Ronda clínica y epidemiológica. Series de tiempo

Alba Luz León-Álvarez[1], Jorge Iván Betancur-Gómez[2], Fabián Jaimes-Barragán[3], Hugo Grisales-Romero[4]

RESUMEN

El análisis de series de tiempo es una técnica que involucra el estudio de individuos o grupos observados en momentos sucesivos en el tiempo. Este tipo de análisis permite estudiar la relación potencialmente causal entre diferentes variables que cambian en el tiempo y que se relacionan entre sí. Es la técnica más importante para hacer inferencias acerca del futuro, predicción, con base en lo que ha ocurrido en el pasado y se aplica en diferentes disciplinas del conocimiento. Se exponen los diferentes componentes, la técnica de análisis y algunos ejemplos específicos en el área de la salud.

PALABRAS CLAVE

Autorregresivo integrado de medias móviles; Nivel; Series de tiempo; Tendencia

SUMMARY

Clinical and epidemiological rounds. Time series

Analysis of time series is a technique that implicates the study of individuals or groups observed in successive moments in time. This type of analysis allows the study of potential causal relationships between different variables that change over time and relate to each other. It is the most important technique to make inferences about the future, predicting, on

[1] Profesora de cátedra, Facultad Nacional de Salud Pública, Universidad de Antioquia. Grupo Académico de Epidemiología Clínica (GRAEPIC), Medellín, Colombia.

[2] Administrador de empresas.

[3] Profesor titular, Grupo Académico de Epidemiología Clínica (GRAEPIC), Departamento de Medicina Interna, Facultad de Medicina, Universidad de Antioquia. Investigador, Unidad de Investigaciones, Hospital Pablo Tobón Uribe, Medellín, Colombia.

[4] Profesor titular, Grupo de Investigación Demografía y Salud, Departamento de Ciencias Básicas, Facultad Nacional de Salud Pública, Universidad de Antioquia, Medellín, Colombia. Correspondencia: Fabián A. Jaimes; fabian.jaimes@udea.edu.co

Recibido: abril 13 de 2016
Aceptado: abril 22 de 2016

Cómo citar: León-Álvarez AL, Betancur-Gómez JI, Jaimes-Barragán F, Grisales-Romero H. Ronda clínica y epidemiológica. Series de tiempo. Iatreia. 2016 Jul-Sep;29(3):373-381. DOI 10.17533/udea.iatreia.v29n3a12.

the basis or what has happened in the past and it is applied in different disciplines of knowledge. Here we discuss different components of time series, the analysis technique and specific examples in health research.

KEY WORDS

Autoregressive integrated moving average; Level; Time series; Trend

RESUMO

Ronda clínica e epidemiológica. Séries de tempo

A análise de séries de tempo é uma técnica que envolve o estudo de indivíduos ou grupos observados em momentos sucessivos no tempo. Este tipo de análise permite estudar a relação potencialmente causal entre diferentes variáveis que mudam no tempo que se relacionam entre si. É a técnica mais importante para fazer inferências sobre o futuro, predição, com base no que há acontecido no passado e se aplica em diferentes disciplinas do conhecimento. Se expõe os diferentes componentes, a técnica de análise e alguns exemplos específicos na área da saúde.

PALAVRAS CHAVE

Autorregressivo Integrado de Médias Móveis; Nível; Séries de Tempo; Tendência

INTRODUCCIÓN

Cada vez es más común encontrar en la literatura científica técnicas de análisis complejas o poco utilizadas en el área de la salud. El análisis de series de tiempo se usa para pronosticar eventos futuros de acuerdo con el comportamiento observado de diferentes variables, llamadas componentes, de ocurrencia en el pasado. Esta condición de variables componentes en vez de las clásicas variables independientes lo clasifica como un modelo dinámico. Es una técnica de modelación utilizada en diferentes áreas del conocimiento como la economía, la física, la química y el mercadeo, entre otras. En el área específica de la salud, su aplicación es bastante amplia y se enfoca principalmente en patrones transitorios de causas de enfermedad y muerte, pronóstico de la incidencia de enfermedades infecto-contagiosas, detección de epidemias y evaluación de intervenciones, todo lo cual convierte este análisis en una poderosa herramienta para comprender diferentes sucesos biomédicos.

