4 | 2.4693601 | 0.09062608 | ✔ |
| No hay raíz unitaria estacional trimestral |
| 5 |  | Hay raíz unitaria estacional semestral |  | 2 | 1.7683136 | 0.10000000 |  |
| No hay raíz unitaria estacional semestral |
| 6 |  | Hay raíz unitaria estacional de cinco ciclos al año |  | 5 | 2.2239858 | 0.10000000 |  |
| No hay raíz unitaria estacional de cinco ciclos al año |
| 7 |  | Hay raíz unitaria estacional anual |  | 1 | 1.6332640 | 0.10000000 |  |

1. **IDENTIFICACIÓN DE MODELOS *SARIMA(p, d, q)(P, D, Q)[12]***

Realizados los análisis anteriores, ahora se van a identificar modelos **SARIMA(p, d, q)(P, D, Q)[s],** esto partiendo del proceso derivado de la serie temporal del índice de ventas del sector manufacturero colombiano con precios en pesos colombianos nominales que ha sido diferenciada por tendencia y estacionalidad, de manera que ya se sabe que para los modelos a presentar se cumple que . Mencionado esto, se partirá de las evaluaciones que se pueden realizar con la ACF, cuya gráfica se presenta nuevamente en la ***figura 3 (a)*** y con la PACF, que se presenta en la ***figura 3 (b).***

|  |  |
| --- | --- |
| **(a)** | **(b)** |

***Figura 3. (a)***Función de autocorrelación parcial (ACF) y ***(b)*** función de autocorrelación parcial muestral (PACF) del logaritmo del índice de ventas del sector manufacturero colombiano en pesos nominales tomando su primera diferencia regular y su primera diferencia estacional.

Así pues, comenzando con el análisis de la parte regular, se observa en la gráfica de la ACF que esta tiene un patrón tipo cola exponencial sinusoidal, mientras que en la PACF se identifica un patrón tipo corte, donde el último rezago estadísticamente diferente de cero es el segundo, por lo que para la parte regular se identifica un ***AR(2),*** pero teniendo presente que se realiza este análisis a partir de la serie diferenciada por tendencia y estacionalidad, entonces se tiene un ***ARIMA(2, 1, 0).*** Después, pasando a la parte estacional, se puede observar en la función de autocorrelación parcial que los dos primeros rezagos estacionales son significativamente diferentes de cero, mientras que el primero no lo es, por lo que se puede pensar que la ACF tiene un patrón tipo corte con último rezago estacional significativo en , mientras que con la PACF se observa que esta es significativa para y que tiene un patrón cola, por lo que se debe cumplir que la parte estacional se pueda modelar como un ***MA(q)[s]***, donde , es decir, un ***MA(2)[12]***,pero teniendo presente la diferencia estacional empleada nuevamente, entonces se llega a un ***ARMA(0, 1, 2)[12],*** y al unir los análisis de la parte estacional y regular se concluye que un primer modelo adecuado es un ***ARIMA(2, 1, 0)(0, 1, 2)[12],*** coincidiendo entonces con el ***modelo uno*** propuesto en el enunciado, y cuya ecuación teórica está dada por:

**[2]**

Hecho esto, se procederá a la identificación de modelos **SARIMA(p, d, q)(P, D, Q)[s]** empleando métodos automáticos, comenzando con la función auto.arima() de la librería **forecast** (versión 8.16) del paquete estadístico ***R***, y cuyas salidas se visualizan en la ***figura 4.***

|  |  |  |
| --- | --- | --- |
| **A screenshot of a computer  Description automatically generated with low confidence(a)** | **Text  Description automatically generated(b)** | **A screenshot of a computer  Description automatically generated with low confidence(c)** |
| **Text  Description automatically generated(d)** | **Text  Description automatically generated(e)** | **Text, letter  Description automatically generated(f)** |

***Figura 4.*** Identificación de métodos **SARIMA(p, d, q)×(P, D, Q)[12]** usando: **(a)** AIC y *ocsb*, **(b)** AIC y *ch*, **(c)** AIC y *seas*, **(d)** BIC y *ocsb*, **(e)** BIC y *ch*, **(f)** BIC y *seas.* El objeto lny que usa en el primer argumento de cada función es la serie temporal de interés transformada usando logaritmo natural.

Así, se presenta en la ***tabla 3*** los modelos que han sido identificados, así como sus ecuaciones. Debe notarse que para los modelos que se identificarán como **uno** y ***dos,*** sus salidas en ***R*** tienen el indicador *with drift* (con deriva), por lo que se le debe agregar el sumando , que corresponde a tal deriva. Para el resto de los modelos no se da esta situación.

***Tabla 3.*** Ecuaciones teóricas de los modelos identificados de forma automática con la función auto.arima().

|  |
| --- |
| **Modelo uno.** Usando el criterio de información AIC y método de selección de diferencias estacionales *ocsb.*  **ARIMA(0, 1, 2)(2, 0, 0)[12] con deriva.** |
| **Modelo dos.** Usando el criterio de información AIC y método de selección de diferencias estacionales *ch.*  **ARIMA(0, 1, 2)(2, 0, 0)[12] con deriva.** |
| **Modelo tres.** Usando el criterio de información AIC y método de selección de diferencias estacionales *seas.*  **ARIMA(4, 1, 0)(1, 1, 2)[12] sin deriva.** |
| **Modelo cuatro.** Usando el criterio de información BIC y método de selección de diferencias estacionales *ocsb.*  **ARIMA(1, 1, 1)(2, 0, 0)[12] sin deriva.** |
| **Modelo cinco.** Usando el criterio de información BIC y método de selección de diferencias estacionales *ch.*  **ARIMA(1, 1, 1)(2, 0, 0)[12] sin deriva.** |
| **Modelo seis.** Usando el criterio de información BIC y método de selección de diferencias estacionales *seas.*  **ARIMA(0, 1, 2)(0, 1, 1)[12] sin deriva.** |

