Análisis Temporal de la Matrícula en Educación Primaria a Nivel Mundial (1970–2023)

Salvador Enrique Rodríguez Hernández (rh06006)

2025-06-15

Índice

1 Introducción

El presente informe tiene como objetivo analizar la evolución histórica de la matrícula en educación primaria a nivel mundial durante el período comprendido entre 1970 y 2023. Para ello, se utiliza un conjunto de datos proporcionado por el **UNESCO Institute for Statistics (UIS)**, disponible públicamente a través del portal de datos del **Banco Mundial** (https://data.worldbank.org/indicator/SE.PRM.ENRL), bajo la licencia **CC BY-4.0**.

El indicador analizado corresponde al número total de alumnos inscritos en el nivel primario, tanto en instituciones públicas como privadas. Los datos se recopilan de los informes oficiales que los países miembros remiten a la UNESCO mediante su encuesta anual de educación. Además, se han armonizado de acuerdo con la **Clasificación Internacional Normalizada de la Educación (CINE)** para garantizar la comparabilidad internacional.

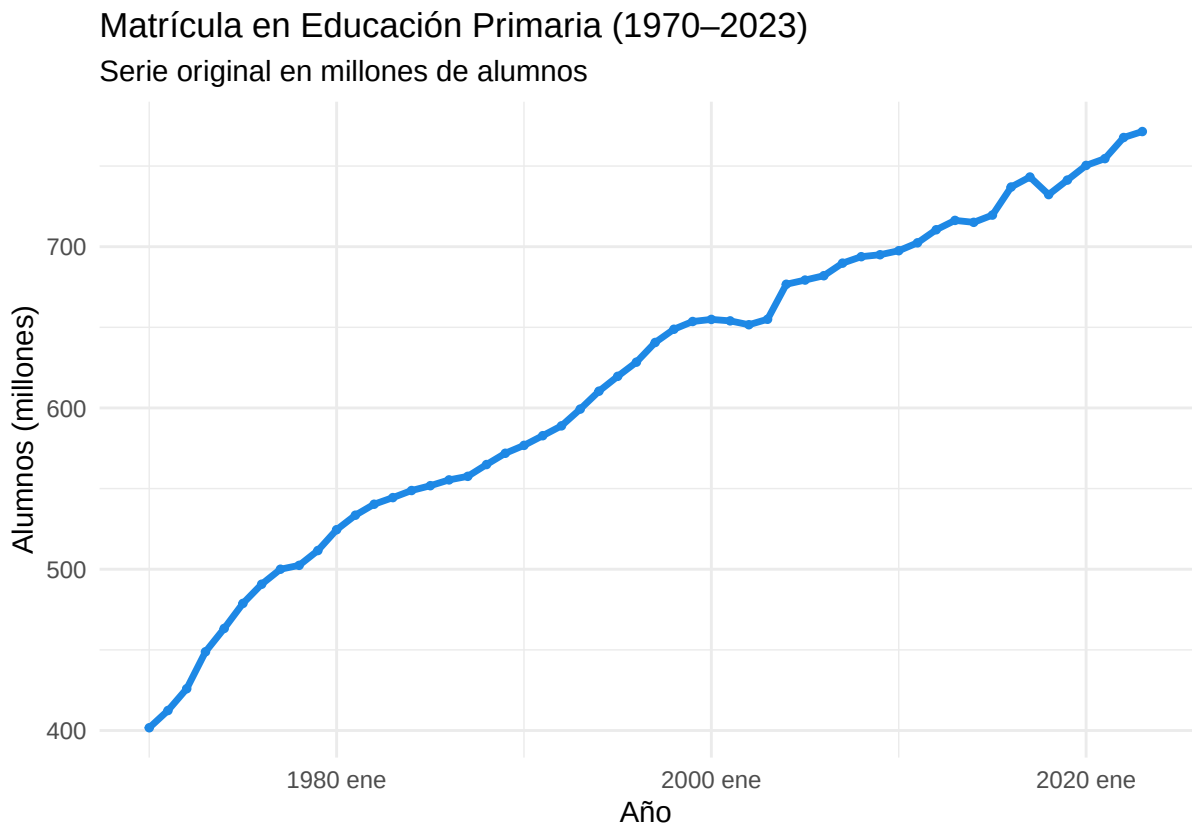
La frecuencia de los datos es anual, y los valores están expresados en millones de alumnos. Este análisis busca identificar patrones de crecimiento, posibles cambios estructurales en la serie, y proponer un modelo ARIMA que permita realizar proyecciones confiables de la matrícula escolar en los próximos años.