GENERALIDADES

El análisis de series de tiempo tiene como ventajas fundamentales que puede integrar datos longitudinales, es decir, de ocurrencia repetida en determinados períodos y también encontrar las tendencias antes y después de un suceso o intervención. Como principales desventajas se describe que su uso es inadecuado cuando las tendencias de datos no son lineales y cuando las intervenciones se realizan más de una vez (1).

Para entender en qué consiste este tipo de análisis revisaremos los elementos que componen una serie de tiempo. Es necesario resaltar que la serie de tiempo se refiere a datos estadísticos que se recopilan, observan y registran en intervalos regulares; y entre sus objetivos más importantes está la capacidad de determinar si se presentan patrones no aleatorios en la ocurrencia de dichos datos y también pronosticar movimientos futuros. Las series constan de los siguientes componentes: tendencia, ciclo, estacionalidad y movimientos irregulares o aleatorios (2) (figura 1).

Tendencia: indica la dirección hacia la cual se dirige la serie de tiempo, característica que lo convierte en el componente más importante. Puede ser creciente, decreciente, constante, lineal, curvilínea, entre otras; se llama también tendencia a largo plazo y se representa con T_t.

Ciclo: indica las variaciones que ocurren en una serie de tiempo en períodos más prolongados. Cuando la métrica es en años son variaciones mayores de un año, comúnmente de 2 a 10 años. La serie sube y baja suavemente a manera de onda siguiendo la tendencia. Dicho ciclo puede ser causado por diversos cambios y se representa con C_t.

Estacionalidad: indica las variaciones que ocurren a corto plazo en una serie de tiempo con respecto a la línea de tendencia general. Ocurre en períodos fijos como días, semanas, meses, trimestres o años y se representa con E_t.

Movimientos irregulares (aleatorios): son oscilaciones de una serie temporal a corto plazo y que se atribuyen a factores imprevisibles o aleatorios. Corresponde al efecto de diversos factores a menudo desconocidos y se representa con A_t.