De este modo, se debe notar que los modelos que se identifican como ***uno, dos, cuatro*** y ***cinco*** no resultan apropiados debido a que solo solo están considerando a la primera diferencia regular y no a la primera diferencia estacional, lo cual no resulta apropiado teniendo en cuenta lo arrojado por el test HEGY, que indica que hay evidencia muestral suficiente para sugerir que existen raíces unitarias regulares y estacionales en el proceso asociado a la serie temporal del índice de ventas del sector manufacturero colombiano en pesos nominales, hecho que también fue verificado al observar la ACF de la serie con transformación logarítmica y sin transformar en la ***figura 2 (b)***, de manera que es necesario incluir tanto la primera diferencia regular gracias a la tendencia, como la primera diferencia estacional, tal y como hacen los modelos ***tres*** y ***seis,*** que justamente se realizaron bajo el mismo método de selecciones de diferencias estacionales: ***“seas”***, pero diferenciándose en el criterio de información, que fue AIC para el modelo ***tres*** y BIC para el modelo ***seis.*** Lo anterior también se evidencia en las gráficas de la ***figura 2***: cuando solo se diferencia por tendencia (primera diferencia regular) o por estacionalidad (primera diferencia regular), los procesos resultantes ***no*** son estacionarios, mientras que el proceso derivado de diferenciar tanto por estacionalidad como por tendencia ***sí*** lo es. Ahora bien, para el modelo ***seis*** se debe notar que para la parte regular sugiere un proceso de medias móviles, lo cual contradice lo identificado con la ACF y la PACF de , que en el sentido de que debería tratarse de un proceso autorregresivo, por lo que se descarta esta opción también, quedando únicamente la ***tres,*** que de hecho coincide con el modelo **dos** del enunciado y con el que se trabajará enseguida.

Ahora, se va a utilizar otra función, que es armasubsets, de la librería TSA (versión 1.3), y la cual genera un tablero en el que se resaltan los parámetros estadísticamente significativos que, de acuerdo con esta función, deben ser incorporados en el modelo. Esto se hará de dos modos: con un tablero , visible en la ***figura 5*** y otro tablero , presentado en la ***figura 6,*** y teniendo en cuenta que estos tableros han sido obtenidos con la serie diferenciada por estacionalidad y tendencia, por lo que se obtendrán modelos ARIMA con .

Así, se tiene que, a partir del tablero de armasubsets de la ***figura 5***, y al ceñirse únicamente el primer renglón, se obtiene un modelo ***ARIMA(2, 1, 10)(0, 1, 1)[12]*** que solo incluye a los términos y , como se puede apreciar en la ***tabla 4:***

***Figura 5.*** Gráfico arrojado por armasubsets sobre usando el método “*ols*”.

***Tabla 4.*** Parámetros identificados a partir del primer renglón del tablero arrojado por armasubsets sobre usando el método “*ols*”.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Parte AR*** | | ***Parte MA*** | |
| **Celda sombreada** | **Parámetros identificados** | **Celda sombreada** | **Parámetros identificados** |
| 1 |  | 10 |  |
| 2 |  | 12 |  |

Así las cosas, la ecuación teórica del modelo identificado es la ***ecuación 3:***

**[3]**

 Sin embargo, si se incorpora el término , el modelo se transforma en un ***ARIMA(6, 1, 10)(0, 1, 1)[12]***, y la ecuación sería como se muestra en la ***ecuación 4***:

**[4]**

A picture containing text, crossword puzzle

Description automatically generated

A picture containing text, crossword puzzle

Description automatically generated Por otro lado, con el primer renglón del tablero de armasubsets de la ***figura 6*** se identifica un modelo ***ARIMA(9, 1, 10)(0, 1, 1)[12]*** que solo emplea los términos y , tal y como se ilustra en la ***tabla 5.***

***Figura 6.*** Gráfico arrojado por armasubsets sobre usando el método “*ols*”.

***Tabla 4.*** Parámetros identificados a partir del primer renglón del tablero arrojado por armasubsets sobre usando el método “*ols*”.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Parte AR*** | | ***Parte MA*** | |
| **Celda sombreada** | **Parámetros identificados** | **Celda sombreada** | **Parámetros identificados** |
| 1 |  | 10 |  |
| 2 |  | 12 |  |
| 9 |  |  |  |

Y se llega entonces a la ***ecuación 6:***

**[6]**

Hecho lo anterior, se puede proceder ahora con el ajuste algunos modelos, como se muestra en la tercera sección.

1. **AJUSTE DE MODELOS DE REGRESIÓN SARIMA(p, q)×(P, Q)[12]**

Se considerarán cuatro modelos SARIMA: **ARIMA(2,1,0)(0,1,2)[12], ARIMA(4,1,0)(1,1,2)[12], ARIMA(6,1,10)(0,1,1)[12]** y **ARIMA(9,1,10)(0,1,1)[12].** En este sentido, se presentan inicialmente las ecuaciones de los modelos propuestos en la ***tabla 5***.

Cada uno de estos cuatro modelos es ajustado con ayuda del software especializado en estadística ***R,*** y los parámetros estimados, así como la evaluación de la significancia estadística de cada uno de los parámetros para los cuatro modelos es exhibido en la ***tabla 6***.

***Tabla 5.*** Ecuaciones teóricas de los modelos propuestos.

|  |
| --- |
| **Modelo 1. ARIMA(2, 1, 0)(0, 1, 2)[12]** |
| **Modelo 2. ARIMA(4, 1, 0)(1, 1, 2)[12]** |
| **Modelo 3. ARIMA(6, 1, 10)(0, 1, 1)[12]** |
| **Modelo 4. ARIMA(9, 1, 10)(0, 1, 1)[12]** |

|  |  |
| --- | --- |
| **Tabla 6.** Parámetros estimados modelos SARIMA | |
| **Tabla 6a.** Parámetros estimados en Modelo 1   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Parámetros | Estimación | Error Std |  |  | |  | -0.721755 | 0.063428 | -11.3791 | < 2.2×10-16 | |  | -0.332388 | 0.063474 | -5.2366 | 1.636×10-7 | |  | -0.784245 | 0.082040 | -9.5593 | < 2.2×10-16 | |  | -0.125373 | 0.086879 | -1.4431 | 0.149 | | AIC= 5.514306, BIC= 5.844665 | | | | | | **Tabla 6b.** Parámetros estimados en Modelo 2   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Parámetros | Estimación | Error Std |  |  | |  | -0.684906 | 0.067280 | -10.1799 | < 2.2×10-16 | |  | -0.292794 | 0.081324 | -3.6003 | 0.0003178 | |  | 0.023450 | 0.081146 | 0.2890 | 0.7725958 | |  | -0.069487 | 0.066520 | -1.0446 | 0.2962056 | |  | -0.668689 | 0.264019 | -2.5327 | 0.0113177 | |  | -0.091140 | 0.248068 | -0.3674 | 0.7133203 | |  | -0.704732 | 0.192679 | -3.6575 | 0.0002546 | | AIC= 5.644018, BIC= 6.248974 | | | | | |
| **Tabla 6c.** Parámetros estimados en Modelo 3   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Parámetros | Estimación | Error Std |  |  | |  | -0.742123 | 0.062208 | -11.9298 | < 2.2×10-16 | |  | -0.344376 | 0.064892 | -5.3069 | 1.115×10-7 | |  | 0.099804 | 0.054453 | 1.8329 | 0.06682 | |  | 0.079693 | 0.069183 | 1.1519 | 0.24936 | |  | -0.860131 | 0.056870 | -15.1245 | < 2.2×10-16 | | AIC= 5.622459, BIC= 6.046614 | | | | | | **Tabla 6d.** Parámetros estimados en Modelo 4   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Parámetros | Estimación | Error Std |  |  | |  | -0.757956 | 0.061594 | -12.3056 | < 2.2×10-16 | |  | -0.367751 | 0.063044 | -5.8333 | 5.436×10-9 | |  | 0.120641 | 0.066607 | 1.8112 | 0.07010 | |  | 0.154306 | 0.087548 | 1.7625 | 0.07798 | |  | -0.849464 | 0.054735 | -15.5196 | < 2.2×10-16 | | AIC= 5.588376, BIC= 6.009960 | | | | | |

