MiniProyecto 03

Juan Camilo Vargas Velez, Samuel Pinzón Valderrutén, Sebastian Diaz Noguera

15 de noviembre de 2023

1. Ejemplo Pasajeros

Para abordar este primer punto, se ha seleccionado un conjunto de datos que abarca el recuento mensual de pasajeros de una aerolínea estadounidense en el período comprendido entre los años 1949 y 1960, conocido como *Air Passengers* [Chi]. El objetivo principal es utilizar este conjunto de datos para realizar pronósticos sobre el número de pasajeros para los próximos cuatro períodos.

1.1. Análisis exploratorio de datos

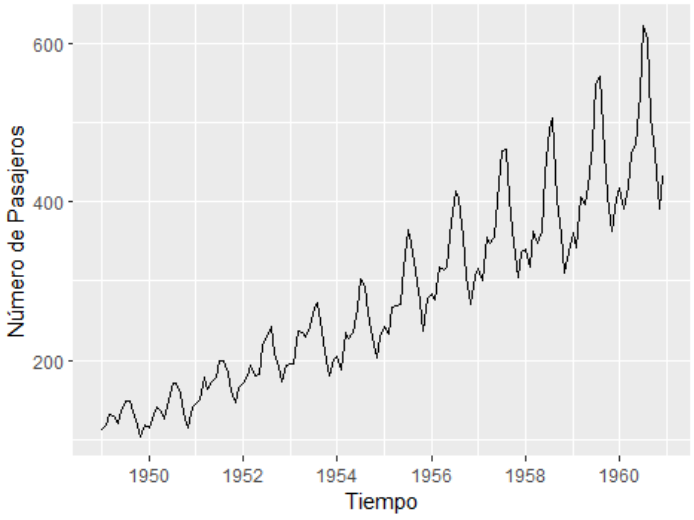

Figura 1: Pasajeros en USA: 1949 - 1961

En la Figura 1, se puede observar un incremento constante en el número de pasajeros a lo largo de los años, lo que sugiere una tendencia positiva evidente. Sin embargo, más allá de esta tendencia general, se aprecia la presencia de patrones estacionales en la serie de datos. Estos patrones estacionales se manifiestan en la mitad de cada año, indicando que hay períodos específicos en los cuales el número de pasajeros tiende a experimentar aumentos y disminuciones de manera recurrente.

La estacionalidad al final de cada año podría deberse a diversos factores, como eventos estacionales, festividades, períodos vacacionales o patrones de comportamiento relacionados con la industria de la aviación.

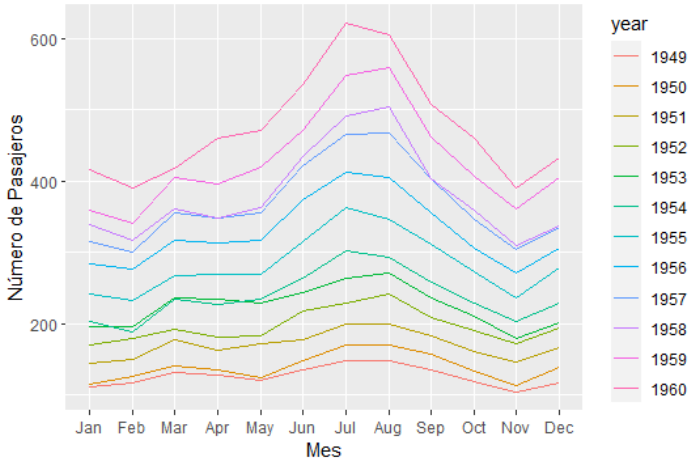

Figura 2: Gráfico Estacional de Pasajeros

En la Figura 2, se puede observar el aumento progresivo en el número de pasajeros conforme pasan los años, destacándose los picos que se evidencian en cada uno de ellos, especialmente en los meses de junio, julio y agosto. Este comportamiento se puede atribuir a la temporada de verano y al atractivo de destinos turísticos, parques de atracciones y lugares de vacaciones.

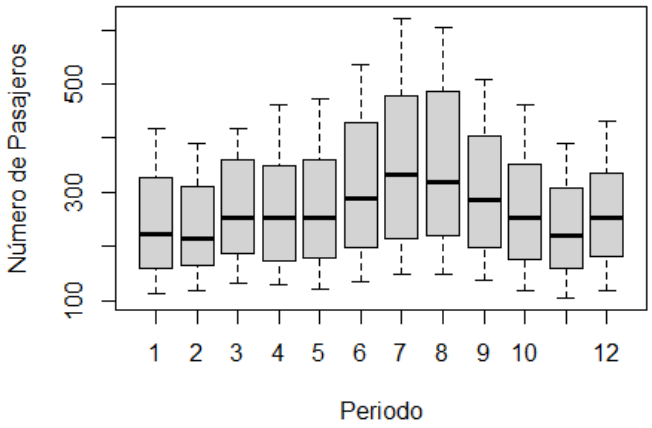

Figura 3: Boxplot de Pasajeros por Periodo

En el Boxplot de la Figura 3, se presenta una representación visual del número de pasajeros divididos en subseries de la serie de tiempo de la aerolínea. Se puede observar que en ninguno de los periodos hay valores atípicos. Por otro lado, los tamaños de las cajas en todos los meses del año sugieren que hay dispersión en los datos. Esta dispersión puede atribuirse a la tendencia ascendente que revela un cambio sistemático en la cantidad de pasajeros a lo largo del tiempo.

