ANÁLISIS DE SERIES DE TIEMPO PARA EL PRONÓSTICO DE INDICADORES EN EDUCACIÓN SUPERIOR: CASO DE LAS TASAS DE ABSORCIÓN, COBERTURA Y ABANDONO EN MÉXICO

Analysis of Time-series for the Forecast of Indicators in Higher education: The Case of Absorption, Coverage and Dropout Rates in Mexico

Yurixhi Andaya Hernández[1], Eric Leonardo Huerta Manzanilla[1*]

[1]Universidad Autónoma de Querétaro

Autor de correspondencia
*eric.huerta@uaq.mx

RESUMEN

En México, las tasas de absorción, cobertura y abandono son importantes indicadores de la capacidad del sistema educativo para atender a la población en edad escolar dentro de los diferentes niveles educativos. A su vez, estas tasas son parte de los indicadores de la llamada "trayectoria escolar" y, debido a los fenómenos que reflejan cada una de ellas, se espera que la evolución de la absorción y la cobertura tenga un constante ascenso a través del tiempo, mientras se espera que el abandono tienda al descenso. En las últimas décadas, se ha observado un aumento sostenido en la tasa de cobertura en el nivel superior; en tanto la tasa de absorción, por el contrario, parece disminuir y el abandono presenta ascensos y descensos constantes. Con base en datos históricos de los últimos 24 años, se realizó un pronóstico de los niveles que estas tasas podrían alcanzar durante los próximos 3 años. Fueron empleados modelos ARIMA debido a su probada capacidad para la proyección de pronósticos completos y bastante aproximados. El objetivo de este trabajo es mostrar la utilidad y pertinencia del empleo de herramientas estadísticas de pronóstico para proyectar el comportamiento de fenómenos en el campo educativo. Del pronóstico realizado, los datos sugieren que se espera un crecimiento sostenido para la tasa de cobertura en los próximos 3 años; mientras que las tasas de absorción y abandono presentan fluctuaciones en su comportamiento, de modo que no es claro si crecerán o decrecerán en este mismo periodo.

Palabras clave: educación superior, indicadores educativos, modelos ARIMA, pronóstico, análisis estadístico.

ABSTRACT

In Mexico, the rates of absorption, coverage and dropout are important indicators of the ability of the education system to attend to the school-age population, at the different educational levels. In turn, these rates are part of the indicators of the "school trajectory" and, due to the phenomena that each of them reflects, it is expected that the evolution over time of absorption and coverage will be in constant growth, while the dropout rate is expected to decrease. In the last decades, there has been a sustained increase in the coverage rate at the higher level, while the absorption rate seems to decrease, and the dropout presents constant ascents and descents. Based on historical data from previous 24 years, it is intended to forecast the levels that such rates could reach during the next 3 years. The ARIMA models were used because of their proven ability to project complete and accurate forecasts. The aim of this work is to show the usefulness and relevance of statistical forecasting tools to project the behavior of phenomena in the educational field. From the forecast made, the data suggests that a sustained growth is expected for the next 3 years for the coverage rate; while absorption and abandonment rates show fluctuations in their behavior, so it is unclear whether they will increase or decrease in this same period.

Keywords: higher education, educational indicators, ARIMA models, forecast, statistical analysis

INTRODUCCIÓN

Los indicadores educativos son herramientas que sirven para analizar el funcionamiento de los sistemas educativos y son de especial importancia debido a que permiten medir y conocer el desempeño de las acciones educativas respecto a una meta o un estándar [1]. En México, hasta hace poco tiempo la información sobre el sistema educativo era escasa y usualmente poco actualizada, por lo que se contaba con pocos indicadores. En consecuencia, se analizaba e interpretaba el funcionamiento del sistema educativo con datos deficientes y limitados. El reciente avance en los campos de la informática y la

computación ha facilitado el procesamiento y distribución de la información, sin embargo, un problema inherente a la información del ámbito educativo es que ésta puede tener interpretaciones diversas debido a su naturaleza fundada en acontecimientos de índole social. En este sentido, el empleo de herramientas estadísticas permite el análisis de los distintos fenómenos educativos desde una perspectiva cuantitativa [2].

El análisis de los indicadores educativos del nivel superior se torna de interés cuando su crecimiento se compara con el de los otros niveles educativos. Como ejemplo, se puede observar que en años recientes se ha logrado una cobertura cercana a un 100 % en el nivel básico, pero los rezagos en nivel superior son evidentes, pues se atiende a poco más del 30 % de la población en la edad escolar correspondiente a este nivel. Esto implica que muchos jóvenes no tienen la oportunidad de continuar sus estudios en el nivel superior y con ello se frena la competitividad, dado que se inhibe la formación de recursos humanos que impulsen la productividad y contribuyan al desarrollo nacional, y esto representa importantes costos sociales [3].

En el escenario mexicano, de acuerdo con el Reporte de Indicadores Educativos [4], los indicadores principales en el nivel superior son: a) absorción, b) cobertura y c) abandono.

La absorción se refiere esencialmente a la proporción de estudiantes que, al final de cierto nivel educativo, acceden de inmediato al siguiente nivel; por su parte, la cobertura mide el porcentaje de estudiantes matriculados en el nivel educativo correspondiente a su edad; mientras que el abandono refleja la cantidad de alumnos que dejan la escuela entre ciclos escolares consecutivos. Estos tres indicadores incluyen a la educación normal y a las licenciaturas y, en el caso de la tasa de cobertura, ésta considera a los estudiantes pertenecientes a una cohorte que va de los 18 a los 22 años. Dentro del reporte Panorama Educativo de México [5], se establecen los referentes para cada uno de estos indicadores. Para el caso de la tasa de absorción, se considera que, los estudiantes deberían mantener una trayectoria regular en educación, por lo que el objetivo de este indicador es alcanzar el 100 %. Por su parte, para la tasa de abandono se establece que el sistema educativo debe retener al 100 % de los alumnos, por lo cual el referente de este indicador queda establecido en 0 %.