Entonces, teniendo en cuenta los valores *p* mostrados en la quinta columna de las tablas 4a, 4b, 4c y 4d, se puede determinar la significancia estadística de cada uno de los parámetros de interés. Así pues, usando una significancia de , se tiene que en el **modelo uno** todos los parámetros son significativos excepto ; en el **modelo dos** todos los parámetros son significativos excepto , y ; en el **modelo cuatro** solo y son no significativos estadísticamente, y por último en el **modelo cuatro**, solo y son no significativos. Para el desarrollo de los análisis de este trabajo se considerarán los modelos con todos los parámetros que han sido estimados, aún si estos no son significativos. Sin embargo, resulta un ejercicio interesante ajustar modelos que incluyan únicamente los modelos que han resultado significativos de acuerdo con los resultados de la ***tabla 7*** y comparar la calidad de sus ajustes y pronósticos, y fundamentalmente, el cumplimiento de los supuestos de estos modelos con los modelos **uno, dos, tres** y **cuatro** para así poder elegir el mejor de todos a la luz de estos análisis.

Por otro lado, vale la pena observar la ***figura 7***, en la cual se observa cómo es el ajuste de cada uno de los modelos comparado con la serie temporal. De las gráficas de ajuste se observa que al modelar se conseguir seguir la componente cíclica en todos los modelos SARIMA. Gráficamente no se logra identificar cuál modelo tiene mejor ajuste porque todas siguen la dinámica de la serie de forma adecuada. Además, se puede observar la presencia de datos atípicos que no logran ser explicados por ningún modelo, esto debido a que son observaciones que se comportan diferente, sin embargo, según las medidas de ajuste AIC y BIC, los modelos con mejor ajuste son el **modelo uno** y el **modelo cuatro**, siendo mejor el **modelo uno** que tiene menor AIC y BIC.

|  |  |
| --- | --- |
| Chart, histogram, scatter chart  Description automatically generated**(a)** | Gráfico, Histograma, Gráfico de dispersión  Descripción generada automáticamente**(b)** |
| Chart, histogram, scatter chart  Description automatically generated**(c)** | Gráfico, Histograma, Gráfico de dispersión  Descripción generada automáticamente**(d)** |

**Figura 7.** Graficas con ayuda de ***R*** de la serie original (en negro) y la serie ajustada (en rojo) para cada uno de los modelos, a saber: **(a)** modelo uno: ARIMA(2,1,0)(0,1,2)[12], **(b)** modelo dos: ARIMA(4, 1,0)(1,1,2)[12], **(c)** modelo tres: ARIMA(6, 1, 10)(0, 1, 1)[12] y **(d)** modelo cuatro: ARIMA(9, 1, 10)(0, 1, 1)[12]

1. **ANÁLISIS DE RESIDUALES Y VALIDACIÓN DE SUPUESTOS**

Ahora bien, se debe tener en cuenta que para el planteamiento de cada uno de los modelos se partieron de varios supuestos sobre implícitos al suponer que estos son un ruido blanco, los cuales son: media cero, varianza constante y ausencia de patrones contrarios a la independencia y la distribución normal. No obstante, teniendo en cuenta que estos no son conocidos, para poder validar los supuestos se va a realizar esta revisión mediante los residuales, que son los estimadores de , para lo cual se tiene la ***figura 8,*** en la que se pueden visualizar a los residuales contra el tiempo, y en la ***figura 9,*** los residuales contra los valores ajustados de cada modelo. Ambas figuras se presentan juntas en la página nueve.

Así pues, vale la pena comenzar evaluando si se cumple el supuesto de media cero y varianza constante para cada modelo, y al revisar cada uno de los gráficos se valida que ninguna aporta evidencia en contra de los supuestos de homocedasticidad y media constante en cero, y tampoco se detecta que haya patrones contrarios a la independencia mediante estos gráficos como ciclos o rachas de signos positivos y negativos.

Asimismo, es necesario evaluar las gráficas del ACF y PACF para cada uno de los modelos en la ***figura 10*** (visible en la página diez),esto con el fin de validar el supuesto de incorrelación para , donde para el **ACF** el estadístico de prueba es , y que con una significancia de aproximadamente se rechaza la hipótesis nula si , y las hipótesis a contrastar en esta son:

Y de igual forma, se presentan las gráficas de autocorrelación parcial, PACF, de cada uno de los modelos en la ***figura 11*** (visible en la página diez),donde el estadístico de prueba es y que con una significancia de aproximadamente se rechaza la hipótesis nula si , y las hipótesis a contrastar en la **PACF** es:

De las figuras ***10*** para la función de autocorrelación muestral y ***11*** para la función de autocorrelación parcial muestral, se puede observar que en ninguno de los modelos se rechaza el supuesto de incorrelación sobre el proceso asociado al , ya que ninguno de estos superan las líneas azules, que son los límites de *Bartlett* para rezagos pequeños, para valores pequeños de rezagos, si bien esto sí sucede para rezagos mayores, pero que pueden ser descartados teniendo presente que la probabilidad de cometer error tipo I aumenta con el valor del rezago.

De la misma forma, se recurre a las pruebas Ljung-Box, cuyos resultados son presentados en la ***tabla 7*** para evaluar el supuesto de incorrelación sobre , de manera que, considerando , se tiene que las hipótesis son:

Y se tiene que el estadístico de prueba es y que tiene como criterio de rechazo que el valor p sea pequeño. Con esto claro, se debe tener presente que se va a realizar seis veces este test conjunto para

Así pues, usando una significancia de 5%, se tiene a la luz de los resultados que en ningún escenario se rechaza la hipótesis nula, por lo que hay evidencia muestral suficiente para sugerir que son incorrelacionados.