2 Análisis gráfico de la serie

En esta sección se **analiza gráficamente** la evolución de la matrícula con el objetivo de detectar la presencia de tendencia, variaciones inusuales y cambios en la dispersión. Estas observaciones servirán de base para determinar las transformaciones necesarias *para estabilizar la serie y asegurar la validez de los modelos estadísticos aplicados posteriormente.*

2.1 Gráfica de la serie

El examen visual constituye el primer paso para entender la dinámica de la matrícula. Se observa la forma general, la posible presencia de tendencia, cambios de variabilidad y cualquier indicio de estacionalidad.



2.2 Análisis de tendencia, variabilidad, estacionalidad etc

Antes de aplicar cualquier transformación conviene verificar si la serie en **nivel** cumple los supuestos básicos de los modelos lineales de series temporales:
(1) *estacionariedad* (media y varianza constantes) y (2) *homocedasticidad* (varianza constante en el tiempo).
Para ello se combinan dos pruebas complementarias de raíz unitaria —**ADF** y **KPSS**— y la prueba **ARCH** de Engle; además se estima el parámetro de **Box–Cox** ( ) como indicador de dependencia de la varianza con el nivel.

Cuadro 1: Pruebas de estacionariedad y homocedasticidad (serie en nivel)

Prueba	Estadistico	p_valor	Decision
ADF	-3.177	0.1006	No rechazar H0
KPSS	1.437	0.01	Rechazar H0
ARCH	41.820	0	Heterocedasticidad
Box-Cox lambda	2.000	NA	

Con base en los resultados estadísticos presentados en la Tabla 1, se concluye que la serie

en nivel **no cumple con el supuesto de estacionariedad** para el modelado de series temporales, ya que su media crece con el tiempo y presenta **varianza no constante**.A continuación se presentan con más detalles los resultados de la Tabla~1.

- **Estacionariedad**
  - **ADF**: $p = 0.1006 > 0.05$ No se rechaza la hipótesis nula de raíz unitaria.
  - **KPSS**: $p = 0.01 < 0.05$ Se rechaza la hipótesis nula de estacionariedad. Ambos resultados apuntan a que la serie **no es estacionaria**.
- **Homocedasticidad**
  - **ARCH (12 rezagos)**: estadístico 41.82 con $p \approx 0$ Se detecta **heterocedasticidad**.
  - **Box-Cox**: $\lambda \approx 2$. Un valor tan alejado de 0 indica que transformaciones logarítmicas (o raíz) difícilmente estabilizarían la varianza.

En resumen, la serie original **presenta raíz unitaria** y **varianza no constante**. Tener una raíz unitaria significa que la serie se comporta como un paseo aleatorio: cada perturbación desplaza permanentemente su nivel, la varianza se expande con el tiempo y la media ya no permanece fija.

3 Transformación para estacionariedad

El análisis previo reveló que la serie en su forma original **no es estacionaria** y presenta **heterocedasticidad**. Trabajar con una serie que incumplan estos supuestos puede llevar a estimaciones sesgadas y predicciones poco fiables; por lo tanto, es imprescindible aplicar una transformación que estabilice la media y la varianza en un nivel sostenido a lo largo del tiempo.

3.1 Justificación del tipo de transformación aplicada

Para decidir el tratamiento adecuado se consideraron cuatro contrastes complementarios (la Tabla 1 recoge los valores obtenidos):

- **ADF** (Augmented Dickey–Fuller)
  $p = 0.1006 > 0.05 \rightarrow$ no se rechaza la presencia de raíz unitaria.
- **KPSS** (nivel)
  $p = 0.0100 < 0.05 \rightarrow$ se rechaza la estacionariedad.
- **ARCH** (12 rezagos)
  $p \approx 0 \rightarrow$ se detecta heterocedasticidad (varianza no constante).
- **Box-Cox**
  $\lambda \approx 2 \rightarrow$ transformaciones logarítmicas o de raíz no estabilizarían la varianza.

