Universidad Panamericana
Maestría en Ciencia de Datos

Desarrollo de un modelo predictivo ARIMA para la evaluación de la tasa de violencia familiar en México durante la pandemia por covid-19

Sarahí Aguilar González
0189970@up.edu.mx

Introducción

Como lo indica la Norma Oficial Mexicana NOM-046-SSA2-2005, la violencia familiar es un problema de salud pública por sus altos niveles de prevalencia y por las repercusiones perjudiciales e incluso letales en las víctimas. La violencia familiar representa un obstáculo significativo en el sano desarrollo de la convivencia social y el ejercicio de los derechos humanos. [1]

A través de múltiples investigaciones, la magnitud y la gravedad de las consecuencias de a violencia familiar se han vuelto evidentes. El abuso de poder daña la integridad del ser humano, desde lo biológico, como en lo psicológico y lo social. [1]

La violencia familiar se ejerce en todos los ámbitos y cualquier persona es susceptible. Sin embargo, las estadísticas indican que los niños, las niñas y las mujeres son los grupos que viven un mayor número de situciones de violencia familiar. La causa está en función de la edad y la desigualdad de género que le otorga un valor superior al género masculino sobre el femenino. [1]

Desde el brote de covid-19, datos e informes emergentes han demostrado que todos los tipos de violencia contra las mujeres y las niñas, en particular la violencia familiar, se han intensificado. A medida de que el alza en los casos de covid-19 continúa ejerciendo presión sobre los servicios de salud, los servicios esenciales, como los refugios para víctimas de violencia familiar y las líneas de ayuda, han alcanzado su capacidad máxima. Además, los gobiernos han priorizado la rapidez en la toma de decisiones, descartando múltiples perspectivas, incluida la de igualdad de género. Por ejemplo, las órdenes de quedarse en casa limitan la propagación del virus, pero pueden resultar en una situación potencialmente peligrosa para las mujeres con parejas violentas. [2]

En el contexto nacional de la pandemia por covid-19, la violencia contra las mujeres no ha desistido. Por el contrario, cifras oficiales demuestran que ha aumentado. En México, se ha registrado un aumento representativo en asesinatos de mujeres y llamados de auxilio relacionados con violencia durante la crisis sanitaria. [3]

Uno de los indicadores de la violencia familiar en México es la tasa de incidencia de violencia familiar por cada 1,000 habitantes. Esta tasa es posible calcularla empleando los datos del Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública (SESNSP) y los datos de proyecciones de población a mitad de año de 2015 a 2020 del Consejo Nacional de

Población. A partir de lo mencionado anteriormente, se podría concluir que este indicador incrementó con la pandemia. Sin embargo, no fue así.

En la Figura 1 se visualiza la serie de tiempo de la tasa de incidencia de violencia familiar en México por cada 1,000 habitantes de enero del 2015 a octubre del 2020, y se puede observar que el comportamiento en el 2020 varía con respecto al de años anteriores. En abril del 2020, mes posterior al mes en el que iniciaron las medidas de confinamiento con la Jornada de Sana Distancia anunciada por el Gobierno de México, la tasa no solo disminuyó con respecto al mes anterior, sino que también fue menor a la tasa del mismo mes en el 2018 y el 2019.

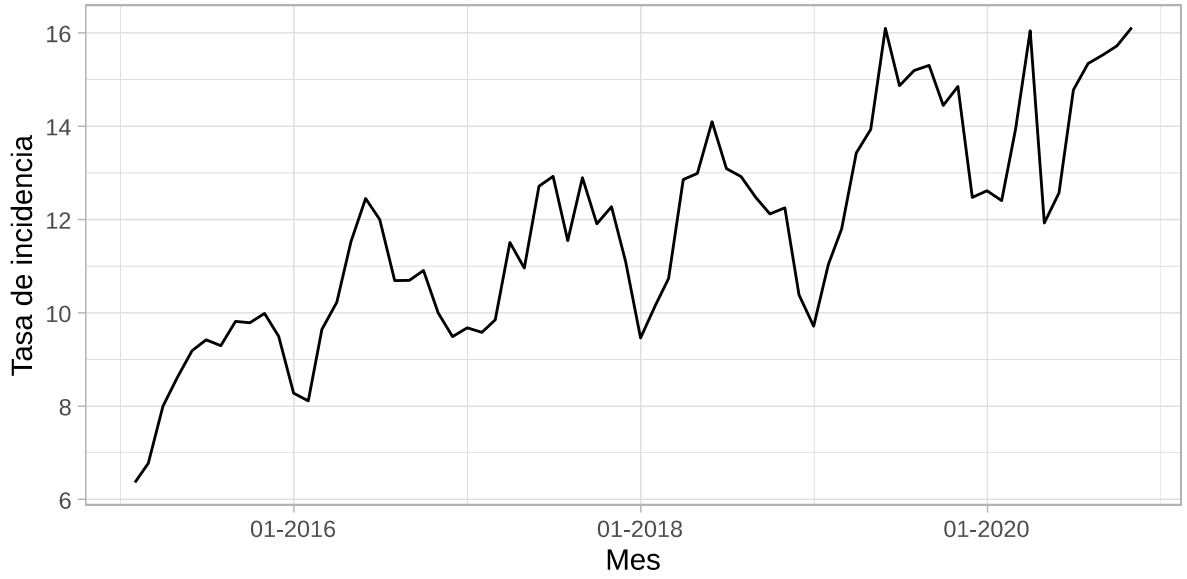

Figura 1. Tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes

Superficialmente, esto podría representar la disminución de casos de violencia familiar nacional, pero los indicadores de violencia alternos sugieren que es otra, y es probable que el decremento observado en la tasa de incidencia de violencia familiar se deba a las limitantes intrínsecas de la recopilación de datos del Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública.