Finalmente, la tasa de cobertura señala que la totalidad de estudiantes de edades normativas para cursar un determinado nivel educativo deberían encontrarse efectivamente cursándolo, a partir de lo cual se establece el referente de este indicador en 100 %. Con base en estos referentes y las técnicas de modelado estadístico, como los pronósticos, se pueden analizar y proyectar patrones de datos pasados para determinar el rango en el que probablemente se incluirán valores futuros y visualizar si, según el comportamiento histórico de estos indicadores, su evolución va en el sentido esperado o si, por el contrario, se requiere la implementación de políticas y/o acciones que las impulsen en el sentido deseado.

Los pronósticos son una herramienta que proporciona una estimación cuantitativa de la probabilidad de que ocurra uno o varios eventos futuros, además permiten llegar a comprensiones más complejas de un fenómeno a través del análisis histórico de datos [6]. Entre los modelos más conocidos se encuentran los de suavización, análisis de correlación y ARIMA (modelo autorregresivo integrado de promedio móvil).

Estos últimos son especialmente populares debido a su flexibilidad para ajustarse a los datos y su capacidad para modelar diversas series de tiempo, con o sin componentes de tendencia o estacionales, para generar pronósticos [7].

Los modelos ARIMA son modelos paramétricos que tratan de obtener la representación de una serie en términos de la interrelación de sus datos a través del tiempo. Entre sus principales ventajas se encuentran: a) se aplican para datos tanto discretos como contin-

uos, b) sólo se pueden aplicar a datos espaciados equidistantemente en intervalos discretos de tiempo, c) son útiles para tratar series que presentan patrones estacionales, y d) pueden ser aplicados a series estacionarias y no estacionarias.

Debido a sus bondades, estos modelos son empleados en diversos campos disciplinares como economía, meteorología, astronomía, demografía, marketing y sociología, entre otros [7] [8]. Tradicionalmente, estos modelos se desarrollan en 4 etapas, que son:

1. Identificación del modelo;
2. estimación de los parámetros implícitos del modelo;
3. verificación de supuestos y
4. uso del modelo (pronósticos).

Para hacer una buena representación, es necesario elegir un intervalo de tiempo que capture un comportamiento descriptivo para el patrón que se desea analizar, y este problema depende usualmente de la periodicidad con la que se obtengan los datos (índices mensuales, anuales, trimestrales, entre otros) [9]. Sus parámetros principales son los que indican el orden de los distintos componentes del modelo: autorregresivo (p), integrado (d) y de media móvil (q).

Un modelo autorregresivo realiza una regresión sobre la misma variable pero en distinto periodo de tiempo (t-1 y t). Un modelo integrado es un proceso no estacionario que se convierte en estacionario después de realizar algunas operaciones de diferencias (diferenciación). Y un modelo de medias móviles describe una serie de tiempos estacionaria [10]. La finalidad de los modelos ARIMA es encontrar el arreglo de sus componentes que mejor se ajuste a una serie de datos para realizar un pronóstico de sus valores futuros. Si bien se cuenta con la metodología tradicional de cuatro etapas para su cálculo, se han desarrollado funciones automáticas para la realización de este proceso.

En este trabajo se empleó el *software* R para el desarrollo de los modelos ARIMA de las tres series de tiempo bajo estudio. En ese sentido, para el establecimiento del modelo ARIMA más adecuado se utilizó una función automática, descrita en la metodología. Se espera que este trabajo pueda ser un apoyo para la aplicación de modelado ARIMA, incluso para las personas que tienen conocimientos básicos sobre el tema. Además, es importante mencionar que está totalmente enfocado a mostrar la aplicación de una herramienta conocida, como lo es este tipo de modelado, a un ejemplo en el campo de la investigación educativa, mas no al desarrollo de una nueva herramienta ni al análisis profundo de alguna problemática educativa en el nivel superior.

METODOLOGÍA

Los datos que se analizan en este trabajo corresponden a un histórico de 24 años, desde el periodo 1994-1995 hasta el periodo 2017-2018. Resulta relevante señalar que, debido a que se trata de datos de ciclos escolares, estos van de agosto de un año a julio del año siguiente. Los datos analizados corresponden al Reporte de Indicadores Educativos de la Secretaría de Educación Pública [4].

Modelado ARIMA

La construcción de modelos ARIMA se lleva a cabo de forma iterativa mediante un proceso en el que se pueden distinguir cuatro etapas [10]:

a. Identificación. El objetivo es determinar los órdenes p, d, q que parecen apropiados para reproducir las características de la serie bajo estudio.
b. Estimación. Se realizan inferencias sobre los parámetros, condicionadas a que el modelo investigado sea apropiado.
c. Validación. Se realizan contrastes para comprobar si el modelo se ajusta a los datos, en caso de no ser así, se analizan las posibles discrepancias del modelo

propuesto para poder mejorarlo.

d. Predicción. Se obtienen pronósticos en términos probabilísticos de los valores futuros de la variable.

Identificación

El objetivo de esta etapa es seleccionar el modelo ARIMA (*p, d, q*) apropiado para la serie, es decir, que reproduce las características de la serie. La identificación del mejor modelo se desarrolla en dos fases:

A. Análisis de estacionariedad. Se determinan las transformaciones que son necesarias para obtener una serie estacionaria, esto es, hacer que los datos posean una media y una varianza constantes a lo largo de la serie.

B. Elección de los órdenes *p* y *q*. Una vez obtenida la serie estacionaria, el objetivo es determinar el proceso estacionario ARMA (*p, q*) que la haya generado.

Para esta fase se realiza una serie de operaciones que van de la identificación de la tendencia y estacionalidad al cálculo de diferencias para lograr la estacionariedad de la serie. En este artículo se mencionan todas las etapas originales de la construcción de los modelos, sin embargo, se emplean algunas operaciones automáticas, de forma que la implementación sea sencilla para cualquier lector, sin importar su área de conocimiento. Se prioriza la interpretación de los datos obtenidos en cada fase sobre los aspectos técnicos de los modelos.