La combinación *ADF no rechaza + KPSS rechaza* confirma la existencia de una **raíz unitaria**, propia de un proceso tipo "paseo aleatorio": las perturbaciones tienen un efecto permanente, la varianza crece con el tiempo y la media no es estable. Al mismo tiempo, el test ARCH muestra que la varianza cambia de forma sistemática y el valor de λ sugiere que **las transformaciones de potencia estándar no son suficientes**.

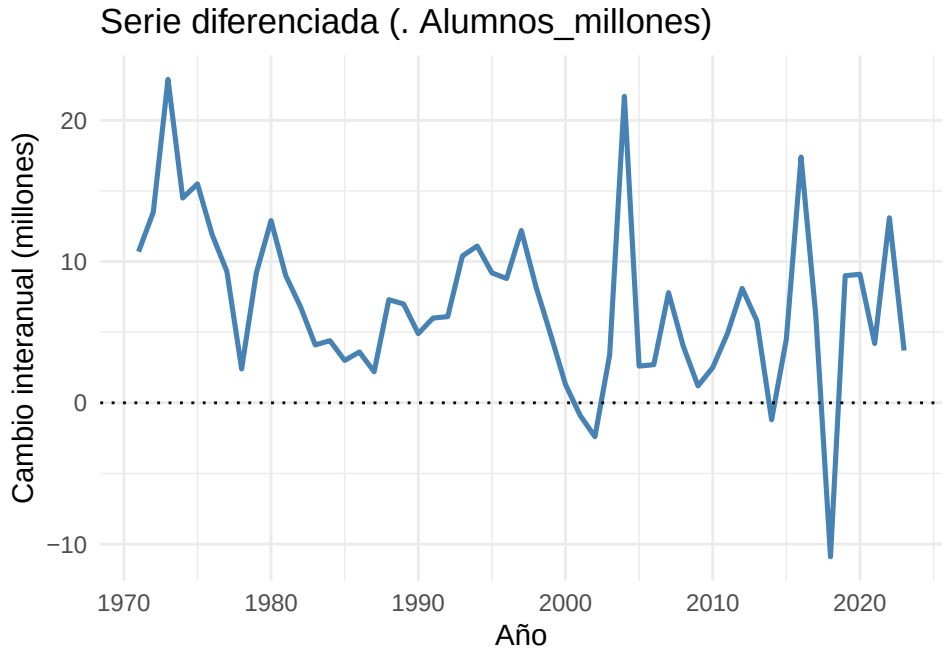
La forma más directa de afrontar ambos problemas —raíz unitaria y heterocedasticidad— es aplicar **una diferencia de primer orden**. La diferenciación elimina la raíz unitaria (resta sucesiva $Y_t - Y_{t-1}$) y, en la práctica, suele amortiguar los cambios de varianza en series anuales como la presente. En la próxima subsección se comprueba empíricamente si esa única diferencia basta o si es necesario un segundo nivel de diferenciación.

3.2 Proceso para convertir la serie en estacionaria

Se aplica la primera diferencia y se repiten las pruebas:

Cuadro 2: Pruebas tras la primera diferencia

Prueba	p_valor	Decision
ADF sobre ΔY	0.0345	Rechazar H0
KPSS sobre ΔY	0.0488	Rechazar H0
ARCH sobre ΔY	0.9086	Homocedasticidad



Tras la primera diferencia, el test ADF rechaza la presencia de raíz unitaria, mientras que el KPSS arroja un p-valor de 0.0488, justo en el umbral del 5 %. Si bien este valor permite rechazar la hipótesis nula de estacionariedad bajo un criterio del 5 %, no constituye

una evidencia contundente, y con un nivel de significancia más exigente, como el 1 %, no se rechazaría dicha hipótesis. Por su parte, el test ARCH no detecta heterocedasticidad. Visualmente, la serie diferenciada fluctúa de forma estable alrededor de su constante, con picos pasajeros que regresan rápidamente a la media. Estos resultados, tanto estadísticos como gráficos, confirman que una sola diferencia es suficiente; aplicar una segunda sólo añadiría complejidad sin mejorar sustancialmente la estacionariedad de la serie.