El objetivo del presente trabajo es la construcción de un modelo ARIMA que estime la tasa de incidencia de violencia familiar por cada 1,000 habitantes en México de enero del 2015 a marzo del 2020 con el fin de teneutilizarlo para tener un pronóstico del comportamiento de dicha tasa de incidencia hasta octubre del 2020. Con ello, se obtendrá el escenario de la violencia familiar sin subregistro y sin efecto de las potenciales consecuencias de las medidas de confinamiento, y se podrá (1) evaluar la gravedad del subregistro, y (2) evaluar el impacto de las medidas de confinamiento en la violencia familiar comparando las cifras pronósticadas con las que el Gobierno de México pudiera actualizar en un futuro sobre la incidencia delictiva del 2020. De esta forma, se pretende contribuir en la visibilidad de la violencia familiar durante la pandemia en México y en su detección, atención, disminución y erradicación.

A continuación se presenta una más detallada descripción de la serie de datos utilizada y un análisis de la misma. Después, se procede a la sección en donde se abordan los detalles del modelado para continuar con la metodología de pronóstico y discusión de resultados. Finalmente, se presentan las conclusiones del trabajo.

Descripción de la serie de tiempo

La serie de tiempo analizada corresponde a la tasa de incidencia de violencia familiar en México por cada 1,000 habitantes de enero del 2015 a marzo del 2020.

Para obtener esta serie de tiempo, primero se emplearon las cifras mensuales de incidencia delictiva estatal publicadas por el Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública el 20 de noviembre del 2020 en el sitio oficial del Gobierno de México. De esta base se filtraron los casos por tipo de delito “Violencia familiar”.

Una vez obtenido el número de casos mensuales, se consultó la base de datos de proyecciones de población a mitad de año de 2015 a 2020 del Consejo Nacional de Población también en el sitio oficial del Gobierno de México.

La tasa de incidencia de violencia familiar en México por cada 1,000 habitantes se obtuvo con la siguiente fórmula:

*Tasa de incidencia = [Casos de violencia familiar/(Proyección de población anual/1000)]*100*

Análisis de la serie de tiempo

En la Figura 2 podemos observar la tasa de incidencia de violencia en México por cada 1,000 habitantes por año. Se puede observar un evidente incremento anual. En el 2015, la tasa de incidencia media era de 8.8, mientras que en 2019 era de de 13.8. En el 2020, la tasa de incidencia media hasta octubre es de 14.4. En la misma figura, se puede observar de 2015 a 2019 un comportamiento similar de la serie de tiempo anual: la tasa de incidencia va en incremento los primeros cinco meses del año y después comienza a bajar para finales de año.

Sin embargo, en el 2020, a apesar de que su valor medio sigue siendo más alto que el del resto de los cuatro años, su comportamiento es particular: hay una caída drástica en el mes de abril, mes siguiente del que se implementaron las medidas de confinamiento en México. De abril a junio del 2020, el valor de la tasa de incidencia es menor que el del año anterior, y es hasta julio cuando la tasa de incidencia vuelve a superar las cifras del 2019.

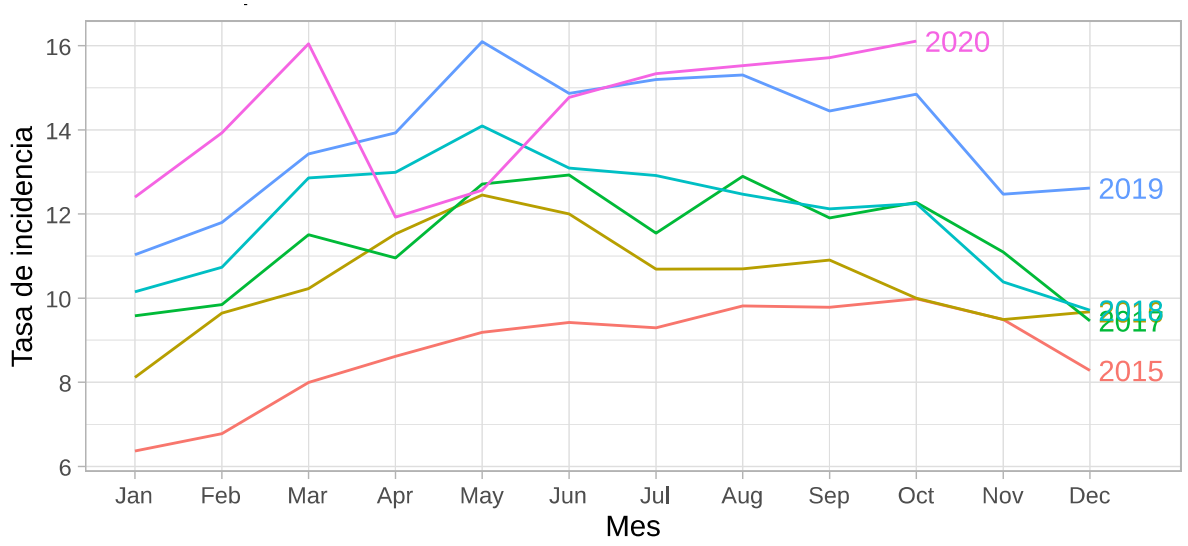

Figura 2. Tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes por año

Descomposición de la serie de tiempo por promedios móviles

Ahora bien, con el objetivo de estudiar los patrones subyacentes de la serie de tiempo de enero del 2015 a marzo del 2020, esta fue descompuesta en un componente de ciclo de tendencia, un componente estacional y un componente de residuos (que contiene cualquier otro ruído adicional en la serie de tiempo). Lo anterior se hizo mediante el método de promedios móviles considerando un componente aditivo.