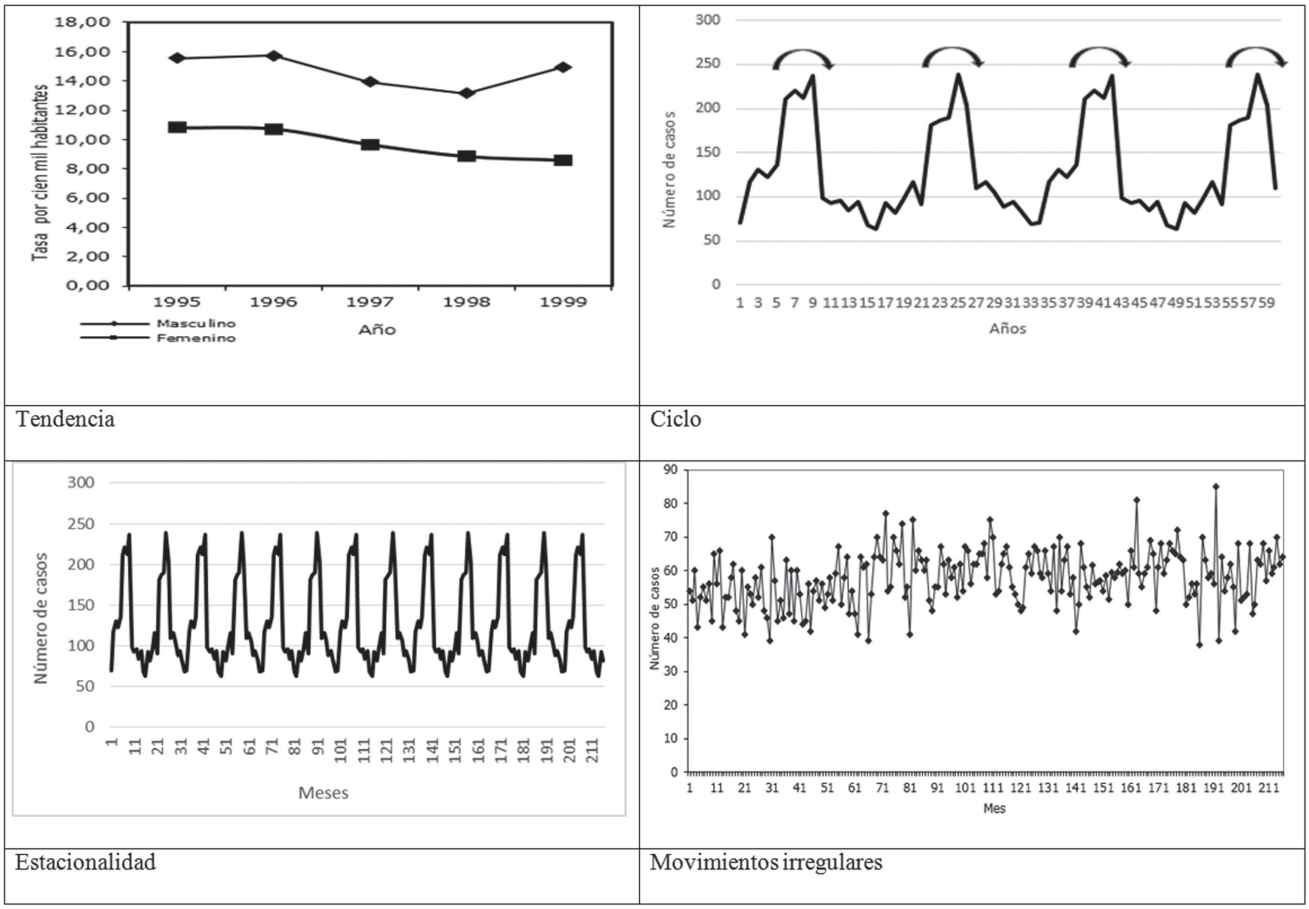


Figura 1. Componentes de una serie de tiempo (Fuente: Análisis de series de tiempo de eventos de mortalidad y morbilidad, Bogotá 1982-1999. Grisales H, Hoyos C, López A, Hincapié D, Bello L. Datos no publicados)

PASOS EN LA CONSTRUCCIÓN DE UNA SERIE DE TIEMPO

Un esquema de la organización para el diseño y análisis de una serie de tiempo es el siguiente:

- Elaborar un gráfico de secuencia, lo que permite observar si la serie tiene todos los componentes mencionados anteriormente.
- Identificar el modelo como una suma o un producto de sus componentes, lo cual depende básicamente del componente estacional. En un modelo aditivo, en el que todos los componentes suman de manera equivalente e independiente, el componente estacional no varía o se mantiene constante a pesar de la tendencia. En el modelo multiplicativo o mixto, por el contrario, la

estacionalidad varía de manera creciente o decreciente en forma proporcional a la tendencia, dado que sus componentes no son independientes entre sí (figura 2).

$$X_{(t)} = T_{(t)} + E_{(t)} + C_{(t)} + A_{(t)}$$

Modelo aditivo

$$X_{(t)} = T_{(t)} \bullet E_{(t)} \bullet C_{(t)} \bullet A_{(t)}$$

Modelo multiplicativo

$$X_{(t)} = T_{(t)} + E_{(t)} \bullet C_{(t)} \bullet A_{(t)}$$

Modelo mixto

Donde:

$X_{(t)}$ serie observada en el instante t

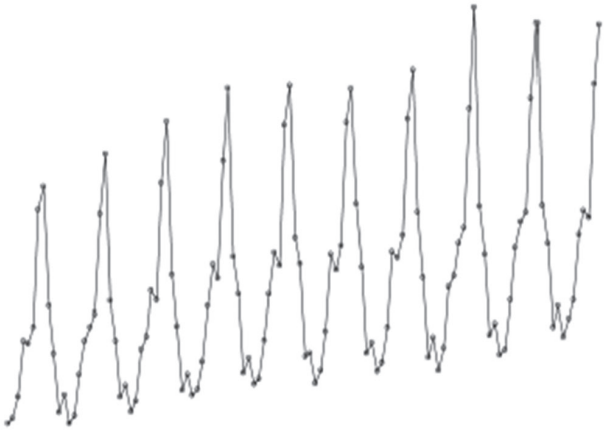
$T_{(t)}$ componente de tendencia

$E_{(t)}$ componente estacional

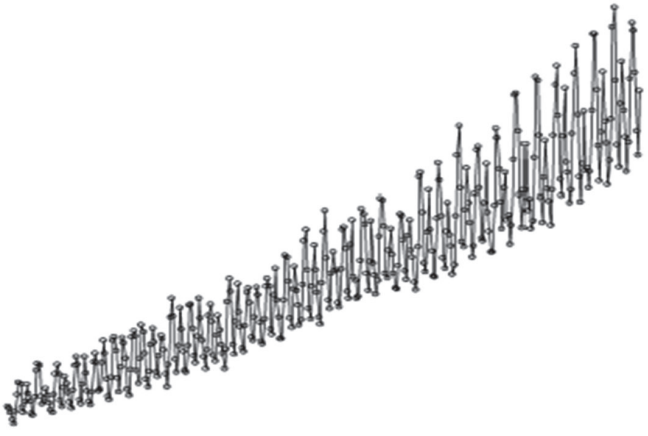
$C_{(t)}$ componente de ciclo

$A_{(t)}$ componente aleatorio

- Descomponer la serie, lo que se consigue calculando la tendencia T_t y su componente estacional E_t a partir de X*t* (el error A*t* se obtiene directamente de los anteriores). Para la descomposición de la serie existen dos grupos de métodos: los paramétricos, específicamente el de mínimos cuadrados y el de modelos ajustados; y los no paramétricos, también conocidos como de medias móviles. Los métodos paramétricos siguen los supuestos y restricciones de los modelos lineales generales, mientras los no paramétricos ofrecen la flexibilidad necesaria para el manejo de los datos con otras distribuciones. En los siguientes apartados profundizaremos en el modelo ARIMA como la más representativa de las técnicas no paramétricas.



Modelo aditivo
Asume que los componentes de la serie son independientes



Modelo multiplicativo o mixto
Asume que los componentes de la serie no son independientes

Figura 2. Modelos aditivo, multiplicativo o mixto en una serie de tiempo (Fuente: Análisis de series de tiempo de eventos de mortalidad y morbilidad, Bogotá 1982-1999. Grisales H, Hoyos C, López A, Hincapié D, Bello L. Datos no publicados)

TÉCNICA DE ANÁLISIS

El modelo ARIMA (Autoregressive integrated moving average)

Su nombre se deriva de sus componentes: AR (Autorregresivo), I (Integrado) y MA (Medias móviles). Es un modelo que utiliza variaciones y regresiones de los datos con el fin de encontrar patrones para una predicción hacia el futuro y es, además, dinámico, es decir, que las estimaciones futuras se explican por los datos del pasado y no por otras variables independientes. En el modelo ARIMA la secuencia convencional se inicia con la exploración de los datos, se continúa con la detección de las características de la serie, se ajusta el modelo, se hace un análisis de sensibilidad y finalmente se realizan las estimaciones y las predicciones.

Es importante aclarar que un modelo ARIMA como estrategia de predicción solo tendrá sentido si las características observadas en la serie permanecen en el tiempo. Lo anterior define el concepto de un modelo *Estacionario*, que se fundamenta en dos requisitos:

1. Que la media y la varianza de los datos sean constantes en el tiempo y,
2. Que la estructura de la covarianza (el grado de variación conjunta de dos variables aleatorias) entre dos períodos diferentes de tiempo dependa solamente de la distancia o rezago entre estos dos períodos y no del tiempo en el que se ha calculado dicha covarianza (3).

Decimos que un proceso o modelo es estacionario en sentido estricto si las funciones de distribución conjuntas como la media, las varianzas, las covarianzas y las funciones de distribución "completas", son constantes o invariantes con respecto al desplazamiento en el tiempo.

La notación utilizada en ARIMA es (*p,d,q*), donde los parámetros *p, d* y *q* son números enteros no negativos que indican el orden de los distintos componentes en el modelo así: *p* es el orden de la autorregresión, *d* es el grado de diferenciación y *q* es el orden de la media móvil considerada.

Autorregresión: indica que cada valor en la serie es una función lineal de momentos anteriores. En otras palabras, el comportamiento de la variable en cualquier momento está influenciado por las observaciones de la propia variable (actual o pasada) (4). De esta manera, el valor del parámetro *p* estará dado por el número de momentos contiguos que están antes del momento observado. Así, si al momento observado solo lo antecede un momento se dirá que hay autocorrelación de orden 1 y esto se expresará en el modelo ARIMA como (1,*d,q*), indicando a su vez solo un coeficiente de correlación en la ecuación. En cambio, si suponemos que al momento observado lo anteceden dos momentos, se dirá que hay autocorrelación de orden 2, la expresión será ARIMA (2,*d,q*) y la ecuación tendrá dos coeficientes de correlación (5). Para la identificación de este parámetro se usa la función de autocorrelación simple (ACF, por la sigla en inglés), que mide la correlación entre los valores de una secuencia temporal X_t, distanciados en un lapso de tiempo k. A este tiempo k se le conoce como retardo, retraso o rezago. El rezago denota el período entre los valores de la secuencia para la cual se miden el tipo y grado de autocorrelación de la variable considerada.