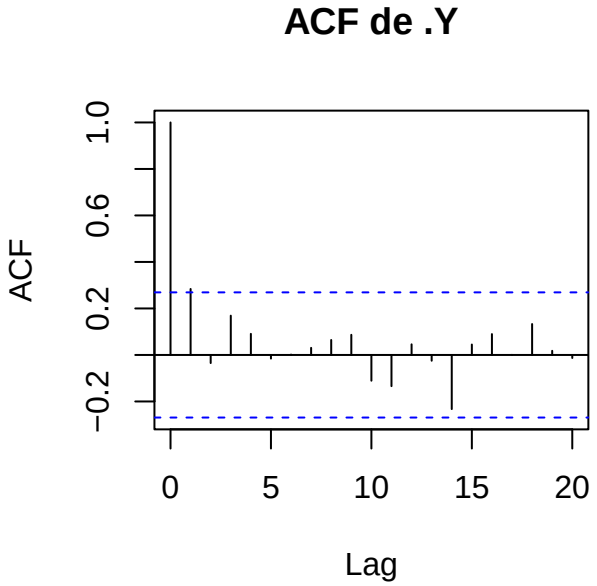
4 Análisis de autocorrelación

El objetivo de este apartado es analizar si los incrementos anuales, una vez eliminada la tendencia, conservan algún tipo de relación temporal. Para ello se utilizan la **Función de autocorrelación simple (ACF)** y la **Función de autocorrelación parcial (PACF)**

Ambas gráficas se construyen a partir de la serie ya diferenciada (ΔY), la cual se ha comprobado que es estacionaria tanto en media como en varianza, según las pruebas previas y el análisis visual.

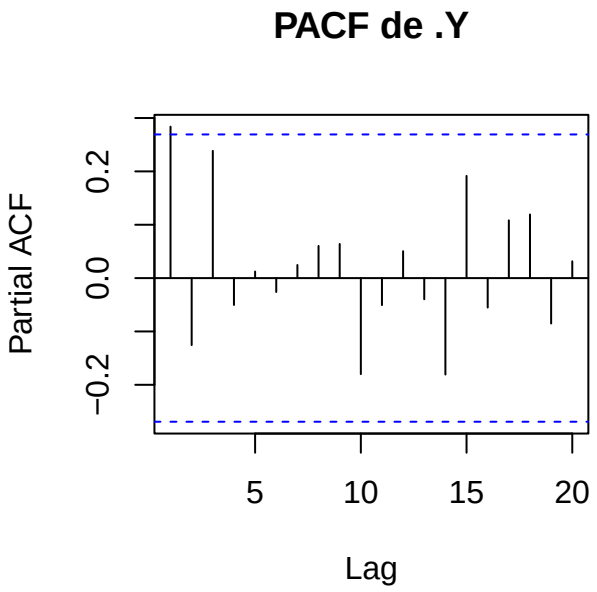
4.1 Gráfica de la función de autocorrelación simple (ACF)

La ACF muestra el grado de relación entre los valores de la serie y sus retardos (valores pasados).



4.2 Gráfica de la función de autocorrelación parcial (PACF)

La PACF muestra la relación directa entre los valores de la serie y sus retardos, eliminando el efecto de los retardos intermedios.



4.3 Análisis de los resultados

En la ACF se observa un pico dominante en el rezago 1, seguido de valores que disminuyen rápidamente y se mantienen dentro de las bandas de confianza, lo que sugiere un patrón típico de un proceso MA(1). La PACF también muestra un único pico significativo en el rezago 1 y luego valores sin importancia, lo que respalda la idea de que la dependencia está concentrada en ese primer rezago. Esta combinación —ACF que cae rápidamente y PACF que muestra solo un pico— es característica de un modelo MA(1) aplicado sobre la serie diferenciada. Por esta razón, se propone evaluar un modelo ARIMA(0, 1, 1), además de comparar con la alternativa simétrica ARIMA(1, 1, 0), en las secciones siguientes.