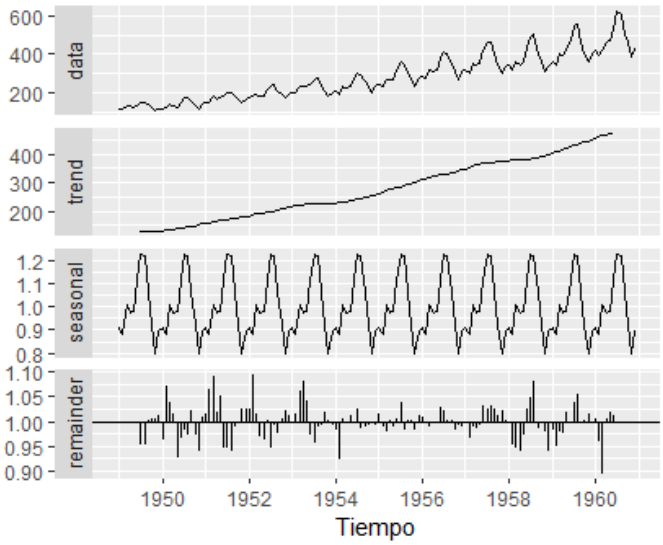

Figura 4: Descomposición multiplicativa del modelo

En la Figura 4 se presenta la descomposición de la serie temporal de pasajeros de la aerolínea. El proceso de descomposición de series temporales es una metodología analítica que separa la serie en distintos componentes, con el propósito de comprender mejor los patrones y las tendencias subyacentes a lo largo del tiempo.

Para la descomposición de esta serie temporal, se ha elegido una descomposición multiplicativa. Esta técnica se fundamenta en la variabilidad proporcional de las fluctuaciones estacionales con respecto a la magnitud total de la serie. En el proceso de descomposición, se identifican tres componentes principales en la serie temporal de pasajeros de la aerolínea.

- **Tendencia:** Se destaca la presencia de una tendencia caracterizada por un aumento constante en el número de pasajeros a lo largo del tiempo.
- **Estacionalidad:** Se observa una estacionalidad que resalta el aumento de viajes durante las vacaciones de verano.
- **Variabilidad:** La variabilidad en los datos, representada por la dispersión alrededor de la media, muestra un incremento a medida que la media global crece, indicando una mayor variación en la demanda de pasajeros a lo largo del periodo analizado.

1.2. Medias móviles

Para el análisis, se decidió implementar medias móviles con un periodo de 4 meses. La elección se fundamenta en la premisa de que un promedio móvil con un "k"(número de períodos) pequeño tiende a capturar los movimientos de la serie temporal a corto plazo, lo cual resulta apropiado para nuestro caso, donde el enfoque del pronóstico se centra exclusivamente en los próximos 4 meses.

El orden de la media móvil determina la suavidad de la estimación de la tendencia-ciclo. En términos generales, un orden más grande implica una curva más suave.

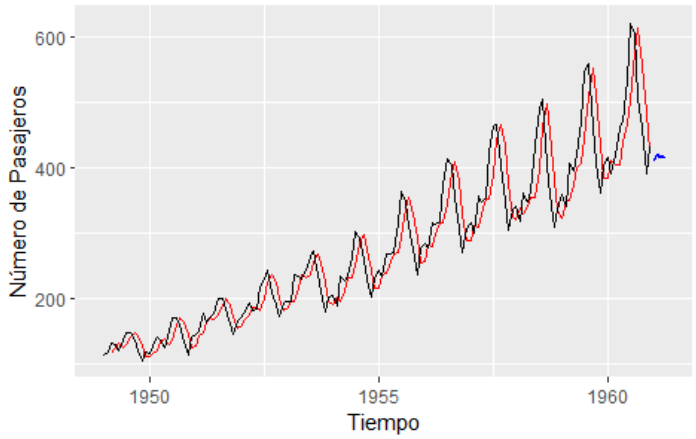

Figura 5: Pronóstico Medias móviles

En la Figura 5, se aprecia que el modelo intenta ajustarse a la serie temporal, capturando adecuadamente la tendencia-ciclo. No obstante, se observa un ligero desfase, sugiriendo que el modelo parece haberse movido un periodo adelante en comparación con la serie temporal real.

1.3. Suavización exponencial simple

Para determinar el valor óptimo de α, se llevó a cabo una búsqueda de valores en un rango específico, calculando el Root Mean Square Error (RMSE) para cada uno de ellos. El valor de α que minimizó el RMSE se seleccionó para llevar a cabo la estimación mediante el método de Suavizado Exponencial Simple (SES).

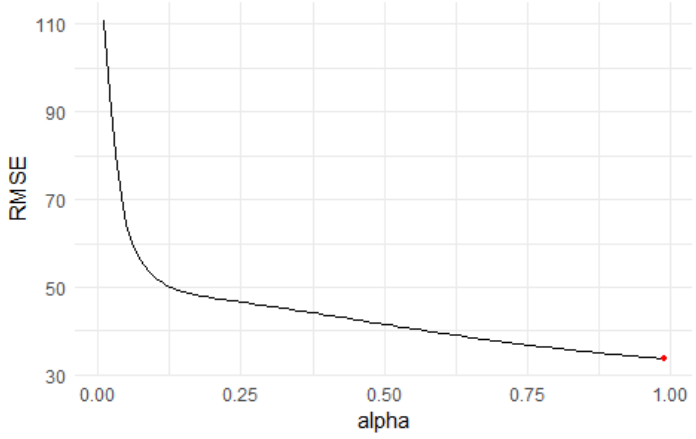

Figura 6: RMSE vs. Alpha

Como se puede observar en la Figura 6, en este caso, el valor de α que minimiza el Root Mean Square Error (RMSE) es $\alpha = \mathbf{0.99}$, lo que implica un aprendizaje rápido en los movimientos diarios.

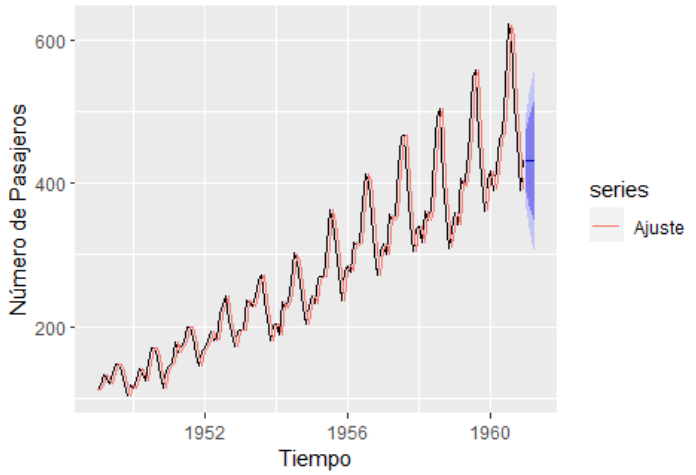

Figura 7: Pronóstico Suavización Exponencial

En la Figura 7, a pesar de la clara tendencia en los datos, se observa que el Suavizado Exponencial Simple (SES), que normalmente se utiliza en situaciones sin tendencia ni estacionalidad, captura adecuadamente la tendencia-ciclo. El desfase que se observaba anteriormente en las medias móviles es menor, y los intervalos de confianza predichos intentan ajustarse a la tendencia.