En la figura 3 se observa cómo cada barra representa el valor tomado por el coeficiente de autocorrelación simple para el retardo correspondiente, desde 1 hasta k. Para considerar un coeficiente como significativamente diferente de 0, su valor debe rebasar los límites de confianza trazados por la línea punteada alrededor del 0. Basta la aparición de autocorrelación simple de la variable en al menos un retardo cualquiera, para considerar la secuencia temporal como correlacionada y de este modo definir la categoría como una serie (6).

Diferenciación: es el proceso que busca detectar el componente estacionario de la serie, con el propósito de no confundir la tendencia propia con el efecto potencial de cualquier intervención o exposición externa. En cierta forma trata de "eliminar" la tendencia, dado que se supone que esta evoluciona lentamente en el tiempo con un efecto acumulativo (5). Diferenciar una serie consiste simplemente en restar a cada valor dentro de ella el anterior valor correspondiente, es decir, la segunda menos la primera, la tercera menos la segunda y así sucesivamente. Por tanto, el parámetro *d* indica el número de veces que una serie ha de ser diferenciada para hacerla estacionaria. Por ello, si solo se hace la resta una vez el modelo se deberá expresar como ARIMA (*p*,1,*q*) (6).

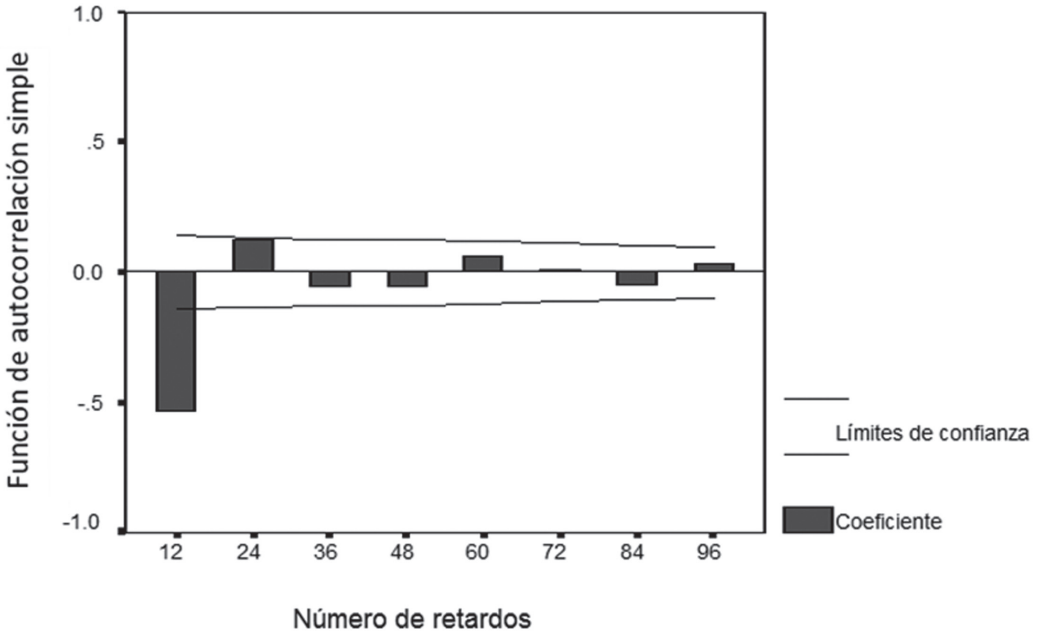


Figura 3. Función de autocorrelación simple (ACF) (Fuente: Análisis de series de tiempo de eventos de mortalidad y morbilidad, Bogotá 1982-1999. Grisales H, Hoyos C, López A, Hincapié D, Bello L. Datos no publicados)