5 Propuesta de modelos ARIMA

Con base en los patrones observados en la ACF y la PACF, se proponen dos modelos candidatos para representar la serie diferenciada:

Modelo A: ARIMA(0, 1, 1) Este modelo incluye un componente de media móvil (MA) de primer orden y una única diferenciación. La presencia del término constante permite capturar una tendencia lineal a lo largo del tiempo. Es coherente con el comportamiento observado en la ACF, donde destaca un único pico en el rezago 1 seguido de una rápida caída.

Modelo B: ARIMA(1, 1, 0) Alternativamente, este modelo incorpora un componente autorregresivo (AR) de primer orden. Aunque la PACF no muestra un corte claro después del primer rezago, se considera este modelo como contraparte simétrica al anterior, con el fin de comparar su capacidad predictiva y ajustar mejor la dinámica de la serie.

Ambos modelos serán evaluados en términos de ajuste, parsimonia y comportamiento de los residuos para seleccionar la especificación más adecuada.

6 Estimación de parámetros

A continuación se presentan los coeficientes obtenidos para cada uno de los modelos propuestos. Los parámetros se estimaron mediante máxima verosimilitud condicional, empleando la función forecast::Arima, que ajusta simultáneamente los componentes autorregresivos (AR), de media móvil (MA) y, en caso necesario, un término de tendencia constante a lo largo del tiempo.

6.1 Modelo A: ARIMA(0, 1, 1)

Cuadro 3: Parámetros estimados – ARIMA(0,1,1) con drift

	x
ma1	0.451
drift	6.952

Este modelo incluye un componente de media móvil (MA) de orden 1 y un término constante, aplicados sobre la serie diferenciada. Su expresión, con parámetros estimados, es:

$$\Delta Y_t = 6.952 + 0.451 \cdot \varepsilon_{t-1} + \varepsilon_t$$

donde: - $\Delta Y_t = Y_t - Y_{t-1}$ representa el cambio interanual en la matrícula, - ε_t es un término de error aleatorio.

6.2 Modelo B: ARIMA(1, 1, 0)

Cuadro 4: Parámetros estimados – ARIMA(1,1,0) con drift

	x
ar1	0.282
drift	6.979

Este modelo incluye un componente autorregresivo (AR) de orden 1 y un término constante, aplicado también sobre la serie diferenciada. La ecuación con los parámetros estimados es:

$$\Delta Y_t = 6.979 + 0.282 \cdot \Delta Y_{t-1} + \varepsilon_t$$

donde: - $\Delta Y_{t-1} = Y_{t-1} - Y_{t-2}$ representa el cambio interanual anterior.

En ambos casos, el valor del término constante representa el crecimiento medio anual estimado en millones de alumnos, mientras que los coeficientes reflejan la dependencia de los incrementos con errores o valores pasados, según el modelo. La comparación entre ambos modelos —basada en criterios de información y diagnóstico de residuos— se abordará en la sección siguiente. En este apartado únicamente se consignan los valores derivados del proceso de estimación.