1.4. Holt

El Método de Holt realiza predicciones para datos con una tendencia utilizando dos parámetros de suavizado, α y β, los cuales corresponden a los componentes de nivel y tendencia, respectivamente.

Por otro lado, se llevó a cabo una búsqueda para encontrar el valor óptimo de β que minimice el método de Holt. Para ello, se utilizó el valor de α encontrado en SES y se estableció un rango para β de 0.0001 a 0.5. El valor de α que minimizó el RMSE se empleó para realizar la estimación de Holt.

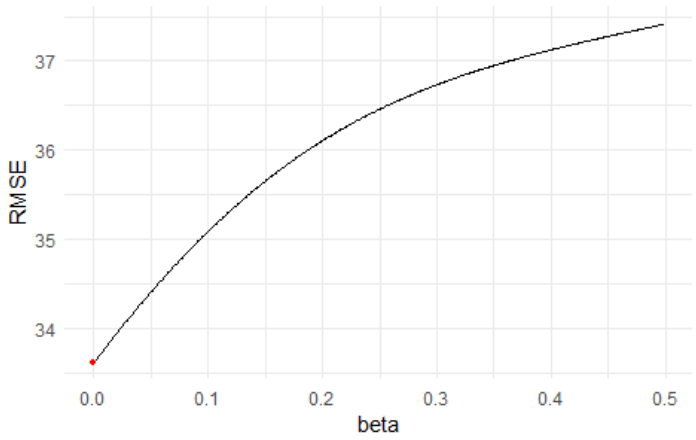

Figura 8: RMSE vs. Beta

Al observar la Figura 8, en este caso, el valor de β que minimiza el RMSE para el método de Holt es **1e-04 (0.0001)**, lo que lo convierte en el valor óptimo para realizar la estimación Holt. En este escenario, $\alpha = \mathbf{0.99}$, indicando un aprendizaje rápido en los movimientos diarios, y $\beta = \mathbf{0.0001}$, lo que significa un aprendizaje lento para la tendencia.

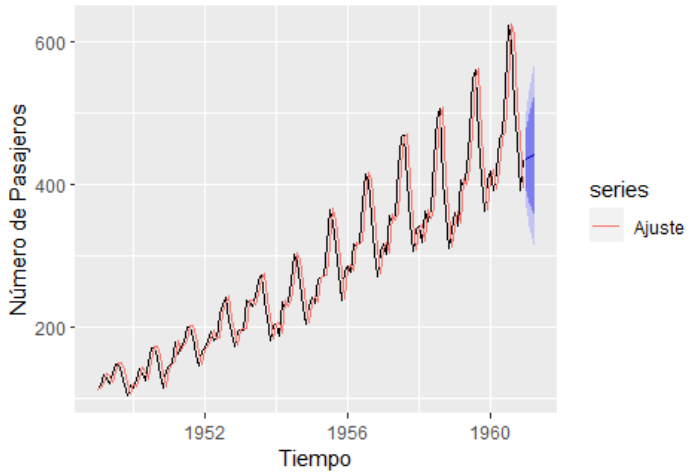

Figura 9: Pronóstico Holt

En la Figura 9, se puede observar que el método de Holt realiza una estimación que se acerca a los valores observados en la serie original. Además, se nota una mejora en la sincronización de los intervalos de confianza predichos con la tendencia.

1.5. Holt-Winters

Para el método de Holt-Winters, además de buscar los valores de α y β, también es necesario determinar el valor de γ. El parámetro γ se utiliza para suavizar la estacionalidad en una serie de tiempo y abordar aquellas que exhiben tanto tendencia como estacionalidad. Este método puede implementarse con una estructura Aditiva o una estructura Multiplicativa, y la elección del método depende de las características del conjunto de datos.

En este caso, en lugar de buscar el valor óptimo de γ, se aplicó una nueva función llamada `ets()`, que representa error, tendencia y estacionalidad. La selección del parámetro del modelo, denotado como `model =` , es crucial para comprender el modelo `ets`. Al especificar el tipo de modelo, siempre se indica el error, la tendencia y luego la estacionalidad, de ahí el término ets. En nuestro caso, utilizamos `model = "MAM"`, como observamos en la descomposición, especificando que la tendencia es aditiva.

Los valores de α, β y γ para el Método Holt-Winters son $\alpha = \mathbf{0.7096}$, $\beta = \mathbf{0.0204}$ y $\gamma = \mathbf{1e\text{-}04}$.

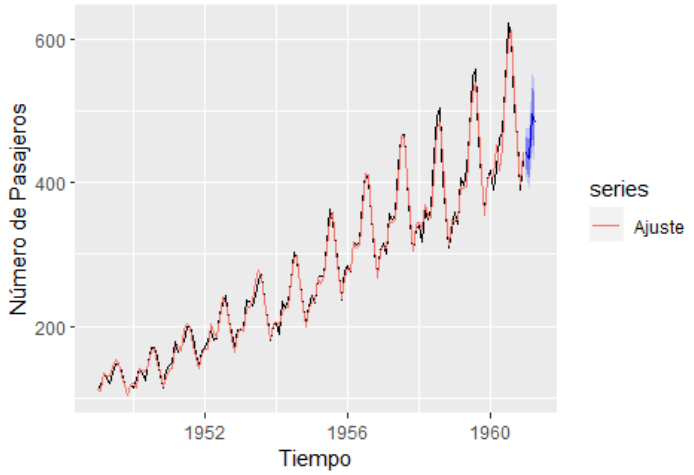

Figura 10: Pronóstico Holt-Winter

Se puede observar en la Figura 10 que el modelo se ajusta casi exactamente al original, y los intervalos de confianza predichos siguen adecuadamente la tendencia-ciclo de la serie. Esto sugiere que el método de Holt-Winters realiza una estimación altamente precisa de los valores previamente observados en la serie original.