Media móvil: indica el número de componentes aleatorios previos que configuran el valor actual en la serie temporal. Dicho de otra manera, se considera que el comportamiento de la variable en cualquier momento está influenciado además por los errores o elementos aleatorios actuales o pasados (4). Este elemento representa la innovación en la serie, la parte que no responde al comportamiento histórico anterior, sino que es nueva en cada período (6). Con este componente se incluye una parte irregular y aleatoria que regula el fenómeno. Así, si existen dos componentes aleatorios el modelo se expresará como ARIMA (*p,d,*2). Para esta identificación se usa la función de autocorrelación parcial (ACFP, por la sigla en inglés), que está constituida por el conjunto de coeficientes de autocorrelación desde el retardo k = 1 hasta el máximo posible de retardos en la mitad de la cantidad de valores contenidos en la secuencia. El coeficiente de autocorrelación parcial no considera las correlaciones acumuladas hasta el retardo k para el que se estima (5). La lectura que se hace de la ACFP es idéntica a la explicada para la ACF; es decir, que para considerar un coeficiente como significativamente diferente de 0, su valor debe rebasar los límites de confianza trazados alrededor del 0 (figura 4).

Un modelo ARIMA puede incluir cualquiera de los anteriores parámetros, aislados o en diferentes combinaciones (4). Es posible también identificar más de un modelo tentativo para una misma serie, es decir, diferentes valores para *p, d* y *q*. La elección del modelo final se establece a partir del análisis de los residuos producidos por el modelo: si este chequeo de residuales conduce al rechazo del modelo tentativo, se debe volver a la identificación inicial de los parámetros (7). En la práctica, un modelo ARIMA se construye desarrollando los siguientes cuatro pasos (7):

Paso 1. Identificación

Se trata de determinar los parámetros *p, d* y *q*. Cabe aclarar que aunque estos pueden tomar cualquier valor, en la mayoría de las situaciones reales dichos

valores serán 0 y 1, lo que hace el proceso de identificación menos complejo de lo que parece. El primer parámetro para determinar es *d*; para ello se analiza gráficamente si la serie es estacionaria. Normalmente es suficiente con observar si los valores de la media y la varianza de la serie de datos se mantienen constantes y si esto no sucede se procede a diferenciarla, una o varias veces hasta que sea necesario.

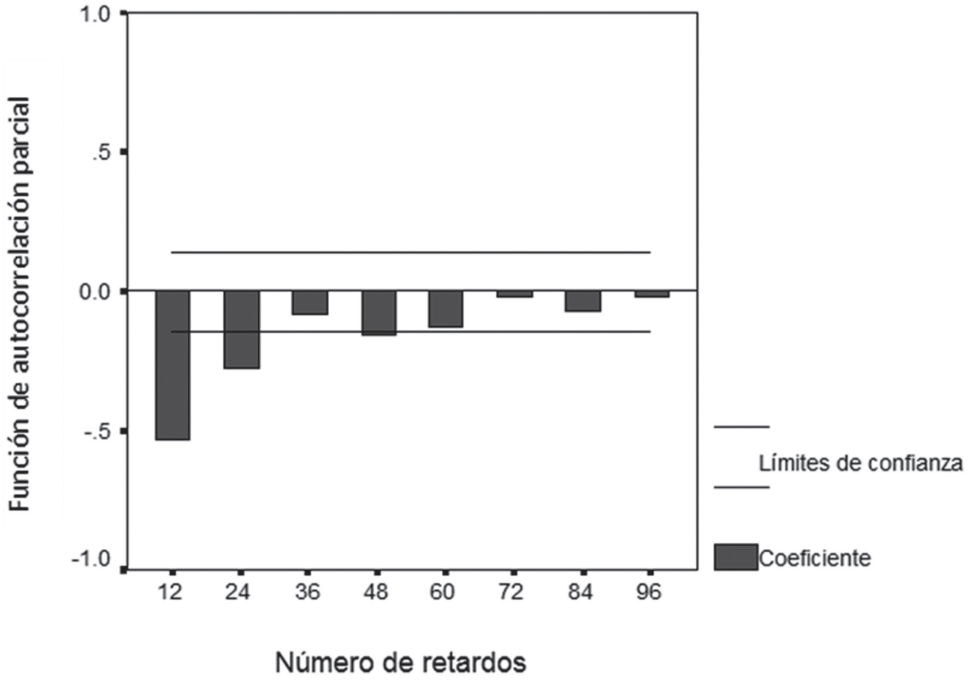


Figura 4. Función de autocorrelación parcial (ACFP) (Fuente: Análisis de series de tiempo de eventos de mortalidad y morbilidad, Bogotá 1982-1999. Grisales H, Hoyos C, López A, Hincapié D, Bello L. Datos no publicados).