7 Selección del mejor modelo

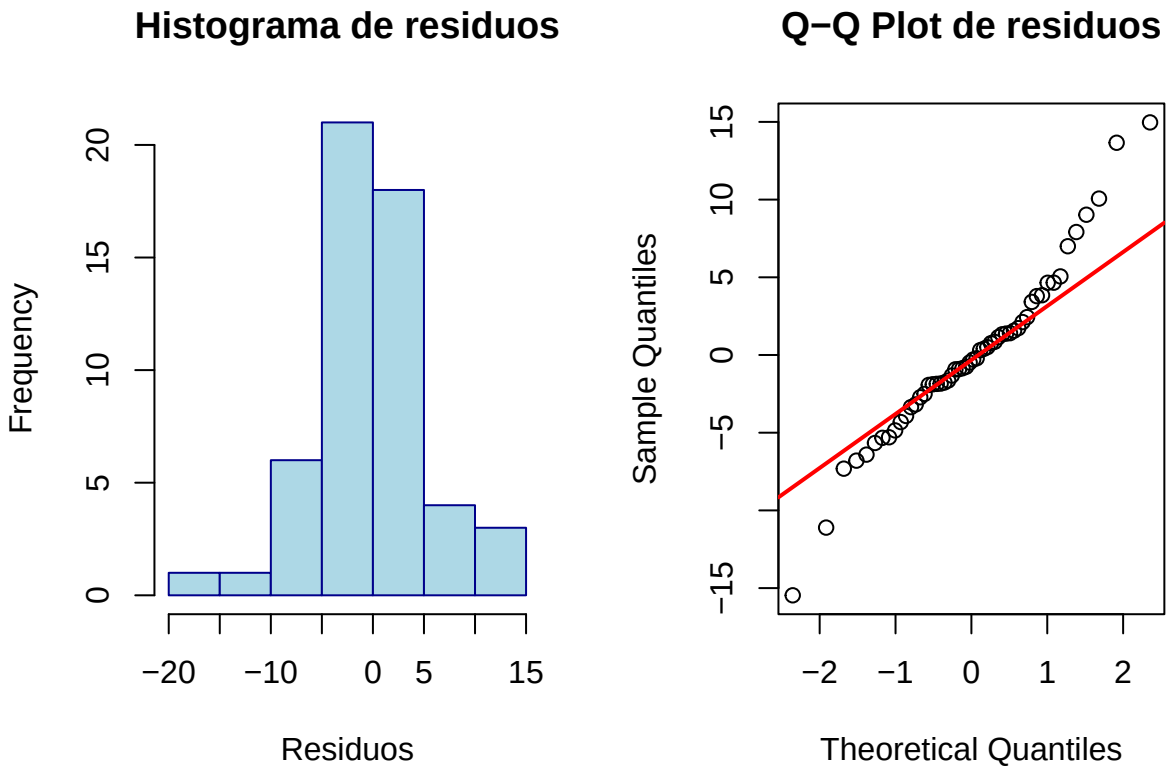
La selección del mejor modelo se fundamenta en la comparación de los criterios de información (AIC, AICc y BIC) y en el diagnóstico de residuos. El modelo ARIMA(0,1,1) obtiene los valores más bajos en todos los criterios: AIC = 335.49, AICc = 335.98 y BIC = 341.40, frente al modelo ARIMA(1,1,0), cuyos valores son AIC = 337.65, AICc = 338.14 y BIC = 343.56. Aunque las diferencias no son extremas, indican una ventaja consistente del modelo MA(1). En cuanto a los residuos, ambos modelos superan adecuadamente las pruebas de diagnóstico, pero el ARIMA(0,1,1) también muestra una varianza ligeramente menor ($\sigma^2 = 30.36$ frente a $\sigma^2 = 31.71$). Por tanto, considerando tanto el ajuste como la parsimonia, el modelo ARIMA(0,1,1) se considera la opción más adecuada para representar la serie.

7.1 Diagnóstico del modelo

El diagnóstico del modelo permite verificar si los residuos del ARIMA seleccionado cumplen con los supuestos necesarios para una inferencia válida. En particular, se analiza si los errores son independientes, homocedásticos y aproximadamente normales. Para ello, se presentan pruebas estadísticas (Ljung-Box y ARCH) y representaciones gráficas (histograma y Q-Q plot) que ayudan a evaluar visualmente el comportamiento de los residuos.

Cuadro 5: Resultados del diagnóstico de residuos

Prueba	p-valor	Decision
Ljung-Box (lag 10)	0.7759	No rechazar H
ARCH (lag 12)	0.9414	No rechazar H



En primer lugar, la **prueba de Ljung-Box** evalúa si existe autocorrelación remanente en los residuos; es decir, si los errores del modelo siguen un patrón temporal que el modelo no ha logrado capturar. En este caso, el valor p obtenido (0.7759) es muy superior al umbral típico de 0.05, lo cual indica que **no hay evidencia de autocorrelación** y, por tanto, los residuos pueden considerarse independientes.