En primer lugar, un promedio móvil de orden m se puede escribir como

$$T_t = 1/m * \sum_{j=-k}^{k} y_{t+j},$$

donde $m = 2k + 1$. Es decir, la estimación del ciclo de tendencia en el tiempo t se obtiene promediando los valores de la serie de tiempo dentro de k períodos de t. Es probable que las observaciones cercanas en el tiempo también tengan un valor cercano. Por lo tanto, el promedio elimina parte de la aleatoriedad en los datos, dejando un componente de ciclo de tendencia uniforme. A esto lo llamamos m-MA, que significa un promedio móvil de orden m. [4]

En segundo lugar, la serie de tiempo posee un componente aditivo pues la magnitud de las fluctuaciones estacionales, o la variación alrededor del ciclo de tendencia, no varía con el nivel de la serie de tiempo. Por lo tanto, si asumimos una descomposición aditiva, entonces podemos escribir

$$y_t = S_t + T_t + R_t,$$

donde y_t es el dato, S_t es el componente estacional, T_t es el componente del ciclo de tendencia y R_t es el componente restante, todo en el período t. [4]

En la Figura 3 podemos observar la tendencia creciente de la serie de tiempo y el componente estacional con incrementos en el mes de mayo de cada año.

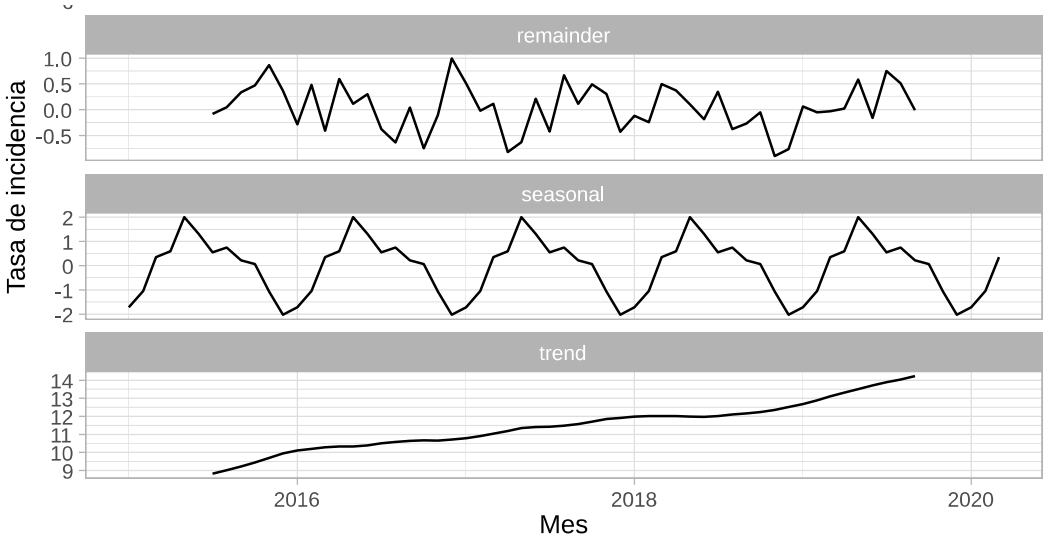

Figura 3. Descomposición de la serie de tiempo mediante promedios móviles considerando un componente aditivo de la tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes

Estacionariedad

Una serie de tiempo estacionaria es aquella cuyas propiedades no dependen del momento en el que se observa la serie. Por lo tanto, las series de tiempo con tendencias, o con estacionalidad, no son estacionarias pues la tendencia y la estacionalidad afectarán el valor de la serie de tiempo en tiempos diferentes. Por otro lado, una serie de ruido blanco es estacionaria; no importa cuando la observe, debería verse muy similar en cualquier momento. [4]

En general, una serie de tiempo estacionaria no tendrá patrones predecibles a largo plazo, y cuando se busca modelarla con un modelo ARIMA, es necesario cumplir con la condición de que se trate de un proceso estocástico estacionario. [4]

Una forma de determinar si la serie de tiempo es estacionaria es mediante una prueba de hipótesis estadísticas de raíz unitaria. [4] Hay varias pruebas de raíz unitaria disponibles, que se basan en diferentes supuestos. Para este análisis, se utilizaron las siguientes tres utilizando el entorno y lenguaje de programación R:

1. La prueba de Augmented Dickey-Fuller
2. La prueba de Phillips-Perron
3. La prueba de Kwiatkowski-Phillips-Schmidt-Shin

Para las dos primeras pruebas (Augmented Dickey-Fuller y Phillips Perron), la hipótesis nula señala que la serie de tiempo presenta una raíz unitaria y, por lo tanto, no es estacionaria. Por consiguiente, se debe buscar evidencia de que la hipótesis nula se rechace, es decir, se buscan valores *p* menores de 0.05. Por el contrario, para la tercera prueba (Kwiatkowski-Phillips-Schmidt-Shin), la hipótesis nula señala que la serie de tiempo es estacionaria; de

modo que se debe buscar evidencia de que la hipótesis nula se acepte, es decir, se buscan valores *p* mayores a 0.05. [5]

Los valores *p* obtenidos al aplicar las pruebas de hispótesis estadísticas en la serie de tiempo se pueden consultar en la Tabla 1.

Prueba de hipótesis estadística	Valor p
Augmented Dickey-Fuller Test	0.01
Phillips-Perron Unit Root Test	0.01442
Kwiatkowski-Phillips-Schmidt-Shin	0.01

Tabla 1. Valores p obtenidos al aplicar las pruebas de hipótesis estadísticas en la tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes

Los valores *p* de las dos primeras pruebas sugieren que la serie de tiempo es estacionaria, pero la tercera, sugiere lo opuesto.