Representación gráfica

En la etapa de identificación, la representación gráfica es de utilidad para visualizar la evolución de las series a lo largo del tiempo y otras características como tendencia, estacionalidad y estacionariedad. La tendencia se entiende como un componente de largo plazo que representa el crecimiento o decrecimiento de una serie histórica; la estacionalidad se refiere a las fluctuaciones estacionales que se pueden encontrar en los datos que son clasificados en periodos de tiempo específicos —años, trimestres, meses o semanas— y no debe confundirse con la estacionariedad, ya que son conceptos distintos. La estacionariedad es una característica en la cual una serie tiene media y varianza constantes a lo largo del tiempo y no presenta tendencia [7].

Análisis de la tendencia

La tendencia es un componente de largo plazo que representa el crecimiento o decrecimiento de una serie histórica. Un primer paso para su análisis es visualizar el comportamiento de las series a través de gráficas.
En la Fig. 1 se puede observar que la tasa de absorción tiene una tendencia en descenso, mientras que la tasa de cobertura refleja una clara tendencia en ascenso. Por su parte, la tasa de abandono parece no presentar tendencia alguna.

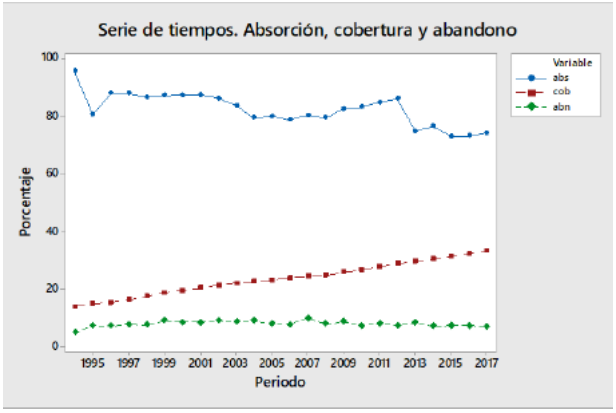

Figura 1. Comportamiento histórico de las tasas de absorción, cobertura y abandono en educación superior.

- **Análisis de la estacionariedad**

El análisis de la estacionariedad se puede realizar a través de los gráficos de la función de autocorrelación (ACF) y de autocorrelación parcial (PACF).

Estas funciones se utilizan para describir la presencia o ausencia de correlación entre los datos de las series de tiempo, señalan-

do si las observaciones pasadas influyen en las actuales. Cuando una serie no presenta estacionariedad, esto se refleja a través de picos que sobrepasan los límites de significancia de manera repetida y en *lags* específicos. Los *lags* son el número de períodos de tiempo que separan a los datos de las series de tiempo. Por el contrario, cuando una serie es estacionaria las barras no sobrepasan los límites de significancia.

La estacionariedad es importante debido a que es una característica requerida previo a la selección de un modelo ARIMA.

Las funciones de ACF y PACF nos ayudan además a conocer a qué *lags* los datos son significativos, es decir, nos ayudan a establecer los ciclos o periodos en los que se presenta estacionalidad.

Los correlogramas para las funciones ACF y PACF de las tasas de absorción, cobertura y abandono se muestran en las Figs. 2-4.

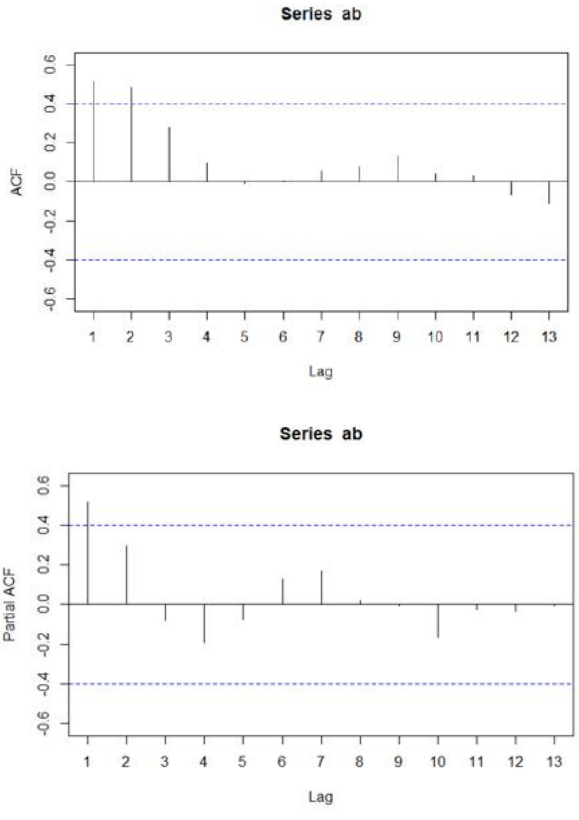

Figura 2. Funciones de autocorrelación y autocorrelación parcial de la tasa de absorción.

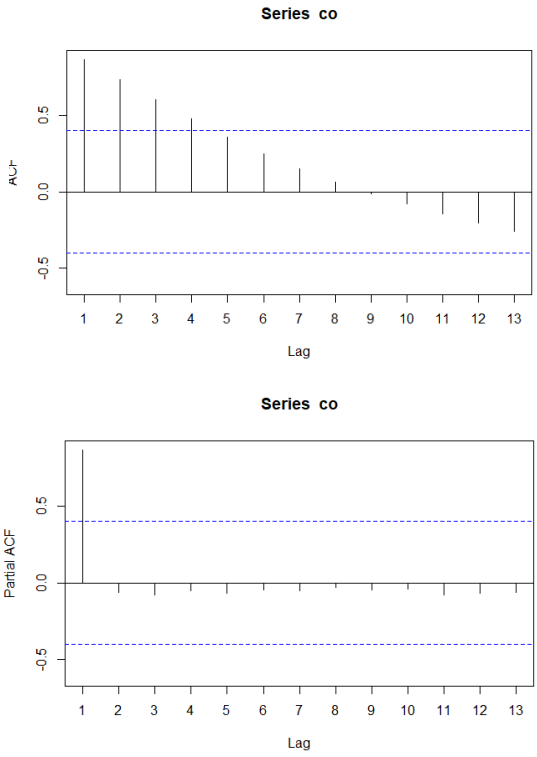

Figura 3. Funciones de autocorrelación y autocorrelación parcial de la tasa de cobertura.