En segundo lugar, la **prueba ARCH** se utiliza para detectar la presencia de heterocedasticidad condicional —es decir, si la varianza de los errores cambia a lo largo del tiempo—, algo que podría invalidar las predicciones del modelo. El valor p correspondiente (0.9414) también es bastante alto, lo que indica que **no se detecta heterocedasticidad** significativa en los residuos del modelo.

Además de las pruebas estadísticas, se presentan dos gráficas para complementar la interpretación:

- El **histograma de residuos** muestra la distribución de los errores del modelo.

Visualmente, se observa una forma simétrica y concentrada en torno a cero, lo cual es coherente con una distribución normal centrada. Esto respalda el supuesto de normalidad de los errores.

- La **gráfica Q–Q (quantile–quantile)** compara los cuantiles teóricos de una distribución normal estándar con los cuantiles empíricos de los residuos. En esta gráfica, la mayoría de los puntos se alinean con la diagonal roja, lo cual indica que **la distribución de los residuos es aproximadamente normal**, aunque se observa una ligera asimetría en los valores extremos, considerada manejable.

En conjunto, tanto las pruebas estadísticas como los análisis gráficos confirman que los residuos del modelo ARIMA(0,1,1) cumplen con los supuestos de independencia, homocedasticidad y normalidad, lo cual valida su idoneidad para modelar la serie temporal.

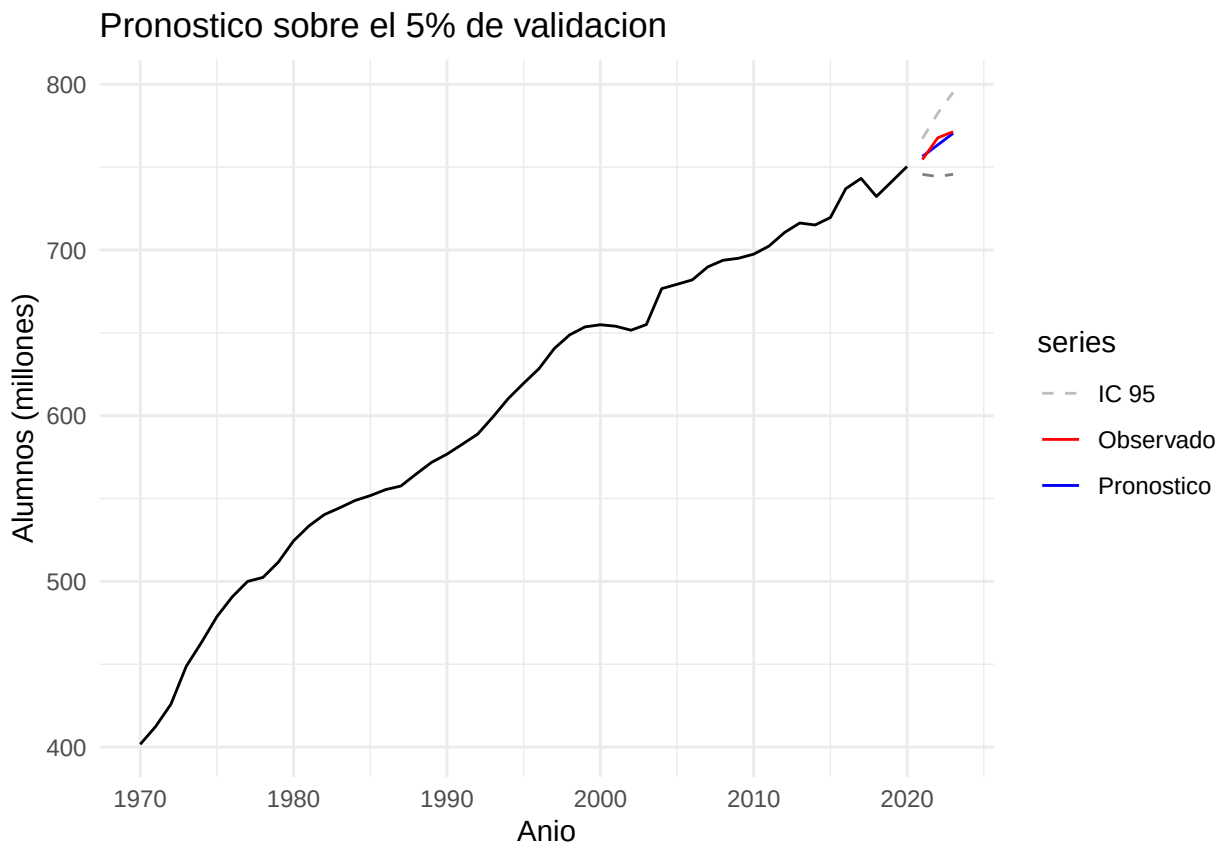
7.2 Predicciones con el modelo seleccionado