Además de estas pruebas de hipótesis estadísticas de raíz unitaria, la visualización de la función de autocorrelación (ACF, por sus siglas en inglés) y la de de la función de autocorrelación parcial (PACF, pr sus siglas en inglés) también son útiles para identificar series de tiempo no estacionarias. [4]

Así como la correlación mide el alcance de una relación lineal entre dos variables, la autocorrelación mide la relación lineal entre los valores rezagados de una serie de tiempo. Hay varios coeficientes de autocorrelación, correspondientes a cada rezago. Por ejemplo, r_1 mide la relación entre y_t e y_{t-1}, r_2 mide la relación entre y_t e y_{t-2}, y así sucesivamente. El valor de r_k puede ser escrito como [4]

$$r_k = \left(\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y}) \right) / \left(\sum_{t=-1}^{T} (y_t - \bar{y})^2 \right)$$

Por su parte, la autocorrelación parcial mide la relación entre y_t y y_{t-k} después de eliminar los efectos de los rezagos 1, 2, 3, …, $k-1$. Por lo tanto, la primera autocorrelación parcial es idéntica a la primera autocorrelación. [4]

Para una serie de ruido blanco, esperamos que el 95% de los picos en la ACF estén cercanos a cero dentro de $\pm 2 / \sqrt{T}$ donde T es la longitud de la serie de tiempo. Si uno o más picos grandes están fuera de estos límites, o si sustancialmente más del 5% de los picos están fuera de estos límites, entonces la serie de tiempo no se considera ruido blanco. [4]

Cuando los datos presentan tendencia, las autocorrelaciones tienden a ser grandes y positivas porque las observaciones cercanas en el tiempo también tienen un tamaño cercano. Entonces, la ACF y la PACF de las series de tiempo con tendencia tiende a tener valores positivos que disminuyen lentamente a medida que aumentan los retrasos. Por otro lado, cuando los datos son estacionales, las autocorrelaciones serán mayores para los retrasos estacionales (en múltiplos de la frecuencia estacional) que para otros retrasos. Por lo tanto, cuando los datos tienen tendencias y son estacionales, se ve una combinación de

estos efectos. [4] Este fenómeno lo podemos en la visualización de la ACF en la Figura 4, lo cual sugiere que la serie de tiempo no es un proceso estoástico estacionario y no se puede proceder a la construcción de un modelo ARIMA.

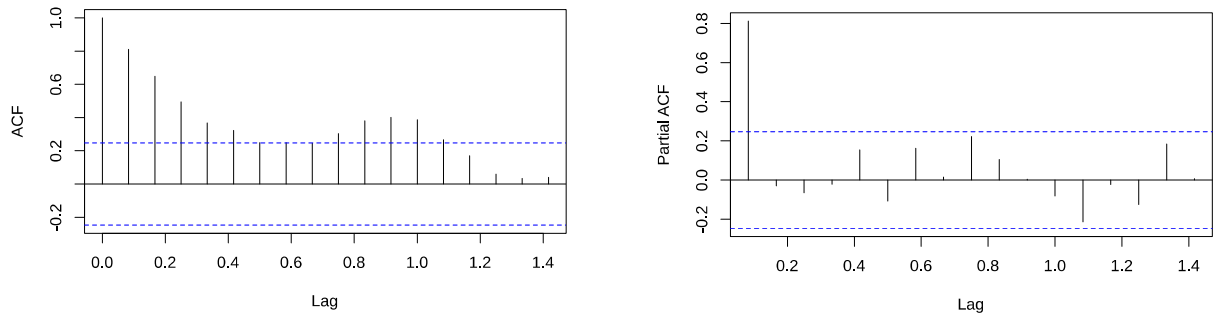

Figura 4. ACF y PACF de la tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes
Límites en azul definidos por ± 2 / √T, donde T es la longitud de la serie de tiempo

Diferenciación

Una forma de convertir una serie de tiempo no estacionaria en estacionaria es calcular las diferencias entre observaciones consecutivas. Esto se conoce como diferenciación. Las transformaciones como los logaritmos pueden ayudar a estabilizar la varianza de una serie de tiempo y la diferenciación puede ayudar a estabilizar la media de una serie de tiempo al eliminar los cambios en el nivel de una serie de tiempo y, con ello, reducir la tendencia y la estacionalidad. Adicionalmente, la diferenciación es una transformación interpretable. [4]

En la Figura 5 se puede observar la serie de tiempo tranformada con el logaritmo y en la Figura 6 se puede observar la serie de tiempo diferenciada. En esta última, visualmente es evidente la estabilización de la varianza y la eliminación de la tendencia y la estacionalidad.

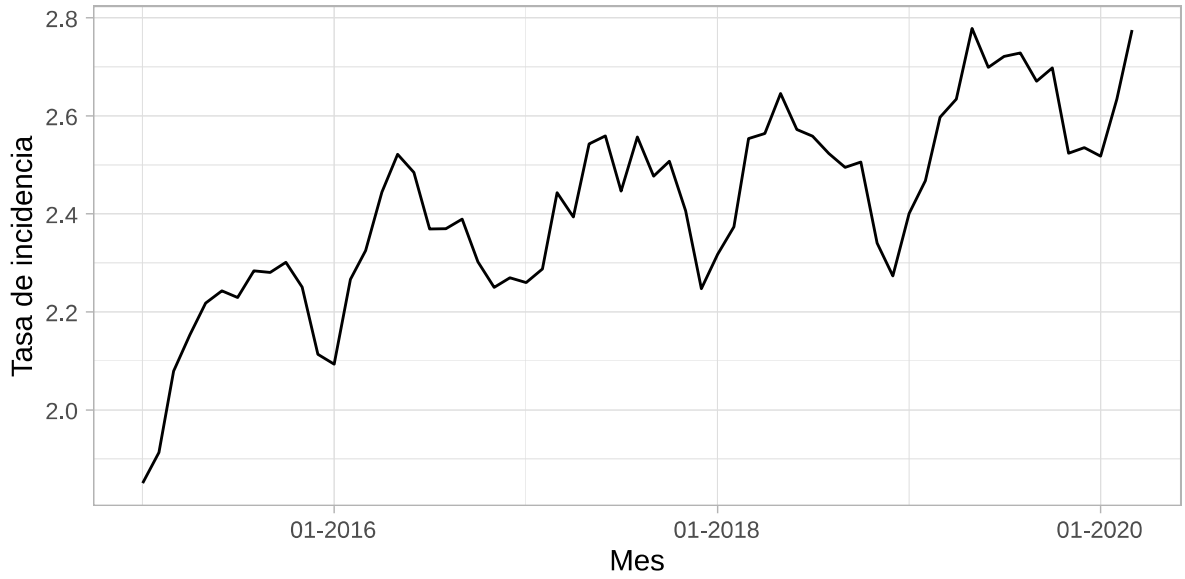

Figura 5. Logaritmos de la tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes

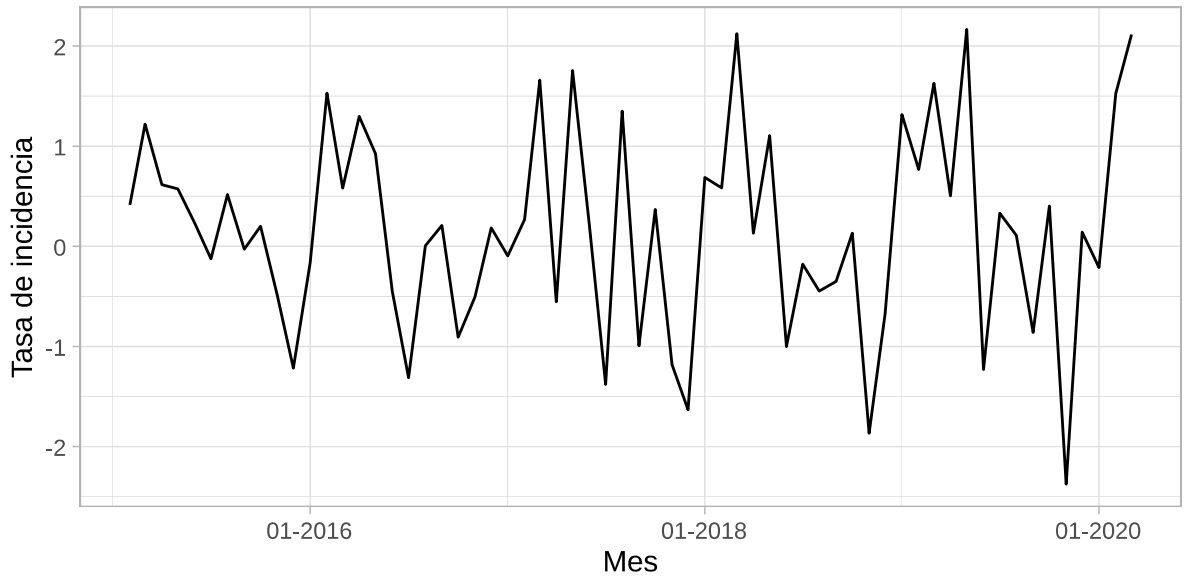

Figura 6. Diferenciación de la tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes

En la Tabla 2 se muestran los valores *p* obtenidos en las pruebas de hipótesis estadísticas aplicadas a la serie de tiempo transformada con el logaritmo y una diferenciación respectivamente.

Prueba de hipótesis estadística	**Valor *p* para la serie transformada con el logaritmo**	**Valor *p* para la serie diferenciada**
Augmented Dickey-Fuller Test	0.01	0.01
Phillips-Perron Unit Root Test	0.01442	0.01
Kwiatkowski-Phillips-Schmidt-Shin	0.01	0.1

Tabla 2. Valores p obtenidos al aplicar las pruebas de hipótesis estadísticas en la tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes trasnformada con el logaritmo y diferenciada respectivamente

Para la serie de tiempo transformada con el logaritmo, los valores *p* de las dos primeras pruebas (Augmented Dickey-Fuller y Phillips Perron) sugieren que la serie de tiempo es estacionaria, pero la tercera (Kwiatkowski-Phillips-Schmidt-Shin), sugiere lo opuesto. En contraste, para la serie de tiempo diferenciada, los valores *p* de las tres pruebas sugieren que la serie de tiempo es estacionaria, e incluso el valor *p* para la última prueba es menor que para el valor *p* de la misma prueba aplicada a la serie transformada con el logaritmo.

Así mismo, en la Figura 7, se puede observar como la ACF y la PACF de la serie de tiempo transformada con el logaritmo son muy similares a las de la serie de tiempo original. No obstante, en la Figura 8, se puede observar como la ACF y la PACF de la serie de tiempo diferenciada caen a valores pequeños con relativa rapidez.

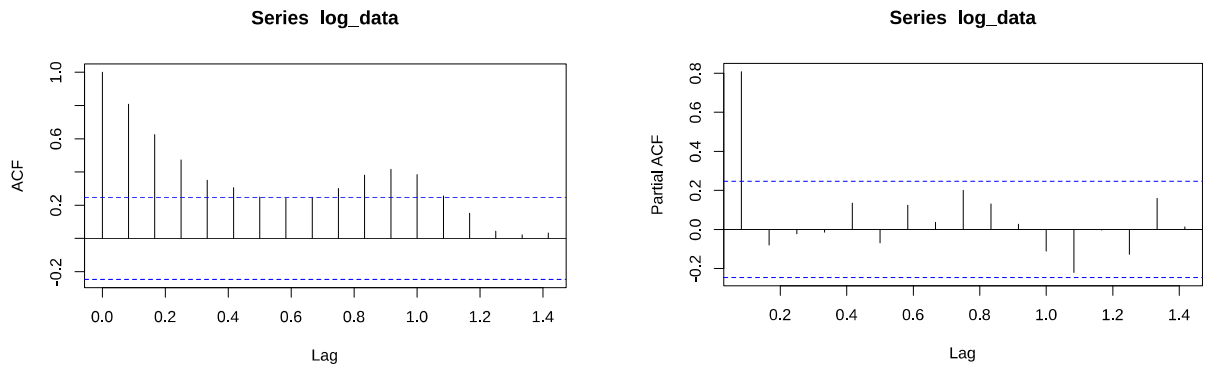

Figura 7. ACF y PACF de la tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes transformada con el logaritmo
Límites en azul definidos por ± 2 / √T, donde T es la longitud de la serie de tiempo