1.6. Comparación de modelos

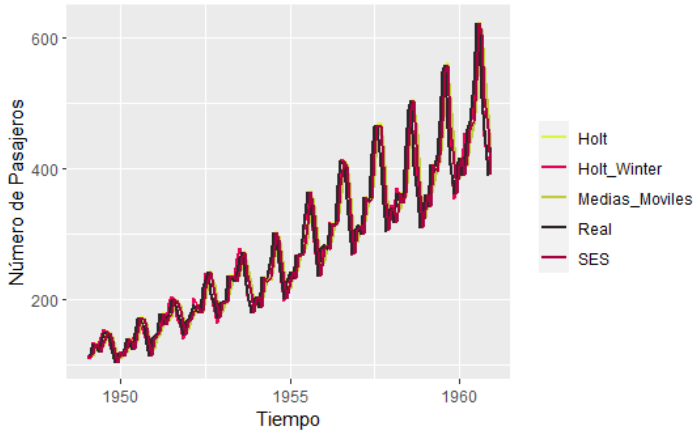

Figura 11: Comparativa de Modelos

En la Figura 11 se puede observar el ajuste de los modelos de Medias Móviles, Suavización Exponencial Simple, Holt y Holt-Winters.

Cuadro 1: Comparativa de Rendimiento

	Modelo	RMSE	MAE	MAPE
1	Medias Móviles	42.15	31.98	11.02
2	Suavización Exponencial Simple	33.70	25.75	8.98
3	Holt	33.62	25.56	8.96
4	Holt-Winters	10.75	7.79	2.86

En la Tabla 1, se presenta una comparativa de rendimiento entre diferentes modelos de pronóstico. Se observa que el modelo Holt-Winters muestra el menor RMSE, MAE y MAPE, sugiriendo que es el modelo más adecuado para este conjunto de datos.

1.7. Pronóstico

Cuadro 2: Pronóstico de 4 Períodos

	Mes	Medias móviles	Suavización Exponencial Simple	Holt	Holt-Winter
1	1961-01	411.00	431.59	433.84	441.80
2	1961-02	421.50	431.59	436.08	434.12
3	1961-03	416.25	431.59	438.31	496.63
4	1961-04	418.88	431.59	440.55	483.24

La Tabla 2 muestra las predicciones para los próximos cuatro períodos utilizando diferentes modelos. Dado el rendimiento superior del modelo Holt-Winters según la Tabla 1, se prefirió utilizar este modelo para realizar las predicciones.

De acuerdo con estas predicciones, se estima que el número de pasajeros para los siguientes meses será de aproximadamente:

- **Enero de 1961** $(1961-01)$**:** 442 pasajeros.
- **Febrero de 1961** $(1961-02)$**:** 435 pasajeros.
- **Marzo de 1961** $(1961-03)$**:** 497 pasajeros.
- **Abril de 1961** $(1961-04)$**:** 484 pasajeros.

Es importante tener en cuenta que, en la práctica, los pasajeros se cuentan en números enteros, por lo que se ajustaron estas cifras para reflejar valores más realistas.

2. Ejemplo Viviendas

En este segundo punto, se dispone de una serie temporal mensual que abarca desde 1959 hasta 2020, y se centra en las nuevas unidades de vivienda de propiedad privada. El objetivo principal es emplear este conjunto de datos para realizar pronósticos sobre el número de viviendas nuevas.

2.1. Análisis exploratorio de datos

En la Figura 12, se puede observar que a lo largo de los años, la serie exhibe una variabilidad significativa en el número de viviendas nuevas, sin mostrar una tendencia clara. Entre 1959 y 1990, se identifican patrones estacionales, los cuales podrían atribuirse a la alta volatilidad de los precios de vivienda a nivel mundial. Este fenómeno puede ser explicado por eventos como guerras, caídas en la economía de diversos países y la implementación de nuevos modelos económicos, como la introducción del euro.

Por otro lado, es relevante destacar la marcada tendencia negativa que se manifestó en el año 2008. Este declive significativo se atribuye a la burbuja inmobiliaria que se presentó a nivel mundial, resultando en un aumento considerable en los precios de las viviendas.

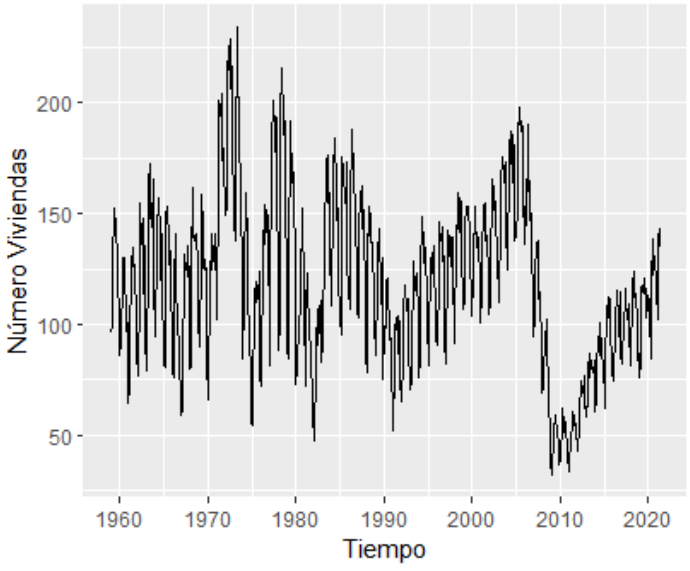

Figura 12: Nuevas Viviendas Mensuales

En el Boxplot de la Figura 13 se puede observar el número de viviendas nuevas por mes divididos en subseries de la serie de tiempo, además, se pueden observar los valores atípicos los cuales resaltan más en los meses de febrero, agosto y octubre. En cambio, las dimensiones de las cajas en todos los meses del año indican una variabilidad, la cual se atribuye a la tendencia creciente que revela una modificación en la cantidad de viviendas.