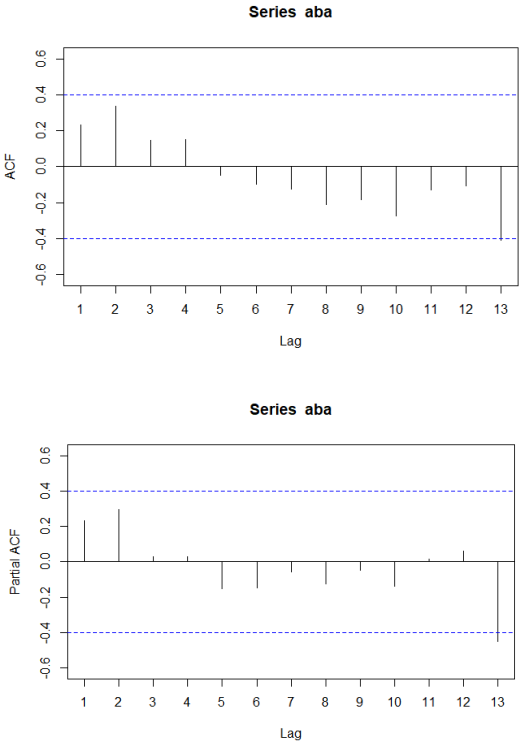

Figura 4. Funciones de autocorrelación y autocorrelación parcial de la tasa de abandono.

De los correlogramas se observa:

Para la tasa de absorción:

a. Análisis de la ACF. Debe omitirse la primera barra, ya que corresponde a la autocorrelación con la misma observación; por ende, la segunda barra sí es significativa debido a que rebasa el límite de significancia, y representa un *lag* al cual la serie puede ser estacional o, en este caso, que la autocorrelación no es significativa para *lags* > 2.

b. Análisis de la PACF. Al igual que en la ACF, se omite la primera barra por corresponder a la autocorrelación con la misma observación. Posteriormente, no se identifican *lags* significativos (Fig. 2).

Para la tasa de cobertura:

a. Análisis de la ACF. Se elimina la primera barra y se observa que la segunda, tercera y cuarta barra son significativas, lo cual podría indicar que la autocorrelación no es significativa para *lags* > 4.

b. Análisis de la PACF. Al igual que en la tasa de absorción, no se identifican *lags* significativos (Fig. 3).

Para la tasa de abandono:

a. a) Análisis de la ACF. No hay barras fuera de los límites de significancia, por lo que no se identifican *lags* significativos.

b. b) Análisis de la PACF. Se observa que la barra en el lag 13 es significativa, lo que podría significar que la autocorrelación no es significativa cuando los *lags* son > 13 (Fig. 4).

La significancia de los lags o retardos es relevante porque indica el número de periodos de tiempo en los cuales los datos se correlacionan.

Por ejemplo, para un lag significativo en 5, si el periodo de tiempo bajo estudio fuera de años, podría significar que los datos analizados tienen una relación que se contamina o confluye de alguna forma cada 5 años y, debido a eso, se genera una relación en ese periodo de tiempo específico, lo que querría decir que cada 5 años se espera un comportamiento inusual que lo hará significativo para esa serie de tiempo.

En el caso de las tasas de absorción, cobertura y abandono, los datos sugieren que no existen *lags* altamente significativos en ninguna de las series, o que se requieren más datos para identificar la existencia de alguno.

Estimación

Una vez que se identificaron las tendencias y elementos estacionales y se determinó si la serie es estacionaria o no, el siguiente paso fue transformar aquellas series que de origen no fueran estacionarias, esto es, conseguir que su media y varianza sean constantes a través del tiempo; además, se requirió eliminar la tendencia y, a partir de lo observado en la transformación, se determinó el mejor modelo ARIMA para las series de tiempo de interés.

De lo observado en las Figs. 2-4 se identificó que las series de absorción y cobertura no son estacionarias y requieren de una o varias transformaciones para su ajuste, mientras que los datos sugieren que la serie de abandono es estacionaria de origen.

En este trabajo se plantea la utilización de una función automática para la realización del proceso de transformación y selección de modelo en el *software* R, como se explica a continuación:

- Para el desarrollo de modelos ARIMA en R, se debe considerar que existen diversas librerías para la realización de pronósticos, por lo que fue indispensable la instalación de aquellas que sean necesarias antes de comenzar a trabajar con las series. Para este trabajo se utilizaron las librerías "*readxl*", "*xts*", "*astsa*" y "*forecast*".

- Una vez que instaladas las librerías, el primer paso fue llamar a la base de datos que contenía la información de las series de tiempo con las que se trabajarán. En este caso, la base de datos provino de un documento de formato *.xls*.

- Después de ser llamada la base de datos, se le asignaron las propiedades de una serie de tiempos a los datos a través de la función "*ts*".

- Después se generó el modelo ARIMA de forma automática con la función *auto.arima*. De donde se obtiene lo siguiente:

Serie: Tasa de absorción

```
ARIMA(1,1,0)

Coefficients:
          ar1
      -0.5288
s.e.   0.2449

sigma^2 estimated as 17.46:  log likelihood=-65.18
AIC=134.35   AICc=134.95   BIC=136.62

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE
Training set -1.29256 4.00039 2.554209 -1.70429 3.206115
                    MASE       ACF1
Training set 0.9763676 0.1129521
```

El modelo automático calculado para la tasa de absorción es de la forma (1, 1, 0), lo que indica que se cuenta con un proceso de autorregresión, uno de integración y ninguno de promedio móvil.

Serie: Tasa de cobertura

```
ARIMA(0,1,0) with drift

Coefficients:
       drift
      0.8397
s.e.  0.0440

sigma^2 estimated as 0.04659:  log likelihood=3.14
AIC=-2.28   AICc=-1.68   BIC=-0.01
```

El modelo automático calculado para la tasa de cobertura es de la forma (0, 1, 0), lo que indica que no se cuenta con procesos autorregresivos, se tiene un proceso de integración y ninguno de medias móviles.

Para este modelo, al no contar con procesos de autorregresión o medias móviles y siendo d = 1, la función automática agrega el parámetro conocido como deriva o *drift* para realizar el cálculo de los coeficientes [11].