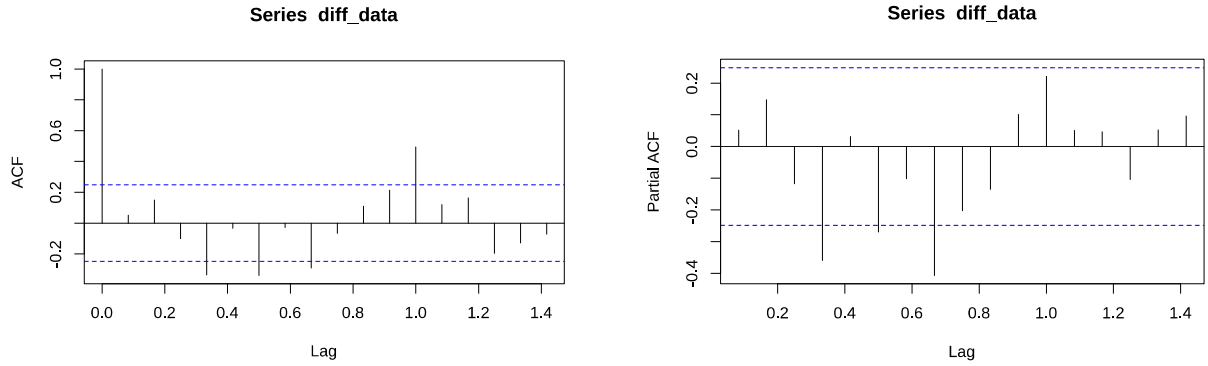

Figura 8. ACF y PACF de la tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes diferenciada
Límites en azul definidos por ± 2 / √T, donde T es la longitud de la serie de tiempo

La serie de tiempo diferenciada es efectivamente un proceso estocástico estacionario y se puede proceder con ella a la modelación.

Modelado de la serie de tiempo

En un modelo de regresión múltiple, pronosticamos la variable de interés utilizando una combinación lineal de predictores. En un modelo de autorregresión, pronosticamos la variable de interés utilizando una combinación lineal de valores pasados de la variable. El término autorregresión indica que es una regresión de la variable contra sí misma. [5]

Así, un modelo de orden autorregresivo de orden p se puede escribir como:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

dónde ε_t es ruido blanco.

Esto funciona como una regresión múltiple pero con valores rezagados de y_t como predictores. Nos referimos a esto como un modelo AR(p), un modelo autorregresivo de orden p. [5]

Los modelos autorregresivos son notablemente flexibles para manejar una amplia gama de patrones de series temporales diferentes. Cambiar los parámetros ϕ_1, ..., ϕ_p da como resultado diferentes patrones de series de tiempo. La varianza del término de error ε_t solo cambiará la escala de la serie, no los patrones. [5]

Para un modelo AR(1): [5]

- cuando $\phi_1 = 0$, y_t es equivalente a ruido blanco;
- cuando $\phi_1 = 1$ y $c = 0$, y_t es equivalente a una caminata aleatoria;
- cuando $\phi_1 = 1$ y $c \neq 0$, y_t es equivalente a una caminata aleatoria con deriva;
- cuando $\phi_1 < 0$, y_t tiende a oscilar alrededor de la media.

Normalmente restringimos los modelos autorregresivos a datos estacionarios, en cuyo caso se requieren algunas restricciones sobre los valores de los parámetros. [5]

- Para un modelo AR(1): $-1 < \phi_1 < 1$.
- Para un modelo AR(2): $-1 < \phi_2 < 1$, $\phi_1 + \phi_2 < 1$, $\phi_2 - \phi_1 < 1$.

Por otro lado, en lugar de usar valores pasados de la variable de pronóstico en una regresión, un modelo de promedio móvil usa errores de pronóstico pasados en un modelo similar a la regresión.

$$y_t = c + \varepsilon_t + \theta 1\varepsilon_{t-1} + \theta 2\varepsilon_{t-2} + \cdots + \theta q\varepsilon_{t-q},$$

donde ε_t es ruido blanco. Nos referimos a esto como un modelo MA(q), un modelo de promedio móvil de orden q. [5]

Cada valor de y_t puede considerarse como un promedio móvil ponderado de los últimos errores de pronóstico. Sin embargo, los modelos de promedio móvil no deben confundirse con el suavizado de promedio móvil. Se usa un modelo de promedio móvil para pronosticar valores futuros, mientras que el suavizado de promedio móvil se usa para estimar el ciclo de tendencia de valores pasados. [5]

Cambiar los parámetros θ_1, …, θ_q da como resultado diferentes patrones de series de tiempo. Al igual que con los modelos autorregresivos, la varianza del término de error ε_t solo cambiará la escala de la serie, no los patrones. [5]

Ahora bien, si combinamos la diferenciación con la autorregresión y un modelo de media móvil, se obitene un modelo ARIMA no estacional. ARIMA es un acrónimo de *AutoRegressive Integrated Moving Average* (en este contexto, "integración" es el reverso de la diferenciación). El modelo completo se puede escribir como:

$$y'_t = c + \phi_1 y'_t - 1 + \cdots + \phi_p y'_t - p + \theta_1 \varepsilon_t - 1 + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

donde y'_t es la serie diferenciada (puede haber sido diferenciado más de una vez). Los "predictores" en el lado derecho incluyen tanto valores retrasados de y_t como errores retrasados. A esto lo llamamos un modelo ARIMA(*p*, *d*, *q*), donde [5]

p = orden de la parte autorregresiva;
d = grado de primera diferenciación involucrado;
q = orden de la parte media móvil. [5]

Las mismas condiciones de estacionariedad e invertibilidad que se utilizan para los modelos autorregresivos y de media móvil también se aplican a un modelo ARIMA. [5]

Ya cumplido el supuesto de estacionariedad en la serie de tiempo diferenciada, se evaluaron múltiples modelos ARIMA mediante el análisis del factor de ajuste de los coeficientes. El factor de ajuste de los modelos se evaluó utilizando la *log likelihood* de los datos y el criterio de información de Akaike (AIC, por sus siglas en inglés).