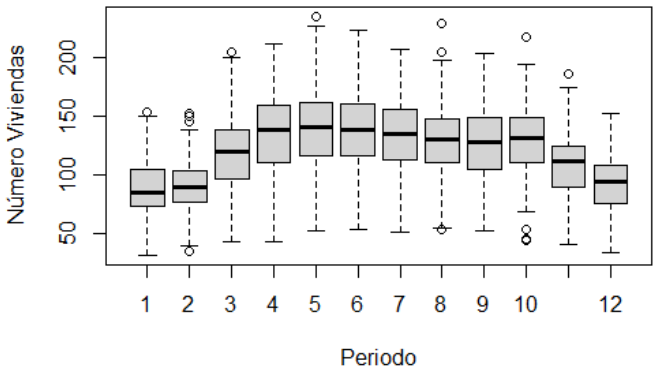

Figura 13: Boxplot del número de viviendas por período

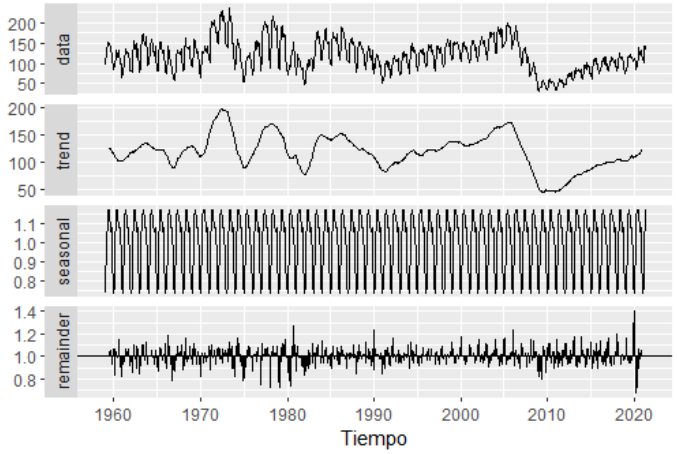

Figura 14: Descomposición Multiplicativa del Modelo

En la Figura 14 se presenta la descomposición multiplicativa de la serie temporal del número de viviendas. En este análisis, se ha elegido la descomposición multiplicativa debido a su capacidad para separar las componentes de la serie temporal, como la tendencia, la estacionalidad y los residuos, de manera multiplicativa. Esta elección se justifica por la variabilidad no constante en la varianza a medida que la serie temporal aumenta. La descomposición multiplicativa es más adecuada para abordar estas variaciones.

2.2. Identificación del modelo

Para la identificación del modelo ideal, se llevó a cabo una evaluación de la estacionariedad de la serie de tiempo. Para este propósito, se empleó el Dickey-Fuller Test, una prueba estadística utilizada para examinar la presencia de raíces unitarias en una serie temporal univariada. El objetivo fundamental de esta prueba es determinar si la serie exhibe una tendencia que la hace no estacionaria.

Las hipótesis planteadas son las siguientes:

$$H_0 : \text{No es estacionaria}$$

$$H_a : \text{Es estacionaria}$$

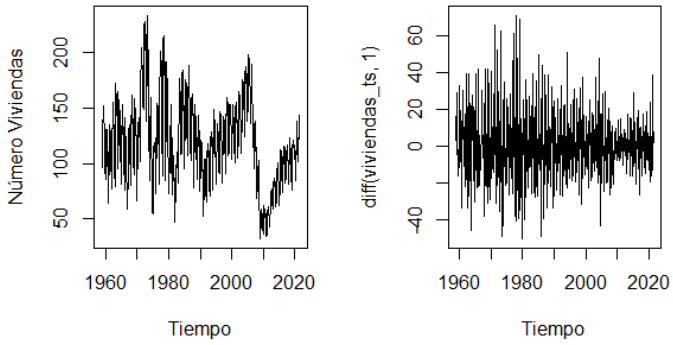

Figura 15: Comparación Serie Original vs. Primera Diferencia del Número de Viviendas

Al realizar la prueba, se obtuvo un valor p-value de 0.54. Al evaluar este valor, se concluye que no hay suficiente evidencia para rechazar la hipótesis nula (H_0), lo que sugiere que la serie temporal no es estacionaria.

Dado que no se puede asumir la estacionariedad de la serie original, se procedió a realizar la primera diferenciación para transformar la serie y lograr la estacionariedad necesaria. Esta acción tiene como objetivo facilitar la identificación del modelo apropiado.

En la Figura 15, se presenta una comparación entre la serie original (a la derecha) y la primera diferenciación (a la izquierda).

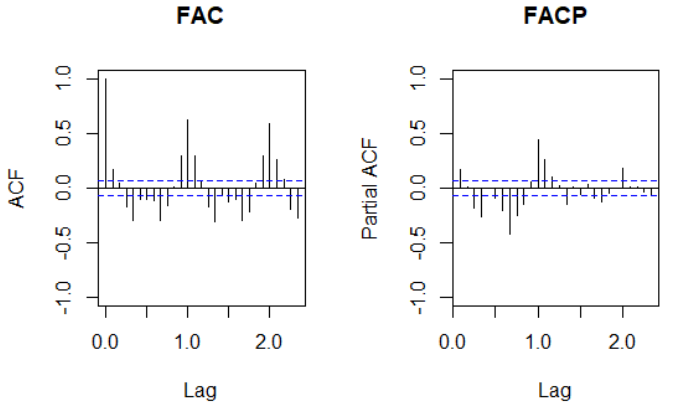

Figura 16: ACF y PACF de la Diff1 del Número de Viviendas

Después de realizar la diferenciación, se obtuvo un *p-value* de 0.01, indicando así que la serie es estacionaria. Sin embargo, al examinar la Figura 16, se aprecia que tanto la función de autocorrelación (ACF) como la función de autocorrelación parcial (PACF) no muestran una disminución clara en sus valores. Esto se debe a los picos que sugieren la persistencia de estacionalidad en los datos. Por lo tanto, para identificar el modelo más apropiado que se ajuste a los datos, es esencial eliminar esta estacionalidad.