La identificación de *p* y *q* se consigue desarrollando la función de autocorrelación simple (ACF) y la función de autocorrelación parcial (ACFP), respectivamente, que ya se explicaron. Estas son las principales herramientas de diagnóstico para identificar conjuntos de modelos y evaluar su poder predictivo (5). De esta manera, cuando se tiene una serie temporal estacionaria se estiman y grafican los valores de la ACF y la ACFP. Adicionalmente, es necesario marcar los intervalos de confianza para poder detectar los valores significativos. Si en la ACF los primeros valores son distintos de cero y los restantes son cero o muy próximos a cero, y la ACF presenta un decrecimiento exponencial y/o un comportamiento sinusoidal, el parámetro *p* corresponderá al número de valores distintos de cero. Para el parámetro *q* se hace el mismo análisis, pero con la gráfica de la función de ACFP.

Paso 2. Estimación

Dado que los modelos ARIMA no son lineales en sus parámetros, se deben utilizar modelos iterativos y de máxima verosimilitud como procedimientos de estimación para obtener los valores exactos de los coeficientes. En este paso se determinan dichos valores, correspondientes al número de parámetros *p, d* y *q*, con sus respectivos errores estándar e intervalos de confianza.

Paso 3. Diagnóstico

Se comprueba si el modelo es adecuado por medio de la ausencia de estructura en los residuales. Esto quiere decir que las diferencias entre los valores observados y los predichos en la serie deben seguir una distribución aleatoria, que es lo que se conoce comúnmente como "ruido blanco". En términos estadísticos, se refiere al caso de procesos estocásticos o no deterministas donde los valores de las variables son independientes e idénticamente distribuidos a lo largo del tiempo, con media cero e igual varianza. Si esto no ocurre, es decir, si se observa una distribución diferente en los residuales, se debe modificar la estructura del modelo y repetir los pasos anteriores.

Paso 4. Predicción

Una vez se obtiene el modelo y se comprueba su validez, se puede proceder a efectuar las predicciones pertinentes.

APLICACIONES

Dunlop y colaboradores (8) publicaron en el año 2014 un estudio llevado a cabo para conocer el impacto de la política de empaquetado genérico para todos los productos de tabaco en Australia. El estudio se hizo entre abril de 2006 y mayo de 2013 y la intervención se aplicó el primero de octubre de 2012 en una población base de 15 745 personas mayores de 18 años. La hipótesis fue que después de la introducción de los nuevos paquetes, los fumadores encontrarían las advertencias sanitarias más llamativas, tendrían un aumento en la respuesta a las mismas y menos percepciones favorables de los cigarrillos por lo menos hasta 6 meses después de la intervención. Los resultados del modelo ARIMA mostraron que después de los 2 y 3 meses de la introducción de los nuevos paquetes hubo un aumento significativo en la proporción de fumadores que tenían fuerte respuesta cognitiva, emocional y de evitación asociada con las advertencias gráficas (tabla 1). El diagnóstico del modelo indicó un adecuado ajuste.

Tabla 1. Resultados del análisis de series de tiempo para la percepción "en desacuerdo" con ciertas consideraciones acerca del tabaco (Australia, octubre 2011- mayo 2013)

Percepciones acerca del producto	"Es atractivo"	"Dice algo bueno de mí"	"Está de moda"	"Coincide con mi estilo"
Incremento en el porcentaje de fuerte desacuerdo	57,5	54,5	44,7	48,1
Intervalo de confianza del 95 %	38,0-77,1	36,9-72,1	28,1-61,2	32,2-64,0

Adaptación de Dunlop SM, et al. BMJ Open 2014;4:e005836

En febrero de 2016 fue publicado un estudio de series de tiempo para la vigilancia de sífilis congénita, primaria, secundaria, terciaria y latente en China entre 2005 y 2012 (9). El modelo ARIMA que se estableció expresaba el carácter estacional de la incidencia de la sífilis en cada serie, con un comportamiento general que mostró mayor incidencia en verano (junio, julio, agosto) que en invierno (diciembre, enero, febrero). Además, basados en un aumento de tres veces la incidencia de sífilis entre 2005 y 2012, el modelo estima un aumento constante en la tendencia a largo plazo de la enfermedad. Las conclusiones del estudio sugieren que el método es una herramienta eficaz para modelar la incidencia histórica y futura de la sífilis en ese país, lo cual puede aportar para la planificación de programas de promoción y prevención.