**Tabla 7.** Test de *Ljung-Box* para para los modelos uno, dos, tres y cuatro con .

|  |  |
| --- | --- |
|  | |
| |  |  |  |  | | --- | --- | --- | --- | | ***Tabla 8a.*** Modelo uno. | | | | | **m** | **QLB** | **Grados de libertad** | **P(χ2>QLB)** | | 6 | 1.982275 | 6 | 0.9213216 | | 12 | 9.194579 | 12 | 0.6862271 | | 18 | 16.002607 | 18 | 0.5923654 | | 24 | 28.494640 | 24 | 0.2397309 | | 30 | 32.057837 | 30 | 0.3648423 | | 36 | 42.033452 | 36 | 0.2258933 | | |  |  |  |  | | --- | --- | --- | --- | | ***Tabla 8b.*** Modelo dos. | | | | | **m** | **QLB** | **Grados de libertad** | **P(χ2>QLB)** | | 6 | 0.7928482 | 6 | 0.9922641 | | 12 | 6.3790678 | 12 | 0.8957813 | | 18 | 13.5704190 | 18 | 0.7566350 | | 24 | 23.8393628 | 24 | 0.4708133 | | 30 | 27.9345550 | 30 | 0.5739049 | | 36 | 37.1007297 | 36 | 0.4180456 | |
| |  |  |  |  | | --- | --- | --- | --- | | ***Tabla 8c.*** Modelo tres. | | | | | **m** | **QLB** | **Grados de libertad** | **P(χ2>QLB)** | | 6 | 3.137233 | 6 | 0.7914361 | | 12 | 6.694835 | 12 | 0.8771039 | | 18 | 11.050634 | 18 | 0.8921956 | | 24 | 26.252068 | 24 | 0.3405000 | | 30 | 29.341449 | 30 | 0.4997173 | | 36 | 40.612890 | 36 | 0.2742943 | | |  |  |  |  | | --- | --- | --- | --- | | ***Tabla 8d.*** Modelo cuatro. | | | | | **m** | **QLB** | **Grados de libertad** | **P(χ2>QLB)** | | 6 | 1.788685 | 6 | 0.9380714 | | 12 | 6.832079 | 12 | 0.8685080 | | 18 | 12.671327 | 18 | 0.8107167 | | 24 | 26.435448 | 24 | 0.3314466 | | 30 | 28.376683 | 30 | 0.5504829 | | 36 | 39.258612 | 36 | 0.3259529 | |

|  |  |
| --- | --- |
| Gráfico  Descripción generada automáticamente**(a)** | Gráfico  Descripción generada automáticamente con confianza baja**(b)** |
| Chart  Description automatically generated with low confidence**(c)** | Gráfico  Descripción generada automáticamente con confianza media**(d)** |

**Figura 8.** Graficas hechas con ayuda de ***R*** de los residuales contra el tiempo para cada modelo, a saber: **(a)** modelo uno: ARIMA(2,1,0)(0,1,2)[12], **(b)** modelo dos: ARIMA(4,1,0)(1,1,2)[12], **(c)** modelo tres: ARIMA(6,1,10)(0,1,1)[12] y **(d)** modelo cuatro: ARIMA(9,1,10)(0,1,1)[12]. Nótese las líneas rojas horizontales presentes en cada modelo. En el medio se marca el residual nulo y las otras dos correspondes a las delimitaciones del positivo y el negativo del doble de la varianza de los residuales asociados con cada modelo.

|  |  |
| --- | --- |
| Gráfico, Gráfico de dispersión  Descripción generada automáticamente**(a)** | Gráfico, Gráfico de dispersión  Descripción generada automáticamente**(b)** |
| Chart, scatter chart  Description automatically generated**(c)** | Chart, scatter chart  Description automatically generated**(d)** |

**Figura 9.** Graficas hechas con ayuda de ***R*** de los residuales contra el índice nominal de ventas para cada modelo, a saber: **(a)** modelo uno: ARIMA(2,1,0)(0,1,2)[12], **(b)** modelo dos: ARIMA(4,1,0)(1,1,2)[12], **(c)** modelo tres: ARIMA(6,1,10)(0,1,1)[12] y **(d)** modelo cuatro: ARIMA(9,1,10)(0,1,1)[12]. Nótese las líneas rojas horizontales presentes en cada modelo. En el medio se marca el residual nulo y las otras dos correspondes a las delimitaciones del positivo y el negativo del doble de la varianza de los residuales asociados con cada modelo.

|  |  |
| --- | --- |
| Chart  Description automatically generated**(a)** | Chart  Description automatically generated**(b)** |
| **Gráfico  Descripción generada automáticamente(c)** | Gráfico  Descripción generada automáticamente**(d)** |

**Figura 10.** Graficas hechas con ayuda de ***R*** de las funciones de autocorrelación, **ACF,** estimadas para cada modelo, a saber: **(a)** modelo uno: ARIMA(2,1,0)(0,1,2)[12], **(b)** modelo dos: ARIMA(4,1,0)(1,1,2)[12], **(c)** modelo tres: ARIMA(6,1,10)(0,1,1)[12] y **(d)** modelo cuatro: ARIMA(9,1,10)(0,1,1)[12]. Nótese las líneas rojas horizontales presentes en cada modelo, las cuales corresponden a los límites de *Bartlett* para evaluar gráficamente la significancia de la ACF para cada rezago (graficados en el eje de las abscisas).

|  |  |
| --- | --- |
| Gráfico, Gráfico de cajas y bigotes  Descripción generada automáticamente**(a)** | Gráfico  Descripción generada automáticamente**(b)** |
| Gráfico  Descripción generada automáticamente**(c)** | Gráfico  Descripción generada automáticamente**(d)** |

**Figura 11.** Graficas hechas con ayuda de ***R*** de las funciones de autocorrelación parciales, P**ACF,** estimadas para cada modelo, a saber: **(a)** modelo uno: ARIMA(2,1,0)(0,1,2)[12], **(b)** modelo dos: ARIMA(4,1,0)(1,1,2)[12], **(c)** modelo tres: ARIMA(6,1,10)(0,1,1)[12] y **(d)** modelo cuatro: ARIMA(9,1,10)(0,1,1)[12]. Nótese las líneas rojas horizontales presentes en cada modelo, las cuales corresponden a los límites de *Bartlett* para evaluar gráficamente la significancia de la PACF para cada rezago (graficados en el eje de las abscisas).