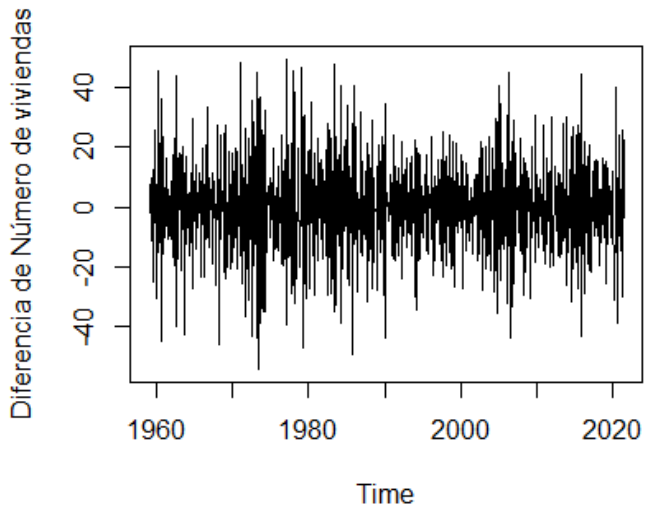

Figura 17: Diferenciación Estacional del Número de Viviendas

En la Figura 17 se observa la Diferenciación Estacional de la serie. Para realizarla, se eliminó el componente estacional y además se realizó la primera diferenciación. Al realizar toda esta diferenciación, se obtuvo un *p-value* de 0.01, indicando así que la serie es estacionaria.

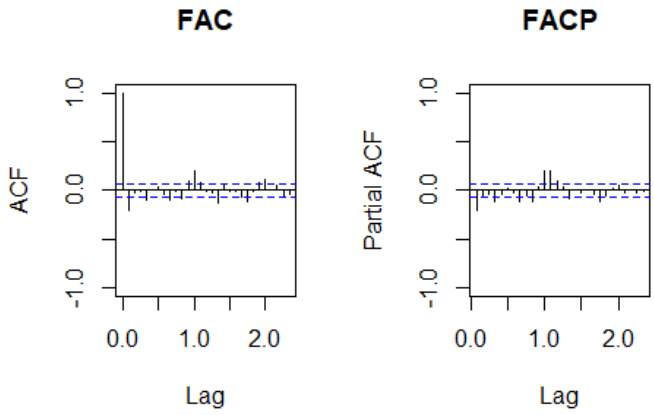

Figura 18: ACF y PACF de la Diferenciación Estacional del Número de Viviendas

En la Figura 18, tanto la ACF como la PACF muestran una tendencia a decrecer rápidamente, indicando la posible presencia de un proceso mixto, es decir, con componente autorregresiva y de medias móviles. Tras el análisis del ACF y PACF, se señala la posibilidad de dos posibles modelos SARIMA: ARIMA(1,1,1)$(1,0,0)_{12}$ y ARIMA(2,1,1)$(1,0,0)_{12}$.

Cuadro 3: Comparativa de Rendimiento

	Modelo	AIC	BIC
1	ARIMA(1,1,1)$(1,0,0)_{12}$	5916.11	5934.58
2	ARIMA(2,1,1)$(1,0,0)_{12}$	5890.05	5913.14

De acuerdo a la Tabla 3, que muestra la comparación de los modelos con AIC y BIC, se determinó el modelo óptimo. Considerando 'd' como 1, se identificó 'p' igual a 2 según el análisis del ACF y 'q'

igual a 1 basado en el PACF. Tras eliminar la estacionalidad, 'P' se estableció en 1, lo que llevó a la selección final del modelo ARIMA(2,1,1)$(1,0,0)_{12}$.

2.3. Validación de Supuestos

Para comprobar la validez del modelo, se realizó una evaluación de la estacionariedad de los residuos del modelo utilizando la prueba de Dickey-Fuller.

Al realizar esta prueba, se obtuvo un valor de p-value $= 0.01$, lo que lleva al rechazo de la hipótesis nula (H_0), sugiriendo que los residuos son estacionarios. Este resultado indica que los residuos no presentan autocorrelación.

Además, para validar el supuesto de normalidad en los residuales se utilizó el Shapiro-Wilk normality test, donde las hipótesis son:

$$H_0 : \text{No se distribuyen normal}$$

$$H_a : \text{Se distribuyen normal}$$

Al realizar esta prueba, el p-value $= 0.1007$, por lo que no se rechazó la hipótesis nula, indicando que los residuos se distribuyen de manera normal.

Al validar los supuestos de estacionaridad y normalidad en los residuales, se concluye que el modelo ARIMA seleccionado es un ajuste adecuado.

2.4. Pronóstico

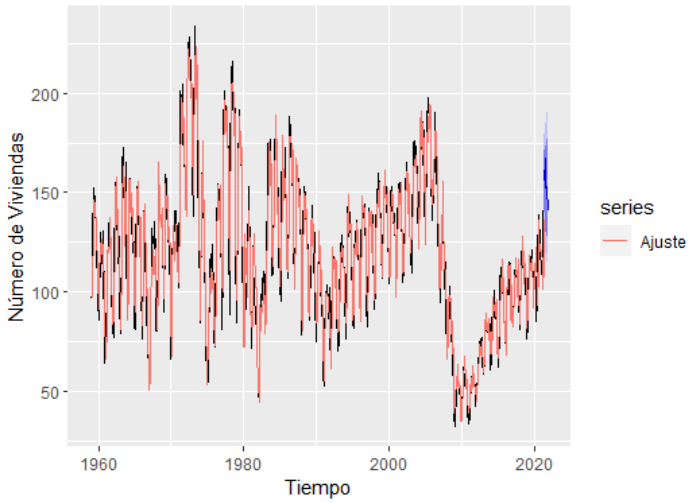

Figura 19: Pronóstico del ARIMA(2,1,1)$(1,0,0)_{12}$

Cuadro 4: Pronóstico de los próximos 6 meses del modelo ARIMA(2,1,1)$(1,0,0)_{12}$

	Mes	# Viviendas	Lo 80	Hi 80	Lo 95	Hi 95
1	2021-06	155.63	139.94	171.32	131.63	179.62
2	2021-07	167.03	148.34	185.72	138.44	195.62
3	2021-08	152.08	130.74	173.42	119.44	184.72
4	2021-09	152.09	128.90	175.28	116.62	187.55
5	2021-10	153.14	128.50	177.77	115.46	190.81
6	2021-11	141.08	115.31	166.84	101.67	180.48

La Tabla 4 presenta el pronóstico del número de viviendas para los próximos 6 meses utilizando el modelo ARIMA$(2,1,1)(1,0,0)_{12}$. En esta tabla, se encuentran los valores pronosticados junto con intervalos de confianza del 80 % y del 95 %. Por ejemplo, para el mes de junio de 2021 se pronostica un número de viviendas de 155.63, con un intervalo del 80 % entre 139.94 y 171.32, y un intervalo del 95 % entre 131.63 y 179.62. Este patrón se repite para los meses subsiguientes.