Serie: Tasa de abandono

```
ARIMA(0,0,0) with non-zero mean

Coefficients:
        mean
      7.6311
s.e.  0.1948

sigma^2 estimated as 0.9504:  log likelihood=-32.93
AIC=69.87   AICc=70.44   BIC=72.22
```

El modelo automático calculado para la tasa de cobertura es de la forma (0, 0, 0), lo que indica que no se tienen procesos autorregresivos, de integración o de promedios móviles. Esto además denota que la serie es aparentemente lo que se conoce como ruido blanco. El ruido blanco es un tipo de serie con un comportamiento aleatorio permanente, es decir, a los datos toman valores sin ninguna relación unos con otros través del tiempo. A partir de su correlograma ACF (Fig. 4), al no haber ningún coeficiente de correlación significativo, se puede decir que los datos son independientes.

Validación

Con una primera propuesta de modelos ARIMA para las tres series de tiempo, se contrastó con otros posibles modelos para confirmar si los modelos propuestos son los mejores o si existe una configuración con un mejor ajuste.

Serie: Tasa de absorción

Agregando un proceso de promedios móviles

```
arima(x = ab, order = c(1, 1, 1))

Coefficients:
          ar1     ma1
      -0.5734  0.0507
s.e.   0.4089  0.3924

sigma^2 estimated as 16.68:  log likelihood = -65.17,  aic = 136.34
```

El valor de máxima verosimilitud (*log-likelihood*) aumenta de -65.18 a -65.17. El valor de AIC aumenta de 134.35 a 136.34, pero se sabe que mientras más pequeño sea el AIC, mejor se ajustará el modelo a los datos.

Eliminando un proceso de autorregresión

```
arima(x = ab, order = c(0, 1, 1))
Coefficients:
          ma1
      -0.4072
s.e.   0.2067
sigma^2 estimated as 17.47:  log likelihood = -65.62,  aic = 135.25
Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE
Training set -1.409953 4.091918 2.532324 -1.860886 3.189769
                   MASE       ACF1
Training set 0.9680019 -0.038376
```

El valor de máxima verosimilitud disminuye de -65.18 a -65.62 pero, como lo que se desea es maximizar la log-verosimilitud, el valor más alto del primer modelo es marginalmente mejor. El valor de AIC aumenta de 134.35 a 135.25.

Por lo tanto, los datos sugieren que el modelo generado en automático es el que se ajusta mejor a los datos debido a que sus valores de máxima verosimilitud y AIC son los más aceptables para la tasa de absorción.

Serie: Tasa de cobertura

Agregando un proceso autorregresivo y un proceso de medias móviles

```
arima(x = co, order = c(1, 1, 1))

Coefficients:
         ar1      ma1
      0.9934  -0.5441
s.e.  0.0115   0.2426

sigma^2 estimated as 0.04929:  log likelihood = 0.41,  aic = 5.17
```

El valor de máxima verosimilitud disminuye de 3.14 a 0.41. El valor de AIC aumenta de -2.28 a 5.17, por lo que el valor del modelo automático es el más aceptable.

Agregando un proceso autorregresivo

```
arima(x = co, order = c(1, 1, 0))

Coefficients:
         ar1
      0.9596
s.e.  0.0429

sigma^2 estimated as 0.06283:  log likelihood = -2.08,  aic = 8.16
```

El valor de máxima verosimilitud disminuye de 3.14 a -2.48. El valor de AIC aumenta de -2.28 a 8.16, por lo tanto, el valor del modelo automático es mejor.

Agregando un proceso de medias móviles

```
arima(x = co, order = c(0, 1, 1))

Coefficients:
          ma1
      0.6702
s.e.  0.1077

sigma^2 estimated as 0.3451:  log likelihood = -20.7,  aic = 45.4
```

El valor de máxima verosimilitud disminuye de 3.14 a -20.7. Respecto a los valores de calidad relativa del modelo, el valor de AIC aumenta de -2.28 a 45.4.

Por lo tanto, los datos sugieren que el modelo generado en automático es el que se ajusta mejor a los datos para la tasa de cobertura.

Serie: Tasa de abandono

Agregando un proceso de autorregresión

```
arima(x = aba, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.3982      7.531
s.e.  0.2496      0.315

sigma^2 estimated as 0.8202:  log likelihood = -31.76,  aic = 69.53
```

El valor de máxima verosimilitud aumenta de -32.93 a -31.76. El valor de AIC disminuye de 69.87 a 69.53.

Agregando un proceso autorregresivo y un proceso de integración

```
arima(x = aba, order = c(1, 1, 0))

Coefficients:
         ar1
      -0.6027
s.e.   0.1905

sigma^2 estimated as 0.7287:  log likelihood = -29.22,  aic = 62.44
```

El valor de máxima verosimilitud aumenta de -32.93 a -29.22. El valor de AIC disminuye de 69.87 a 62.44.

Agregando un proceso autorregresivo, un proceso de integración y un proceso de promedios móviles

```
arima(x = aba, order = c(1, 1, 1))

Coefficients:
          ar1      ma1
      -0.5089  -0.1581
s.e.   0.3072   0.3448

sigma^2 estimated as 0.7222:  log likelihood = -29.13,  aic = 64.27
```

El valor de máxima verosimilitud aumenta de -32.93 a -29.13. El valor de AIC disminuye de 69.87 a 64.27.

Por lo tanto, de acuerdo con la evidencia, el modelo generado en automático no es el modelo que mejor se ajusta a los datos para la tasa de abandono. La evidencia sugiere que mediante procesos de autorregresión, integración y promedios móviles, se pueden obtener mejores ajustes.