Por último, sobre la normalidad de , donde se quiere contrastar las siguientes hipótesis: se distribuye siguiendo una **normal** contra no se distribuye siguiendo una **normal**, se puede observar que en los tests de Shapiro-Wilk en la ***tabla 8***, todas las pruebas se rechazan la normalidad usando una significancia de además, en los gráficos de normalidad de la ***figura 12,*** se puede observar que los residuos de todos los modelos se da un alejamiento en la cola izquierda, lo que indica que la distribución real de los errores podría ser alguna que tenga mayor probabilidad de valores atípicos en comparación con una distribución normal.

|  |  |
| --- | --- |
| Gráfico, Gráfico de líneas  Descripción generada automáticamente**(a)** | Gráfico, Gráfico de líneas, Gráfico de dispersión  Descripción generada automáticamente**(b)** |
| **Gráfico, Gráfico de líneas  Descripción generada automáticamente(c)** | Gráfico, Gráfico de líneas  Descripción generada automáticamente**(d)** |

***Figura 12.*** Tests de *Shapiro-Wilks* para verificar normalidad de los para los modelos uno, dos, tres y cuatro.

***Tabla 8.*** Tests de *Shapiro-Wilks* para verificar normalidad de los para los modelos uno, dos, tres y cuatro.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Modelo** | **Errores estructurales** | **Estadístico W** | **Valor p** | **Validez R.B.** | **Validez normalidad** |
| Uno | ARIMA(2,1,0)(0,1,2)[12] | 0.98552 | 0.01577 | Sí | No |
| Dos | ARIMA(4,1,0)(1,1,2)[12] | 0.98415 | 0.009155 | Sí | No |
| Tres | ARIMA(6,1,10)(0,1,1)[12] | 0.98811 | 0.04529 | Sí | No |
| Cuatro | ARIMA(9,1,10)(0,1,1)[12] | 0.98572 | 0.01711 | Sí | No |
| **NA.** No aplica la evaluación del test de normalidad por no cumplirse la incorrelación. | | | | | |

De esta manera, se concluye que ninguno de los modelos propuestos cumple con el supuesto de normalidad.

1. **ANÁLISIS DE PRONÓSTICOS CON LA ESTRATEGIA DE VALIDACIÓN CRUZADA**

A continuación, se verán los pronósticos realizados para el periodo *ex post* para llevar a cabo validación cruzada, lo cual se hace teniendo en cuenta que el origen ocurre en. Además, para los intervalos de pronóstico se usará un nivel de confianza del 95% y se presentará la gráfica de los pronósticos contra los datos reales en los tiempos de pronóstico *ex post.* De esta manera, vale la pena comenzar estudiando la ***tabla 9.***

Teniendo los pronósticos que se muestran en la ***tabla 9*** para todos los modelos propuestos, vale la pena interpretar el resultado con algún periodo particular, como lo puede ser el mes de febrero de 2021, para el cual el ***modelo uno*** pronostica que el índice de ventas del sector manufacturero colombiano en pesos nominales será de 107.034 puntos y se situará entre los 99.232 y los 115.449 puntos con un nivel de confianza del 95 %. Por su parte, el ***modelo dos*** pronostica que el índice de ventas del sector manufacturero colombiano en pesos nominales de febrero de 2021 será de 107.980 y se ubicará entre los 100.005 y 116.591 puntos, mientras que el ***modelo tres*** proyecta que será de 106.799 puntos y se situará entre los 99.030 y los 112.177 puntos con un nivel de confianza del 95 %. Por último, el ***modelo cuatro*** pronostica que el índice de ventas del sector manufacturero colombiano en pesos nominales de febrero de 2021 será de 106.827 puntos y con un nivel de confianza del 95% el índice de ventas del sector manufacturero colombiano en pesos nominales estará entre 99.104 y 115.152 puntos. Ahora bien, es importante comparar los cuatro modelos a partir de la tabla de diferentes medidas de error gracias al conocimiento de los valores reales para el periodo ex post, para lo que se presenta la ***tabla 10***.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tabla 9.** Pronósticos puntuales y por I.P del 95% de confianza | | | | | | | | | | | | |
|  | Modelo 1 | | | Modelo 2 | | | Modelo 3 | | | Modelo 4 | | |
| Período | Pronóstico | Lim. Inf | Lim. Sup | Pronóstico | Lim. Inf | Lim. Sup | Pronóstico | Lim. Inf | Lim. Sup | Pronóstico | Lim. Inf | Lim. Sup |
| Dec 2020 | 125.004 | 116.960 | 133.600 | 125.912 | 117.839 | 134.538 | 125.877 | 117.754 | 134.560 | 126.074 | 117.920 | 134.791 |
| En 2021 | 104.026 | 97.088 | 111.461 | 104.564 | 97.546 | 112.086 | 104.015 | 97.091 | 111.433 | 103.913 | 97.006 | 111.313 |
| Feb 2021 | 107.034 | 99.232 | 115.449 | 107.980 | 100.005 | 116.591 | 106.799 | 99.030 | 112.177 | 106.827 | 99.104 | 115.152 |
| Mar 2021 | 113.871 | 104.628 | 123.931 | 114.277 | 104.834 | 124.570 | 113.601 | 104.401 | 123.612 | 113.495 | 104.331 | 123.464 |
| Abr 2021 | 11.305 | 101.784 | 121.716 | 112.172 | 102.520 | 122.732 | 111.057 | 101.603 | 121.390 | 111.115 | 101.716 | 121.382 |
| May 2021 | 116.444 | 105.854 | 128.093 | 116.870 | 106.057 | 120.785 | 116.142 | 105.631 | 127.700 | 115.996 | 105.574 | 127.447 |
| Jun 2021 | 114.782 | 103.765 | 126.968 | 115.432 | 104.201 | 127.874 | 114.568 | 103.388 | 126.957 | 114.580 | 103.723 | 126.575 |
| Jul 2021 | 117.292 | 105.522 | 130.375 | 116.968 | 105.097 | 130.179 | 117.014 | 105.202 | 130.152 | 117.041 | 105.471 | 129.881 |
| Ago 2021 | 119.176 | 106.690 | 133.124 | 120.196 | 107.415 | 134.498 | 119.584 | 106.894 | 133.781 | 119.632 | 107.292 | 133.392 |
| Sep 2021 | 121.367 | 108.148 | 136.201 | 121.901 | 108.461 | 137.007 | 121.101 | 107.690 | 136.181 | 121.133 | 107.884 | 136.009 |
| Oct 2021 | 123.385 | 109.461 | 139.081 | 123.752 | 109.608 | 139.720 | 123.262 | 108.991 | 139.403 | 123.295 | 109.165 | 139.255 |
| Nov 2021 | 127.263 | 112.415 | 144.073 | 127.839 | 112.729 | 144.975 | 126.865 | 111.600 | 144.219 | 126.819 | 111.640 | 144.062 |