3. Ejemplo cambio porcentual del consumo personal

En este tercer punto, se cuenta con una serie temporal trimestral que abarca desde 1960 hasta 2016, enfocada en el cambio porcentual del consumo personal. El objetivo principal es utilizar este conjunto de datos para realizar pronósticos sobre dichos cambios a lo largo del tiempo.

3.1. Análisis exploratorio de datos

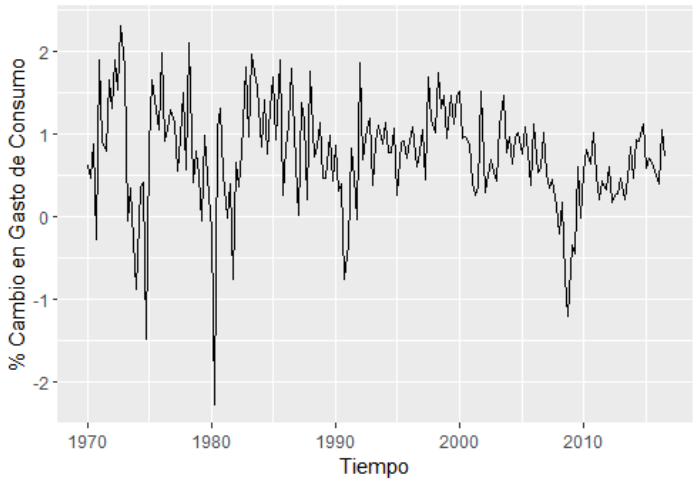

Figura 20: Cambio Porcentual Trimestral en el Gasto de Consumo Personal (1960-2016)

La Figura 20 muestra picos de tendencias negativas en los años 1975, 1980, 1990 y 2008, los cuales fueron originados por factores económicos clave. En 1975, una marcada recesión afectó el poder adquisitivo, provocando una disminución en el gasto. En 1980, la restricción crediticia y la alta inflación contribuyeron a otra tendencia negativa. Durante la década de 1990, se experimentó una reducción en el gasto debido a crisis financieras regionales y cambios en las condiciones económicas globales. En 2008, la crisis financiera mundial y la recesión generalizada impactaron significativamente en el comportamiento de gasto de los consumidores.

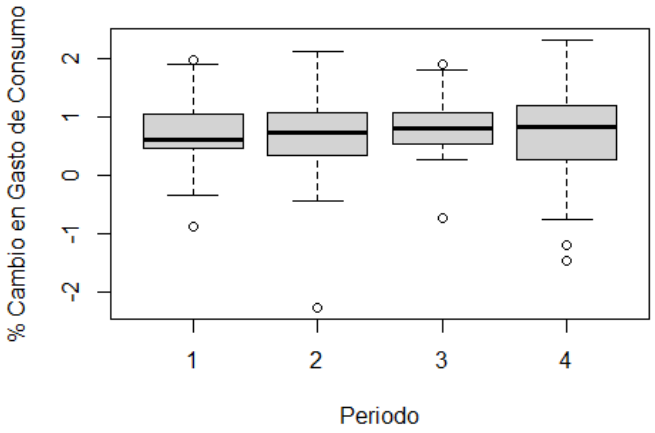

Figura 21: Boxplot del % Cambio en Gasto de Consumo por Período

En la Figura 21 se puede observar el gasto de consumo personal por trimestre en subseries de la serie de tiempo, por otro lado, en ella se pueden observar la existencia de datos atipicos los cuales se resaltan mas en los trimestres 1,3 y 4. Adicionalmente las dimensiones de las cajas en los trimestres 1,2 y 4 indican una variabilidad en los datos, el cual puede atribuirse a los cambios bruscos de tendencias positivas a negativas.

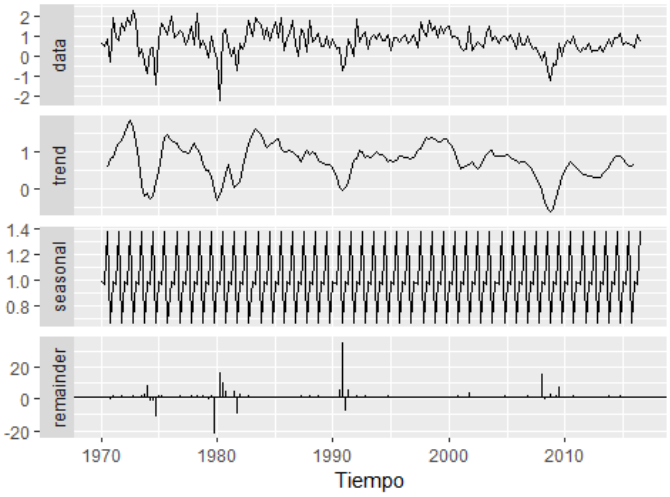

Figura 22: Descomposición Multiplicativa del Modelo

En la Figura 22 se ha optado por utilizar una descomposición multiplicativa al analizar esta serie temporal. La selección de esta opción se fundamenta en la variabilidad no constante en la varianza a medida que la serie temporal se extiende. La descomposición multiplicativa resulta más apropiada para manejar estas variaciones

3.2. Identificación del modelo

Para la identificación del modelo adecuado, se utilizó la misma metodología que en el punto anterior. Se llevó a cabo una evaluación de la estacionariedad de la serie mediante la prueba de Dickey-Fuller.