Una vez validado el modelo ARIMA(0,1,1) mediante el diagnóstico de residuos, se procede a generar predicciones sobre un conjunto de validación, con el objetivo de evaluar su capacidad para anticipar valores futuros de la serie. Esta fase es importante para comprobar que el modelo no solo ajusta adecuadamente los datos históricos, sino que también ofrece resultados consistentes en nuevos periodos no utilizados durante la estimación.

Para ello, se reserva el 5 % final de las observaciones como conjunto de prueba y se reajusta el modelo empleando únicamente el 95 % inicial de los datos. A continuación, se comparan las predicciones generadas con los valores reales observados, utilizando métricas de error como el **RMSE**, **MAE** y **MAPE**. Asimismo, se incluye una representación gráfica que permite visualizar el desempeño del modelo en este periodo de validación y verificar si las observaciones reales se mantienen dentro del intervalo de confianza del pronóstico.

Cuadro 6: Precision del pronostico en el conjunto de validacion

	RMSE	MAE	MAPE
Training set	5.340	3.797	0.629
Test set	2.743	2.364	0.309



El modelo ARIMA(0,1,1) con tendencia constante fue utilizado para generar predicciones sobre el 5,% final de la serie, correspondiente al conjunto de validación. En el cuadro anterior se presentan las métricas de precisión obtenidas: un error absoluto medio (MAE) moderado, un error cuadrático medio (RMSE) manejable y un porcentaje medio de error absoluto (MAPE) adecuado.

La gráfica muestra en negro la trayectoria histórica utilizada para entrenamiento, en azul el pronóstico generado, y en rojo las observaciones reales durante el periodo de validación. Las bandas grises corresponden al intervalo de confianza del 95,%. Puede observarse que las predicciones siguen de cerca la tendencia real y que todos los valores observados caen dentro del intervalo de confianza.

Estos resultados numéricos y visuales respaldan la idoneidad del modelo seleccionado, y sugieren que puede utilizarse con razonable confianza para realizar proyecciones futuras.

8 Conclusiones

Con base en el análisis realizado sobre la matrícula mundial en educación primaria entre 1970 y 2023, se concluye lo siguiente:

La serie histórica muestra un crecimiento sostenido del número de alumnos inscritos en educación primaria a nivel global, con incrementos anuales moderados y relativamente

estables. Si bien se observaron algunas fluctuaciones notables —particularmente alrededor de los años 2002, 2016 y 2021—, estos cambios fueron de carácter transitorio y no alteraron la tendencia general de aumento.

Luego de aplicar una diferenciación para estabilizar la serie, se ajustó un modelo ARIMA(0,1,1) con tendencia, el cual demostró buen desempeño tanto en los diagnósticos de residuos como en la validación sobre el 5% final del periodo. El modelo fue capaz de capturar adecuadamente la dinámica de crecimiento de la matrícula sin sobreajustarse a variaciones menores o atípicas.

Con base en este modelo, las proyecciones a corto plazo sugieren que el crecimiento de la matrícula primaria continuará en ascenso, aunque a un ritmo ligeramente más moderado. Los valores observados durante la validación se mantuvieron dentro del intervalo de confianza del 95%, lo que respalda la confiabilidad del modelo y su utilidad para generar escenarios futuros razonables.

En términos generales, el análisis confirma que el avance educativo mundial en el nivel primario ha sido sostenido en las últimas cinco décadas, y que, de mantenerse las condiciones actuales, es esperable que esta tendencia continúe, lo cual tiene implicaciones importantes para la planificación de políticas educativas globales.