***Tabla 10.*** Métricas para el análisis de precisión de los pronósticos puntuales y de los intervalos de predicción (I.P.) del 95 % de confianza.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Medidas | Modelo 1 | Modelo 2 | Modelo 3 | Modelo 4 |
| RMSE | 0.5962 | 0.1083 | 0.6586 | 0.6737 |
| MAE | 0.5462 | 0.1019 | 0.5886 | 0.6174 |
| MAPE (%) | 0.4683 | 0.0874 | 0.5091 | 0.5338 |
| Amplitud. Media I.P | 22.7103 | 23.1035 | 22.9409 | 22.6579 |
| Cobertura (%) I.P | 100% | 100% | 100% | 100% |

A partir de las métricas presentadas en la ***tabla 10*** se ve que el ***modelo dos*** presenta el menor valor tanto como en RMSE, MAE y MAPE. Siguiendo con esta idea, según el RMSE el modelo dos se equivocó en promedio en cada pronóstico del periodo *ex post* en 0.1083 puntos del índice de ventas del sector manufacturero colombiano en pesos nominales y según MAE este mismo modelo se equivoca en promedio 0.1019 puntos del índice, mientras que según MAPE el ***modelo dos*** erró en promedio en cada pronóstico en un 0.0874% respecto al valor real del índice de ventas del sector manufacturero colombiano en pesos nominales. En general los cuatro modelos presentan buenos resultados en cuanto pronóstico según las medidas mencionadas anteriormente.

Luego, en cuanto a los intervalos de predicción, se ve que todos contienen el valor real de la serie para cada uno de los periodos *ex post*, en este caso el modelo con un intervalo de predicción más estrecho es el ***modelo cuatro*** seguido por el ***modelo uno***. Para ver más fácilmente la calidad de la predicción se presenta en la figura 1 las predicciones puntuales de cada modelo y los valores reales del índice de ventas para el periodo ex post

|  |
| --- |
| Chart, line chart  Description automatically generated |
| ***Figura 13.*** Comparación de los pronósticos. |

Podemos concluir con la ***figura 13*** que no hay diferencias prácticas de importancia entre los pronósticos puntuales y los valores reales. Como estamos interesados en escoger el modelo que mejor pronostica sin dejar de lado la validez de supuestos de dicho modelo, y teniendo en cuenta que no hay diferencias importantes, se preferirán los modelos más parsimoniosos, ya que ninguno de los modelos paso el supuesto de que sus errores tienen una distribución normal, nos quedamos con el modelo uno por las razones comentadas anteriormente.

1. **COMPARACIÓN CON MODELOS GLOBALES, LOCARLES Y ARMA(p, q)×(P, Q)[12] Y CONCLUSIONES**
   1. ***Conclusiones relacionadas con los modelos SARIMA.***

A partir del análisis descriptivo se evidencia que la serie tiene tanto componente de tendencia como componente de estacionalidad. Adicionalmente, su varianza no es constante y aumenta en la misma dirección de la tendencia, por lo que se procede a modelar el logaritmo de la serie, con lo que se consigue que la varianza se estabilice, pero la serie sigue sin ser estacionaria por que aún se mantiene la componente de tendencia y la componente estacional. Estas mismas observaciones se reflejan en la respectiva gráfica ACF, y por tanto es pertinente analizar el comportamiento del logaritmo natural de la serie cuando se le aplica la primera diferencia regular, aunque con esto se logra eliminar de la tendencia, en la ACF asociada se tiene un decaimiento lento para, es decir, indica necesidad de aplicar también la diferencia estacional. Con lo anterior se concluye que este proceso no es estacionario en sentido débil. Entonces, al aplicar solo la primera diferencia estacional sobre el logaritmo de la serie se obtiene que aunque para la gráfica ACF en los múltiplos de 12 ya no se encuentran correlaciones significativas, no se puede considerar el proceso como estacionario en covarianza porque se encontraron evidencias en contra de los supuestos media constante y homocedasticidad; después, al aplicar el filtro mixto (primera diferencia regular y primera diferencia estacional), se puede comentar de la gráfica ACF que a pesar de que se encuentran tanto para la parte regular como para la parte estacional autocorrelaciones muestrales estadísticamente significativas, se pueden considerar ambas partes como ergódicas, y adicionalmente no se encuentran evidencias gráficas en contra de los supuestos de homocedasticidad y media constante, y así, se concluye que el logaritmo natural de la serie diferenciado por tendencia y estacionalidad es un proceso ergódico, lo cual muestra además que ya no hay evidencia de existencia de raíces unitarias regulares y estacionales, por lo que no es necesario diferenciar más por tendencia o estacionalidad.

A continuación, según el test HEGY, la serie transformada usando logaritmo natural tiene tanto raíz unitaria regular como estacional, luego, es apropiado diferenciar regular y estacionalmente a esta serie, y con ello se confirma lo evaluado con los análisis descriptivos que se realizaron justo antes en cuanto a la necesidad de tomar la primera diferencia regular y estacional para el logaritmo natural del índice de ventas del sector manufacturero colombiano en pesos nominales.

Ahora, ante la necesidad de ajustar modelos SARIMA, se procede con la identificación de los patrones de las gráficas ACF y PACF para la serie diferenciada por tendencia y estacionalidad. De este modo, para la parte regular se identifica un patrón tipo cola exponencial sinusoidal mientras que la PACF se identifica un patrón tipo corte donde el último rezago estadísticamente diferente de cero es el segundo, por lo que para la parte regular se identifica un AR(2), lo que implica que para la parte regular de se tiene un ARIMA(2,1,0); hecho esto, se prosigue con el análisis de la parte estacional, y comenzando con la ACF es posible identificar un patrón tipo corte con último rezago estacional significativo en y en la gráfica PACF un patrón tipo cola por lo que los modelos propuestos deben cumplir la condición de que la parte estacional se modela con un MA(2)[12] y en el caso de un ARMA(0,1,2)[12]. Entonces, mezclando ambos análisis se propone el ***modelo uno:*** ARIMA(2,1,0)(0,1,2)[12] para .

A continuación, pasando a la identificación de modelos SARIMA(p, d, q)(P, D, Q)[s] con ayuda de métodos automáticos, con el método auto.arima(), solo se consideraron aquellos que contemplaban tanto la diferencia regular como la estacional, y también se descartó aquel modelo que proponía un proceso de medias móviles para la parte regular debido a que no concuerda con las evidencias gráficas que se mencionaron anteriormente, por lo que el modelo candidato resultante de auto.arima() es: ***modelo 2****:* ARIMA(4,1,0)(1,1,2)[12] sin deriva para .