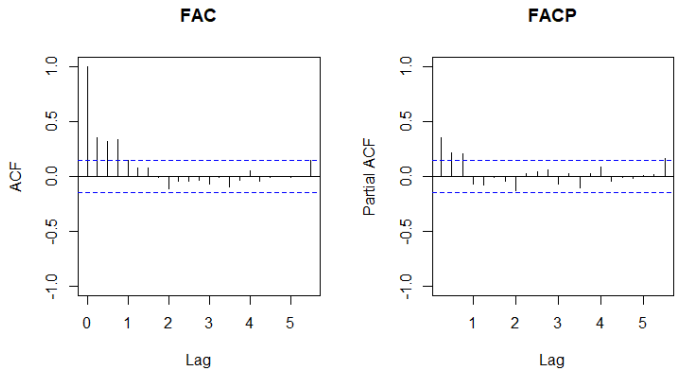

Figura 23: ACF y PACF Cambio Porcentual en Gasto de Consumo

Al realizar la prueba, se obtuvo un valor p-value de 0.01, llevando a la conclusión de que hay evidencia para rechazar la hipótesis nula (H_0). Esto sugiere que la serie temporal es estacionaria, por lo que no era necesario realizar ninguna diferenciación.

En la Figura 23, tanto la ACF como la PACF muestran una tendencia a decrecer rápidamente, indicando la posible presencia de un proceso mixto, es decir, con componente autorregresiva y de medias móviles. Tras el análisis del ACF y PACF, se señala la posibilidad de cinco modelos ARIMA: ARIMA(1,0,0), ARIMA(0,0,1), ARIMA(0,0,2), ARIMA(1,0,1) y ARIMA(1,0,2).

Cuadro 5: Comparativa de Rendimiento

	Modelo	AIC	BIC
1	ARIMA(1,0,0)	353.33	363.02
2	ARIMA(0,0,1)	360.37	370.06
3	ARIMA(0,0,2)	356.80	369.72
4	ARIMA(1,0,1)	343.82	356.74
5	ARIMA(1,0,2)	343.33	359.49

La Tabla 5 compara los modelos utilizando los criterios AIC y BIC. Se seleccionó el modelo ARIMA(1,0,1) como el más óptimo, ya que presenta el menor valor en BIC y un valor relativamente bajo en AIC.

3.3. Validación de Supuestos

Para comprobar la validez del modelo, se realizó una evaluación de la estacionariedad de los residuos del modelo utilizando la prueba de Dickey-Fuller.

Al realizar esta prueba, se obtuvo un valor de p-value = 0.01, lo que lleva al rechazo de la hipótesis nula (H_0), sugiriendo que los residuos son estacionarios. Este resultado indica que los residuos no presentan autocorrelación.

Además, para validar el supuesto de normalidad en los residuales se utilizó el Shapiro-Wilk normality test.

Al realizar esta prueba, el p-value = 7.948e-05, por lo que se rechazó la hipótesis nula, indicando que los residuos no se distribuyen de manera normal.

Al no validarse los supuestos de estacionaridad y normalidad en los residuales, se concluye que el modelo ARIMA seleccionado no es un ajuste adecuado.

3.4. Pronóstico

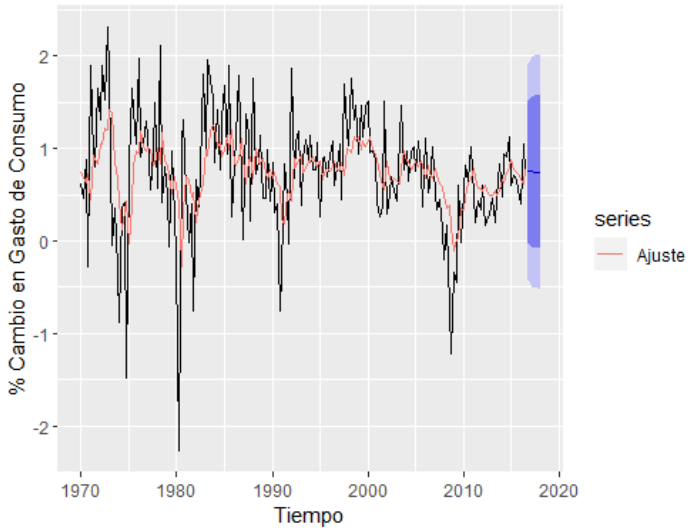

Figura 24: Pronóstico del ARIMA(1,0,1)

En la Figura 24, se puede observar que, aunque los intervalos de confianza predichos intentan seguir el comportamiento de la serie, el modelo no se ajusta muy bien. En muchas partes, el modelo se queda corto, y sería mejor si pudiera estirarse para mejorar la precisión de las predicciones.

Cuadro 6: Pronóstico de los próximos 6 trimestres del modelo ARIMA(1,0,1)

	Trimestre	% Cambio en Gasto de Consumo	Lo 80	Hi 80	Lo 95	Hi 95
1	2016 Q4	0.75	-0.01	1.51	-0.41	1.91
2	2017 Q1	0.75	-0.04	1.54	-0.46	1.96
3	2017 Q2	0.75	-0.06	1.56	-0.49	1.98
4	2017 Q3	0.75	-0.07	1.57	-0.50	2.00
5	2017 Q4	0.75	-0.08	1.57	-0.52	2.01
6	2018 Q1	0.75	-0.08	1.58	-0.52	2.02

La Tabla 6 muestra el pronóstico del % Cambio en el Gasto de Consumo para los próximos 6 trimestres utilizando el modelo ARIMA(1,0,1). Se presentan los valores pronosticados junto con intervalos de confianza del 80 % y del 95 %. Por ejemplo, para el cuarto trimestre de 2016, se pronostica un % Cambio en el Gasto de Consumo de 0.75 %, con un intervalo del 80 % entre -0.01 y 1.51, y un intervalo del 95 % entre -0.41 y 1.91. Este patrón se repite para los trimestres subsiguientes.

Referencias

[afi13] Exponential smoothing - afit data science lab r programming guide, 2013.

[Chi] Chirag19. Air passengers dataset. https://www.kaggle.com/datasets/chirag19/air-passengers.

[fpp19] 9.9 seasonal arima models — forecasting: Principles and practice (3rd ed), 2019.

[fpp19] [afi13]