Ruiz y colaboradores (10) evaluaron el impacto de un programa de reemplazo e intercambio de jeringas para la prevención del VIH en Washington DC. Utilizando los datos de vigilancia de casos de VIH por uso de drogas inyectables entre septiembre de 1996 y diciembre de 2011, los autores construyeron un modelo ARIMA (0,0,1) x (0,0,1) para predecir el número de casos asociados a esta conducta en los 24 meses siguientes a la aplicación del programa. Se compararon los valores pronosticados con los casos observados durante el mismo período para calcular el número de casos

evitados. La inspección del gráfico mostró una tendencia a la disminución de nuevos casos de VIH asociados al uso de drogas inyectables antes y después de la implementación del programa. Los datos de vigilancia reportaron 176 casos de VIH asociados a la utilización de drogas inyectables en los 2 años siguientes a la aplicación del programa. En contraste, el modelo ARIMA predijo 296 infecciones por VIH, lo cual sugiere que gracias al programa se evitaron 120 casos de infección.

CONCLUSIONES

El análisis de series de tiempo es un diseño que cada vez toma más fuerza en diferentes disciplinas del conocimiento y su gran capacidad de predicción lo hace particularmente atractivo para el área de la salud. Esta técnica puede ayudar a entender y explicar muchas situaciones médicas y de salud pública de difícil manejo. Con una adecuada aplicación es posible hacer inferencias adecuadas acerca de fenómenos desconocidos o poco explorados en el ámbito de las ciencias biomédicas.

AGRADECIMIENTOS

Agradecemos muy especialmente a las estudiantes Catalina Hoyos y Ana María López; además, a los profesores Doracelly Hincapié, León Darío Bello y Hugo Grisales, por facilitarnos las figuras para esta publicación.

FINANCIACIÓN

Trabajo apoyado parcialmente por la Estrategia de Sostenibilidad de la Universidad de Antioquia, 2013-2014.

REFERENCIAS BIBLIOGRÁFICAS

1. Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BMJ. 2015;350:h2750. DOI 10.1136/bmj.h2750.
2. Box GE, Jenkins GM, Reinsel GC. Operational research quarterly. In: Time Series Analysis: Forecasting and Control. 4th ed. Hoboken, NJ: Wiley; 2008. p. 137-91.
3. Asteriou D, Hall SG. ARIMA Models and the Box–Jenkins Methodology. In: Applied Econometrics. 2nd ed. London: Palgrave Macmillan; 2011. p. 265-86.
4. Coutin Marie G. Utilización de modelos ARIMA para la vigilancia de enfermedades transmisibles. Rev Cubana Salud Pública. 2007 Abr-Jun;33(2):1-11.
5. Grisales H. Una aplicación de los modelos ARIMA en la predicción de la mortalidad por ataque con arma de fuego y explosivos para la ciudad de Medellín, de 1997 al año 2000. Rev Fac Nac Salud Pública. 1999 Ene-Jun;16(2):30-49.
6. Guisande C, Vaamonde A, Barreiro A. Series temporales. En: Tratamiento de datos con R, Statistica y SPSS. España: Díaz de Santos; 2011. p. 585-637.
7. Aguirre A. La modelación ARIMA. ¿Es la serie estacionaria? En: Introducción al tratamiento de series temporales: aplicación a las ciencias de la salud. España: Díaz de Santos; 1994. p. 173-213.
8. Dunlop SM, Dobbins T, Young JM, Perez D, Currow DC. Impact of Australia's introduction of tobacco plain packs on adult smokers' pack-related perceptions and responses: results from a continuous tracking survey. BMJ Open. 2014 Dec;4(12):e005836. DOI 10.1136/bmjopen-2014-005836. Erratum in: BMJ Open. 2015;5(8):e005836corr1.
9. Zhang X, Zhang T, Pei J, Liu Y, Li X, Medrano-Gracia P. Time Series Modelling of Syphilis Incidence in China from 2005 to 2012. PLoS One. 2016 Feb;11(2):e0149401. DOI 10.1371/journal.pone.0149401.
10. Ruiz MS, O'Rourke A, Allen ST. Impact Evaluation of a Policy Intervention for HIV Prevention in Washington, DC. AIDS Behav. 2016 Jan;20(1):22-8. DOI 10.1007/s10461-015-1143-6.