Pasando a la identificación con armasubsets con el primer renglón del tablero 1212 sobre usando el método “ols” y agregando al parámetro se obtiene el ***modelo 3*:** ARIMA(6, 1, 10)(0, 1, 1)[12], con únicos parámetros no nulos . Ahora con el primer renglón del tablero 18 × 18 de armasubsets sobre usando el método “ols” se identifica el **modelo 4. ARIMA(9, 1, 10)(0, 1, 1)[12]** que solo emplea los términos

Pasando al ajuste de estos modelos se puede comentar de la gráfica de la serie real y el modelo ajustado que se observa que sigue adecuadamente la tendencia y la estacionalidad presente en la serie; además, para los cuatro modelos no se observan diferencias significativas en cuanto al ajuste, pero, aun así, no está de más comentar que el ***modelo uno*** es el que cuenta con un menor AIC y BIC.

Asimismo, considerando la validación de los supuestos bajo los cuales fueron ajustados los modelos, para ninguno de los cuatro se encontraron evidencias gráficas en contra de que sus medias fueran iguales a cero, las varianzas sean constantes o que los errores sean un proceso ergódico. Con el test Ljung-Box se confirmó que no hay evidencia suficiente en contra de que los errores sean incorrelacionados, y como para todos los modelos se obtuvo que los errores eran un proceso de ruido, luego se pasó a probar si para todos los modelos los errores son un proceso ruido blanco mediante el chequeo del supuesto de normalidad. De esta forma, con ayuda de los gráficos cuantil-cuantil se rechazó el supuesto de normalidad para los cuatro modelos, concluyendo así que ningún modelo SARIMA tiene errores ruido blanco. Con esto en mente, se escoge como mejor modelo al ***modelo uno***, ya que en términos de pronósticos no existen diferencias prácticas de importancia entre los modelos y además este es el modelo que contiene un menor número de parámetros. Por último, es importante comentar que, ya que el ***modelo uno*** no es un modelo válido en términos de cumplimiento de supuestos, sus pronósticos no son confiables.

* 1. ***Conclusiones asociadas a todos los modelos***

Ahora vale la pena hacer un repaso sobre algunos de los modelos obtenidos a lo largo de los tres trabajos: mejores modelos local y global de regresión global del trabajo uno, mejor modelo de regresión global con errores ARMA en el trabajo dos, mejor modelo SARIMA del trabajo tres, cuyas ecuaciones son presentadas en la ***tabla 11.***

|  |
| --- |
| ***Tabla 11*.** Ecuaciones de los mejores modelos ajustados. |
| **Mejor modelo global.** Logpolinomial de grado seis estacional con funciones trigonométricas en cinco frecuencias Fj =j/12 , j = 1, 2, 3, 4, 5. |
| **Mejor modelo local.** Descomposición multiplicativa y *loess* lineal  En la vecindad de un tiempo donde se quiere el ajuste  con para todo t en la vecindad de , y  con parámetros de la recta local. |
| **Mejor modelo con errores ARMA .**  Logpolinomial de grado seis estacional con funciones trigonométricas en cinco frecuencias Fj =j/12 , j = 1, 2, 3, 4, 5; con error estructural ARMA: ARMA(12,10) con 𝜙7 y 𝜃10.  donde |
| **Mejor modelo SARIMA** ARIMA(2, 1, 0)(0, 1, 2)[12]  **,** |

Hecho esto, en la ***figura 14*** se presentan las gráficas que comparan la serie en su escala original comparadas con los modelos ajustados. Así pues, es importante rescatar que para los mejores modelos global, ARMA y SARIMA se consigue en general seguir de forma adecuada la dinámica de la serie, lo cual se refleja en el hecho que las respectivas líneas rojas, asociadas a los modelos ajustados, se apegan de forma aceptable a la línea negra, correspondiente a los datos reales de la serie temporal; empero, para el mejor modelo global se puede observar que, si bien logra seguir bien la dinámica de la serie en cuanto a su tendencia y estacionalidad, cuando se tienen ciclos el ajuste no es adecuado, como se puede verificar por ejemplo en el año 2009, cuando el modelo sobreestima los índices de ventas del sector manufacturero colombiano en pesos nominales para los periodos de dicho año. Este análisis se verifica claramente al observar las métricas de ajuste con los criterios de información de Akaike y bayesiano presentados en la ***tabla 12,*** pues este par de métricas son máximas para el mejor modelo global, dejando claro entonces que es el modelo que peores ajustes consigue a la luz de los índices realmente observados. Ahora bien, en cuanto a los otros tres modelos, se tiene que según el AIC el modelo que mejores ajustes realiza es el modelo local de descomposición multiplicativa y *loess* lineal, y los mejores modelos ARMA y SARIMA tienen AIC semejantes y no muy distantes entre sí, y ambos separados una unidad aproximadamente del AIC del mejor modelo local; no obstante, con el criterio BIC el análisis cambia en cuanto el modelo con el menor valor de criterio de información bayesiano es ahora el modelo SARIMA y con una mayor distancia respecto al segundo y tercer modelo en orden de acuerdo con este criterio que en el caso del AIC. Así pues, considerando que prefiere al criterio de información bayesiano, se concluye que el modelo que mejores ajustes realiza es el modelo SARIMA, que corresponde al ***modelo uno*** presentado en este trabajo.

En cuanto a los pronósticos puntuales de estos modelos, se ve en la ***figura 15*** que los pronósticos del modelo global son los que más se alejan de los valores reales que toma la serie en el periodo *ex post*, pero el resto de modelos alcanzan un conjunto de predicciones aceptable, pero destaca la buena predicción del modelo SARIMA para el periodo *ex post*. Para respaldar estas evidencias gráficas se recurre a los criterios para la precisión de los pronósticos puntuales y de los intervalos de predicción al 95% de confianza que se muestran en la ***tabla 13***, donde, a excepción del modelo global, los demás modelos tienen valores similares y resulta interesante que el modelo local es el cuenta con menor MAE y MAPE, mientras que el modelo SARIMA presenta el menor RMSE. A propósito de los intervalos de predicción, es importante tener en cuenta que por la forma en que se ajustaron el mejor modelo local y el mejor modelo ARMA, no es posible obtener para estos intervalos de predicción, sino pronósticos puntuales únicamente, pero de todos modos, entre el mejor modelo local y el mejor modelo SARIMA, vale la pena rescatar que ambos cubren el 100 % de los índices de ventas del sector manufacturero colombiano en pesos nominales para los meses del periodo *ex post,* pero se diferencian porque el mejor modelo SARIMA permite obtener intervalos de predicción más angostos, lo cual es claramente más beneficioso. Así, se puede determinar que el mejor modelo en cuanto a los pronósticos es el SARIMA ya que tiene el menor RMSE, su MAE y su MAPE no difieren mucho de los del mejor modelo local y además permite obtener intervalos de predicción que permiten tener una mejor idea del rango de valores que puede tomar el índice de interés.

|  |  |
| --- | --- |
| Chart, histogram  Description automatically generated  (a) | Chart, histogram, scatter chart  Description automatically generated  (b) |
| Chart, line chart, histogram  Description automatically generated  (c) | Chart, histogram, scatter chart  Description automatically generated  (d) |

***Figura 14.*** Gráficas de los ajustes (a) en mejor modelo global (modelo uno del trabajo uno); (b) en mejor modelo local (modelo cuatro del trabajo uno); (c) en mejor modelo con errores ARMA(modelo cuatro del trabajo dos); (d) en modelo 2b.