Predicción

Para realizar los pronósticos se utilizan como base los modelos resultantes de las etapas de estimación y validación mediante la función predict. Esta función permite establecer la cantidad de datos futuros o periodos que se desean predecir, así como agregar elementos estacionales si es que se consideran relevantes en alguna de las series. En este caso, la cantidad de periodos a predecir fue de 3 años para cada una de las tasas y no se agregaron elementos estacionales, ya que, como se estableció en el análisis de los correlogramas (Figs. 2-4), la evidencia sugiere que ninguna de las series cuenta con lags significativos como para suponer elementos de estacionalidad. Los resultados se muestran a continuación:

Pronóstico: Tasa de absorción

```
$pred
Time Series:
Start = 2018
End = 2020
Frequency = 1
[1] 73.46595 73.72448 73.58778

$se
Time Series:
Start = 2018
End = 2020
Frequency = 1
[1] 4.178271 4.618955 5.583593
```

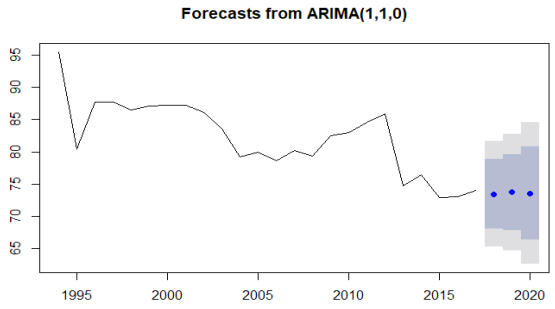

Figura 5. Valores pronosticados a tres años para la tasa de absorción.

La predicción se realizó sobre la base de un modelo de la forma (1, 1, 0). Es de notar que en la predicción se esperan ascensos y descensos en el comportamiento de la serie para los 3 años pronosticados, descendiendo de 73.95 (valor del periodo 2017-2018) a 73.47 (valor para 2018-2019), ascendiendo de nuevo a 73.72 (valor para 2019-2020) y luego descendiendo a 73.59 (valor para 2020-2021).

Pronóstico: Tasa de cobertura

```
$pred
Time Series:
Start = 2018
End = 2020
Frequency = 1
[1] 33.901381 34.650864 35.595439

$se
Time Series:
Start = 2018
End = 2020
Frequency = 1
[1] 0.8536112 0.9185096 1.1247839
```

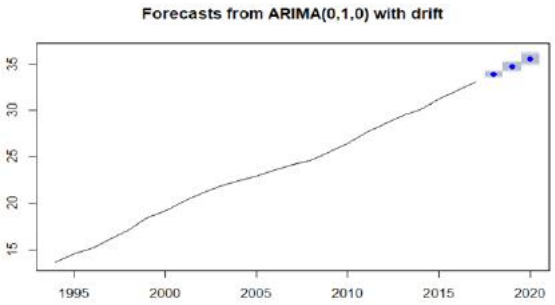

Figura 6. Valores pronosticados a tres años para la tasa de cobertura.

La predicción se realizó sobre la base de un modelo de la forma (1, 1, 0), y se espera que la cobertura crezca en un promedio de 0.8 % anual durante los siguientes tres años. Además, se observa en la Fig. 6 que el error esperado —señalado por los cuadros en color azul— es muy reducido respecto a los valores pronosticados.

Pronóstico: Tasa de abandono

```
$pred
Time Series:
Start = 2018
End = 2020
Frequency = 1
[1] 7.011581 6.867813 6.954464

$se
Time Series:
Start = 2018
End = 2020
Frequency = 1
[1] 0.8536112 0.9185096 1.1247839
```

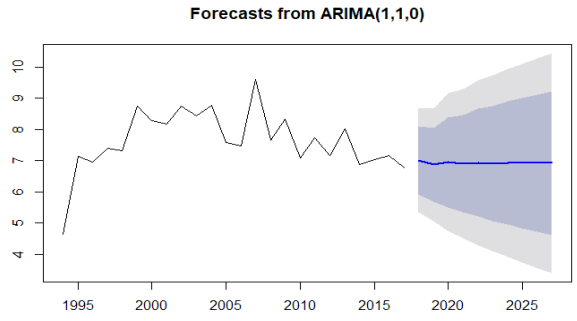

Figura 7. Valores pronosticados a tres años para la tasa de abandono.

La predicción se realizó sobre la base de un modelo de la forma (1, 1, 0), y se esperan ascensos y descensos en el comportamiento de la serie para los 3 años pronosticados, ascendiendo de 6.77 (valor del periodo 2017-2018) a 7.01 (valor para 2018-2019), descendiendo de nuevo a 6.87 (valor para 2019-2020) y luego ascendiendo a 6.95 (valor para 2020-2021).

RESULTADOS Y DISCUSIÓN

Del análisis de los datos históricos, las tasas de absorción y cobertura mostraron un comportamiento con tendencia pero sin elementos estacionales aparentes, mientras la tasa de abandono no reflejó tendencia ni estacionalidad, lo cual se comprobó a través del análisis gráfico de las funciones ACF y PACF. En el caso de la cobertura, su pronóstico es un reflejo de su comportamiento con tendencia creciente. Por otra parte, la absorción parece encontrarse en una etapa de estancamiento en la que su comportamiento es fluctuante, pero sin rebasar cierto nivel a pesar de tener una tendencia histórica decreciente en el pronóstico realizado. El abandono, de manera congruente con su comportamiento histórico, presenta un comportamiento fluctuante en el pronóstico. La ausencia de estacionalidad en las tres tasas indica que no existe un comportamiento cíclico.

Un primer paso para la determinación de los mejores modelos fue la eliminación de los elementos de tendencia y estacionalidad que pudieran contener las series, esto es, lograr obtener series estacionarias, lo que significa que las series deben tener una media y varianza constantes a través del tiempo. Esta eliminación es la llamada "fase de transformación" y, de manera usual, se realiza por medio de operaciones de diferenciación, donde se encuentran los logaritmos como una de las operaciones de diferenciación más comunes. La fase de transformación en este trabajo se realizó de manera automática utilizando la función *auto.arima*, en el *software* R. Esta función devuelve el mejor modelo ARIMA según los valores de calidad relativa del modelo (AIC, AICc, BIC) y del valor de máxima verosimilitud.

Los valores de calidad relativa de los modelos automáticos y de validación se muestran en la Tabla 1.