La estimación de máxima verosimilitud implica encontrar los valores de los parámetros que maximizan la probabilidad de obtener los datos que hemos observado. Para los modelos ARIMA, la estimación de máxima verosimilitud (MLE, por sus siglas en inglés) es similar a las estimaciones de mínimos cuadrados que se obtendrían minimizando: [4]

$$\sum_{t=1}^{T} \varepsilon^2_t.$$

Sin embargo, los modelos ARIMA son mucho más complicados de estimar que los modelos de regresión. Además, es importante tener presente que en caso de que se esté empleando software, los resulados pueden variar ligeramente de otro software ya que utilizan diferentes métodos de estimación y diferentes algoritmos de optimización. En el presente trabajo, se utilizó el entorno y lenguaje de programación R, que informa el valor de la *log likelihood* de los datos; es decir, el logaritmo de la probabilidad de que los datos observados provengan del modelo estimado. Para valores dados de *p*, *d* y *q*, R intentará maximizar la probabilidad logarítmica al encontrar estimaciones de parámetros. [4]

Por otro lado, el criterio de información de Akaike, se puede escribir como [4]

$$AIC = -2\log(L) + 2\,(p + q + k + 1),$$

donde *L* es la probabilidad de los datos, $k = 1$ si $c \neq 0$ y $k = 0$ si $c = 0$. Es importante tener presente que el último término entre paréntesis es el número de parámetros en el modelo (incluido σ^2, la varianza de los residuos). [4]

Para los modelos ARIMA, el AIC corregido se puede escribir como [4]

$$AICc = AIC + 2\,(p + q + k + 1)\,(p + q + k + 2)\,T - p - q - k - 2.$$

Lo que se busca entonces es minimizar el AIC o AICc.

Tomando en cuenta los criterios descritos, los valores de la *log likelihood* y el criterio de información de Akaike obtenidos de los diferentes modelos ARIMA, sugieren que el modelo

ARIMA(0, 0, 1) es el que mejor reproduce el comportamiento de la serie de tiempo diferenciada de la tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes de enero del 2015 a marzo del 2020. Este modelo es de primer orden del componente de promedio móvil. Estos valores se pueden consultar en la Tabla 3 y la visualización del ajuste del modelo se puede observar en la Figura 9.

Modelo	***Log likelihood***	**Criterio de información Akaike**
ARIMA(1, 0, 0)	-53.89	113.78
ARIMA(1, 1, 0)	-70.69	147.39
ARIMA(0, 1, 0)	-85.87	175.74
ARIMA(0, 1, 1)	-52	130
ARIMA(0, 0, 1)	-53.69	113.38
ARIMA(1, 0, 1)	-53.14	114.28
ARIMA(1, 1, 1)	-56.05	120.1

Tabla 3. Valores de la *log likelihood* y el criterio de información de Akaike de distintos modelos ARIMA para la serie de tiempo diferenciada de la tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes diferenciada

	ma1	**sar1**
	-0.5152	-0.5557
s. e.	0.1193	0.1201

Tabla 5. Coeficientes del modelo ARIMA(0, 0, 1) para la serie de tiempo diferenciada de la tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes

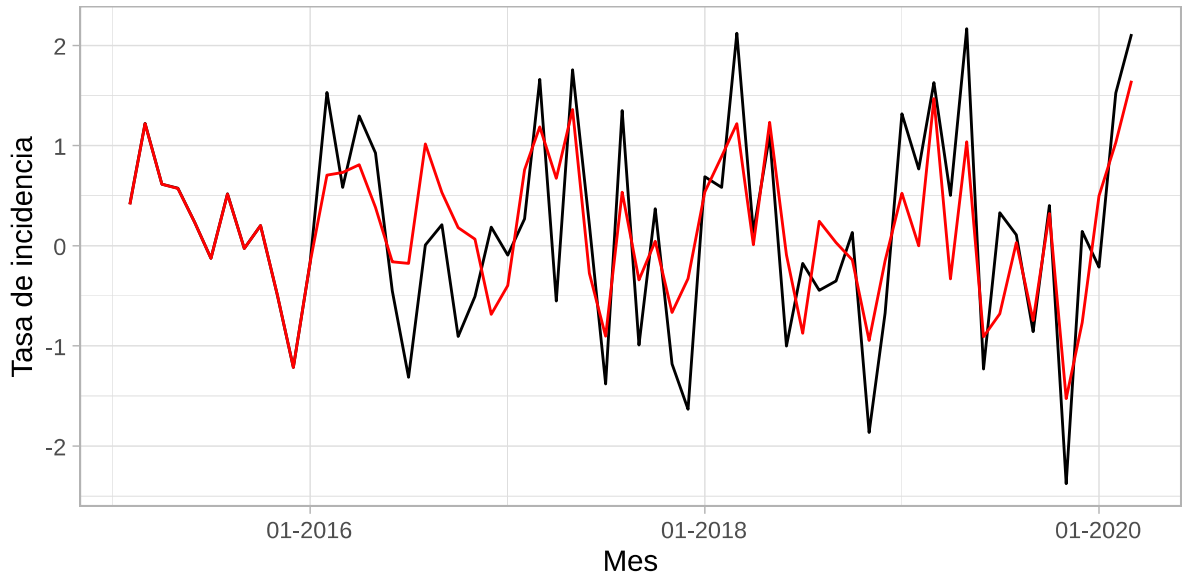

Figura 9. En rojo, ajuste del modelo ARIMA(0, 0, 1) para la serie de tiempo diferenciada de la tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes diferenciada, y en negro, serie de tiempo original diferenciada

Después de que se identificó un modelo óptimo, se evaluaron los residuales de este.