**Tabla 12.** Valores criterios AIC y BIC

|  |  |  |
| --- | --- | --- |
| Medidas | AIC | BIC |
| Mejor modelo global | 9.11703 | 11.674705 |
| Mejor modelo local | 4.241348 | 7.265074 |
| Mejor modelo con errores ARMA | 5.303542 | 7.629496 |
| Mejor modelo SARIMA | 5.514306 | 5.844665 |

|  |
| --- |
| Chart, line chart  Description automatically generated |

**Figura 15.** Comparación de los pronósticos de los mejores modelos con respecto a los índices realmente observados en el periodo *ex post*

*.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tabla 13.** Precisión de los pronósticos puntuales y de los I.P del 95% | | | | |
| Medidas | **Mejor modelo global** | Mejor modelo local | Mejor modelo con errores ARMA | Mejor modelo SARIMA |
| RMSE | 2.906349 | 0.6728501 | 0.7601474 | 0.5962 |
| MAE | 2.510166 | 0.5283713 | 0.6555441 | 0.5462 |
| MAPE (%) | 2.108491 | 0.4439505 | 0.5586861 | 0.4683 |
| Amplitud. Media I.P | 26.42722 | NA | NA | 22.7103 |
| Cobertura (%) I.P | 100% | NA | NA | 100% |

Por último, se va a hacer un recuento del cumplimiento de los supuestos bajo los cuales se construyen todos los modelos en los tres trabajos, que estriban básicamente en el hecho de que sus errores sean procesos de ruido blanco, lo cual implica que sean estacionarios en sus medias sean nulas, sean homocedásticos e independientes entre sí y que se distribuyan siguiendo una normal. Para ello, se presenta la ***tabla 14.***

***Tabla 14.*** Resumen del cumplimiento del supuesto de ruido blanco para los mejores modelos ajustados.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Modelo | Supuestos  sobre | | ¿Hay evidencia fuerte en contra de? | | | | | | | | | | Modelo  Válido | |
| Media cero | | Homocedasticidad | | Independencia | | | Normalidad | | |
|  |  |  | Sí | No | Sí | No | Sí | No | *NA* | Sí | No | *NA* | Si | No |
| Mejor modelo global | ✘ |  | ✘ |  | ✘ |  |  |  | ✘ |  |  | ✘ |  | ✘ |
| Mejor modelo local | ✘ |  |  | ✘ |  | ✘ | ✘ |  |  |  |  | ✘ |  | ✘ |
| Mejor modelo con errores ARMA |  | ✘ |  | ✘ |  | ✘ |  | ✘ |  | ✘ |  |  |  | ✘ |
| Mejor modelo SARIMA | ✘ |  |  | ✘ |  | ✘ |  | ✘ |  | ✘ |  |  |  | ✘ |

De esta manera, se debe iniciar mencionando que aunque el modelo global hubiera cumplido con los supuestos aún no sería recomendable para modelar la serie temporal asociada le índice de ventas del sector manufacturero colombiano en pesos nominales, ya que los modelos globales asumen que las relaciones establecidas por el modelo son estables en el tiempo lo que no es un hecho viable para índices como, puesto que las relaciones comerciales, industriales y económicas evolucionan con el paso del tiempo. Así pues, una solución a esto serían los modelos locales, que son precisamente aquellos que admiten cambios en los parámetros; aunque un modelo local nos proporciona buenos pronósticos dado que estos buscan la última estimación que se ajustó, esto es peligroso ya que nada asegura que el último patrón local y el patrón futura van a coincidir.

En cuanto a los problemas enfrentados en la modelación, se puede empezar comentando que para el modelo global fue relativamente fácil ajustar el modelo ya que desde los enunciados se especificaba el orden del polinomio que modelaba la tendencia, pero determinar dicho orden no es una tarea fácil ya que para llegar a él se requiere una serie de pruebas que comparen diferentes órdenes para determinar cuál resulta más apropiado. Por el lado del modelo local, afortunadamente se cuentan con métodos automáticos que ayudan a proponer dichos modelos ya que encontrar un parámetro de suavizamiento es una tarea exhaustiva. Por su lado, en el modelo con errores ARMA, además de que se debe encontrar cómo representar tanto tendencia como estacionalidad (trabajo que ya se había adelantado con el modelo global), se debe definir cómo modelar el error, lo que tal vez puede representar problemas ya que no siempre es fácil identificar los órdenes de y sin la ayuda de métodos automáticos. Este último inconveniente también lo comparte el modelo SARIMA, y en este caso identificar patrones y órdenes en gráficas ACF y PACF llega a ser más retador aún por la interacción que se da entre la parte regular y la parte estacional, además que se limita a pocos rezagos (lo cual es muy complejo para modelos trimestrales para el análisis asociado a la parte regular).

Mencionado todo lo anterior, se puede concluir que los modelos ajustados, si bien son adecuados siguiendo la dinámica de la serie (con las salvedades explicadas para el mejor modelo global), todos fallan en el cumplimiento de los supuestos sobre los errores, pero como nunca se van a encontrar modelos perfectos, no está demás comentar que **el modelo SARIMA arrojó unos buenos pronósticos a corto plazo y aunque la confiabilidad de estos no está asegurada sobre todo para los intervalos de predicción, no es una mala idea considerar dicho modelo.**

**BIBLIOGRAFÍA**

Álvarez, N. G. (2022). *Notas de Clase Estadística III*.

1. Estudiante de Estadística, Universidad Nacional de Colombia – Sede Medellín. [↑](#footnote-ref-1)
2. Estudiante de Estadística, Universidad Nacional de Colombia – Sede Medellín [↑](#footnote-ref-2)
3. Estudiante de Estadística, Universidad Nacional de Colombia – Sede Medellín [↑](#footnote-ref-3)