De acuerdo con los valores de calidad relativa observados en la Tabla 1, y considerando que se desea obtener los valores máximos de la log-verosimilitud y los valores más pequeños de AIC, los modelos resultantes de la función automática para las tasas de absorción y cobertura parecen ajustarse de manera adecuada a los datos de las series. Por su parte, para la función de abandono en la etapa de validación, los comparativos realizados demostraron que los modelos que contienen por lo menos un proceso *p*, *d*, *q* tienen mejores ajustes que el modelo automático; por ejemplo, el modelo de la forma (1, 1, 0). No debe olvidarse en este punto que una característica de los modelos ARIMA es que puede ajustar una gran cantidad de modelos distintos a una misma serie de datos, por lo que es necesario que se tengan claros los criterios que se considerarán como óptimos para la aceptación e implementación de un modelo, ya que, aunque en ocasiones pueden tener valores de máxima verosimilitud y de calidad relativa similares, la configuración de cada modelo describe procesos con características distintas. Los pronósticos generados a partir de aquellos modelos ARIMA determinados con la función automática y en la etapa de validación siguen un comportamiento que parece ajustarse al comportamiento natural de las series.

En la Tabla 2 se muestra el comparativo del comportamiento general de los datos

Tabla 1. Valores de calidad relativa de los modelos.

	Log Verosimilitud	AIC	
Tasa de absorción			
Modelo automático (1,1,0)	-65.18	134.35	
Modelo de validación 1 (1,1,1)	-65.17	136.34	
Modelo de validación 2 (0,1,1)	-65.62	135.25	
Tasa de cobertura			
Modelo automático (0,1,0)	3.14	-2.28	
Modelo de validación 1 (1,1,1)	0.41	5.17	
Modelo de validación 2 (1,1,0)	-2.08	8.16	
Modelo de validación 3 (0,1,1)	-20.7	45.4	
Tasa de abandono			
Modelo automático (0,0,0)	-32.93	69.87	
Modelo de validación 1 (1,0,0)	-31.76	69.53	
Modelo de validación 2 (1,1,0)	-29.22	62.44	
Modelo de validación 3 (1,1,1)	-29.13	64.27	

históricos contra los datos pronosticados a partir de sus estadísticos básicos.
De la Tabla 2 puede observarse que los valores del error estándar promedio de los datos pronosticados son relativamente pequeños, lo que podría ser un indicio del adecuado ajuste de los datos a los modelos de pronóstico seleccionados. Por su parte, las medias y medianas tanto de los datos históricos como de los datos pronosticados son muy similares para cada una de las tasas, lo que parece indicar que las distribuciones de los datos son simétricas.

De las desviaciones estándar de los datos pronosticados, parece que los datos no se encuentran muy dispersos entre sí debido a sus valores pequeños, a diferencia de las desviaciones estándar de los datos históricos, que muestran dispersiones mayores en cada una de las tasas.

De lo obtenido en el desarrollo de modelos ARIMA y de lo observado en las Figs. 5-7 sobre el pronóstico del comportamiento de las tasas de absorción, cobertura y abandono en el nivel superior educativo mexicano, la evidencia sugiere que existirá una fluctuación en el comportamiento de las series en un pronóstico a tres años tanto para las tasas de absorción y abandono, mientras que se espera un crecimiento sostenido en el caso de la tasa de cobertura.

De manera general, el comportamiento de las tasas derivado del pronóstico realizado en este trabajo podría tener importantes implicaciones en el escenario educativo de acuerdo con lo siguiente:

1. Para la tasa de absorción: no existe evidencia de una tendencia de crecimiento dentro de los próximos 3 años, por lo que no se podría esperar que este indicador rebasara el 74% de absorción en educación superior de acuerdo con los datos pronosticados y esto representa una brecha importante con su nivel deseado del 100 %. De su comportamiento histórico (Fig. 1), la evidencia sugiere que incluso podría esperarse un decrecimiento en esta tasa que representaría un retroceso para alcanzar el nivel deseado.

2. Para la tasa de cobertura: la evidencia sugiere que esta tasa podría presentar un crecimiento sostenido dentro de los próximos 3 años, con un crecimiento anual promedio de 0.8 %, lo que parece indicar que esta tasa se encuentra evolucionando de manera positiva hacia su nivel ideal del 100 %.

3. Para la tasa de abandono: no existe evidencia de una tendencia de decrecimiento dentro de los próximos 3 años

Tabla 2. Comparativos de estadísticos básicos de datos históricos vs. datos pronosticados.

	Estadísticos descriptivos de los datos históricos (1994-2017)	Estadísticos descriptivos de los datos pronosticados (2018-2020)
Tasa de absorción	Media= 82.23 Error estándar promedio= 1.15 Desviación estándar= 5.66 Mediana= 82.76	Media= 73.59 Error estándar promedio=0.074 Desviación estándar= 0.129 Mediana= 73.59
Tasa de cobertura	Media= 23.32 Error estándar promedio= 1.17 Desviación estándar= 5.75 Mediana= 23.30	Media= 34.72 Error estándar promedio= 0.490 Desviación estándar= 0.849 Mediana= 34.65
Tasa de abandono	Media= 7.63 Error estándar promedio= 0.199 Desviación estándar= 0.975 Mediana= 7.53	Media= 6.94 Error estándar promedio= 0.042 Desviación estándar= 0.072 Mediana= 6.95

que indique que esta tasa tendrá un comportamiento que le permita alcanzar su nivel ideal de 0%. De los datos pronosticados se observa un comportamiento más bien errático, ya que crece y decrece sin una tendencia clara.

Finalmente, es importante señalar que una consideración sobre este análisis es la cantidad de datos que se tuvieron disponibles para la realización del pronóstico. Es sabido que, entre mayor cantidad de datos se tengan para realizar un pronóstico, este será más preciso, ya que con un amplio referente histórico se pueden visibilizar los patrones de comportamiento con mayor facilidad y, por ende, replicarlos en el pronóstico. También es importante mencionar que se realizó el pronóstico de una cantidad pequeña de años, debido a que entre mayor es la cantidad de periodos que se pretenden pronosticar, el nivel de error aumenta.

CONCLUSIONES

Los modelos ARIMA se encuentran entre los modelos más versátiles para realizar pronósticos debido a que pueden ser empleados para estimar una amplia variedad de series de tiempo, tanto estacionales como no estacionales. A través de su proceso iterativo de cuatro etapas permiten identificar modelos tentativos, estimar parámetros, validar los modelos tentativos y, finalmente, calcular un pronóstico válido. Su utilización para estimar valores futuros en el campo de los indicadores educativos aporta herramientas de análisis cuantitativo que permiten tener un panorama general del comportamiento que tendrán ciertos fenómenos educativos en el corto y largo plazo. Si bien es importante resaltar que estos fenómenos se encuentran influenciados por una gran cantidad de variables cualitativas, el análisis numérico de su comportamiento histórico es apenas un acercamiento para entender su evolución, y comprender la mejor manera de orientar su desarrollo en el sentido esperado.

El ejercicio desarrollado en este trabajo es una evidencia de que la aplicación de los modelos ARIMA para el pronóstico de datos del campo educativo es factible y puede ser accesible a personas de todos los campos del conocimiento.

De los resultados obtenidos en los pronósticos se observan algunos retos a enfrentar en el escenario educativo mexicano de nivel superior, ya que aunque la tasa de cobertura parece crecer hacia el nivel deseado, su crecimiento podría ser lento si se considera que en el nivel básico ya se ha alcanzado un nivel aproximado al 100 % desde hace algunos años. Por su parte, la tasa de abandono , aunque aparentemente oscila en un mismo nivel, no parece decrecer hacia su nivel esperado, lo que podría indicar que las acciones implementadas para atacar este indicador no van en el sentido correcto. Y en el caso de la tasa de absorción, aparentemente no ha dejado de decrecer a través de los años, lo que muestra que tiene un comportamiento totalmente opuesto al deseado y podría requerir algunas estrategias de intervención si lo que se desea es lograr que comience a crecer.

AGRADECIMIENTOS

Los autores agradecen al Consejo Nacional de Ciencia y Tecnología por la beca para los estudios de posgrado.

APÉNDICE

Código R

```
#Librería para exportar documentos formato .xls y .xlsx
library(readxl)
#Librerías para pronósticos
library(xts)
library(astsa)
library(forecast)
#Exportar y visualizar datos de excel
data1<-read_excel("data1.xlsx")
View(data1)
#Estampa de tiempo
```

```
ab<-ts(data1$abs,start = 1994,end = 2017)
co<-ts(data1$cob,start = 1994,end = 2017)
aba<-ts(data1$abn,start = 1994,end = 2017)
#Funciones ACF y PACF
Acf(ab)
Pacf(ab)
Acf(co)
Pacf(co)
Acf(aba)
Pacf(aba)
#Modelos y pronóstico
fit1<-auto.arima(ab) #Encontrar modelo automático para absorción
summary(fit1)
plot(forecast(fit1,h=3))
fit2<-auto.arima(co) #Encontrar modelo automático para cobertura
summary(fit2)
plot(forecast(fit2,h=3))
fit3<-auto.arima(aba) #Encontrar modelo automático para abandono
summary(fit3)
plot(forecast(fit3,h=3))
plot(forecast(Arima(y=aba,order=c(1,1,0)),n.ahead=3)
co.m2<-arima(x=co, order=c(1,1,1)) #Evaluar modelo ARIMA (1,1,1) para cobertura
summary(co.m2)
co.m3<-arima(x=co,order = c(1,1,0)) #Evaluar modelo ARIMA (1,1,0) para cobertura

summary(co.m3)
co.m4<-arima(x=co,order = c(0,1,1)) #Evaluar modelo ARIMA (0,1,1) para cobertura
summary(co.m4)
ab.m2<-arima(x=ab, order = c(1,1,1)) #Evaluar modelo ARIMA (1,1,1) para absorción
summary(ab.m2)
ab.m3<-arima(x=ab, order = c(0,1,1)) #Evaluar modelo ARIMA (0,1,1) para absorción
summary(ab.m3)
aba.m2<-arima(x=aba, order = c(1,0,0)) #Evaluar modelo ARIMA (1,0,0) para abandono
summary(aba.m2)
aba.m3<-arima(x=aba, order = c(1,1,0)) #Evaluar modelo ARIMA (1,1,0) para abandono
summary(aba.m3)
aba.m4<-arima(x=aba, order= c(1,1,1)) #Evaluar modelo ARIMA (1,1,1) para abandono
summary(aba.m4)
ab.pred<-predict(fit1,n.ahead = 3) #Predicción para absorción
ab.pred
co.pred<-predict(fit2,n.ahead = 3) #Predicción para cobertura
co.pred
aba.pred<-predict(aba.m3,n.ahead = 3) #Predicción para abandono
aba.pred
```

REFERENCIAS

[1] CEPAL, "Confiabilidad y utilidad para la evaluación de indicadores." Chile, 2018.

[2] A. Márquez Jiménez, "Sistemas de indicadores educativos: su utilidad en el análisis de los problemas educativos," *Rev. Electrónica Sinéctica*, vol. 2006, no. 35, pp. 1–25, 2010.

[3] O. Hernández *et al.*, "La educación superior en México: un estudio comparativo," *Cienc. Ergo Sum*, vol. 21, no. 3, pp. 181–192, 2014.

[4] SEP, "Reporte de Indicadores Educativos," Mexico, 2018.

[5] V. Medrano Camacho, R. R. Rojas Olmos, E. E. Valencia López, C. Mexicano Melgar, E. G. Ángeles Méndez, and R. M. Bautista Espinosa, "Panorama educativo de México 2016. Indicadores del Sistema Educativo Nacional," pp. 34–35, 2017.

[6] C. E. Montenegro Marín, A. P. Gallego Torres, and P. Rocha Salamanca, "Evaluation model for stages of training in statistics in engineering," *Rev. Logos Cienc. Tecnol.*, vol. 8, no. 1, 2016.

[7] M. P. González Casimiro, *Análisis de series temporales económicas: modelos ARIMA*. 2017.

[8] J. Hernandez, *Modelación ARIMA*. 2014.

[9] D. A. López, N. Y. García, and J. F. Herrera, "Desarrollo de un modelo predictivo para la estimación del comportamiento de variables en una infraestructura de red," *Inf. Tecnol.*, vol. 26, no. 5, pp. 143–154, 2015.

[10] S. De la Fuente, *Series Temporales: Modelo Arima*. 2016.

[11] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: The forecast package for R," *J. Stat. Softw.*, vol. 27(1), pp. 1–22, 2008.