Los residuales en un modelo de serie de tiempo son iguales a la diferencia entre las observaciones y los valores ajustados correspondientes: [4]

$$e_t = y_t - \overline{y}_t.$$

En la visualización de la ACF de la Figura 9, se puede observar que los residuos no están correlacionados. Esto es relevante pues correlación entre los residuales indica que aún hay información en los residuales que el modelo podría aún estar capturando. Así mismo, en el histograma de la Figura 9, se puede observar que su distribución es normal, su varianza es constante y su media es 0. De esta forma, se puede concluir que el modelo producirá pronósticos no sesgados. [4]

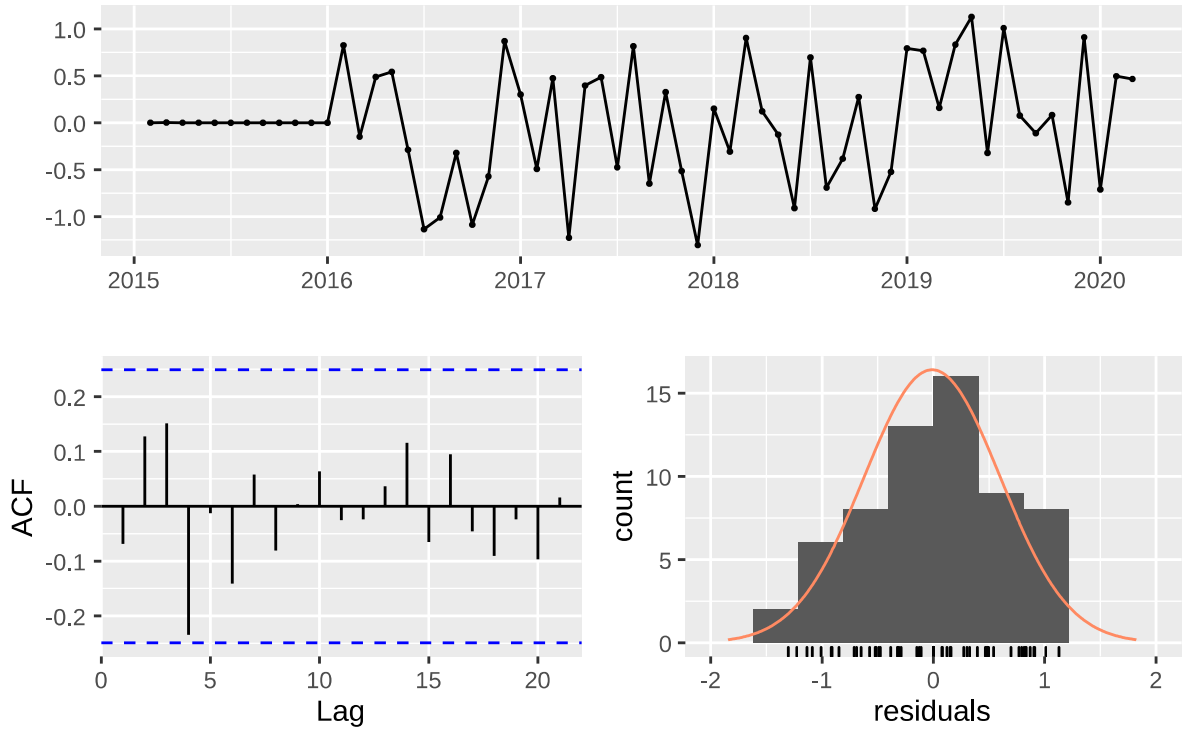

Figura 10. En la visualización superior, la serie de tiempo de los residuales del modelo ARIMA(0, 0, 1). En la visualización inferior izquierda, la ACF de los residuales del modelo ARIMA(0, 0, 1), en donde los ímites en azul definidos por ± 2 / √T, donde T es la longitud de la serie de tiempo.En la visualización inferior derecha, la disttribución de los residuales del modelo ARIMA(0, 0, 1).

Adicionalmente, utilizando el entorno y lenguaje de programación R, se realizaron tres pruebas de hipótesis estadísticas para evaluar la normalidad de los residuales. Las tres pruebas de normalidad empleadas en esta evaluación fueron las siguientes:

1. La prueba de Ljung-Box
2. La prueba de Ljung-Box
3. La prueba de Shapiro-Wilk

La hipótesis nula de las tres pruebas señala que los datos se distribuyen normalmente. Por consiguiente, se debe buscar evidencia de que la hipótesis nula se acepte, es decir, se buscan valores *p* mayores de 0.05. [4]

Los valores p de las tres pruebas de normalidad aplicadas a los residuales del modelo ARIMA(0, 0, 1) sugieren que estos tienen una distribución normal, lo cual refuerza la validación de este modelo como uno óptimo para la reproducción de la serie de tiempo de la variable estimada.

Prueba de hipótesis estadística	Valor p
Ljung-Box Ljung-Box	0.5804
Ljung-Box	0.4613
Shapiro-Wilk	0.1737

Tabla 5. Valores p obtenidos al aplicar las pruebas de hipótesis estadísticas en normarlidad a los residuales del modelo ARIMA(0, 0, 1) para la serie de tiempo diferenciada de la tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes

Prónostico de la serie de tiempo

Finalmente, con los coeficientes del modelo ARIMA(0, 0, 1), se procedió a realizar un pronóstico hasta octubre del 2020, mismos con los que se contaba con información reportada por el Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública. Los coeficientes del modelo se pueden consultar en la Tabla 4.

En la Figura 11, se puede observar la serie de tiempo diferenciada con el próstico para los siguientes siete meses.

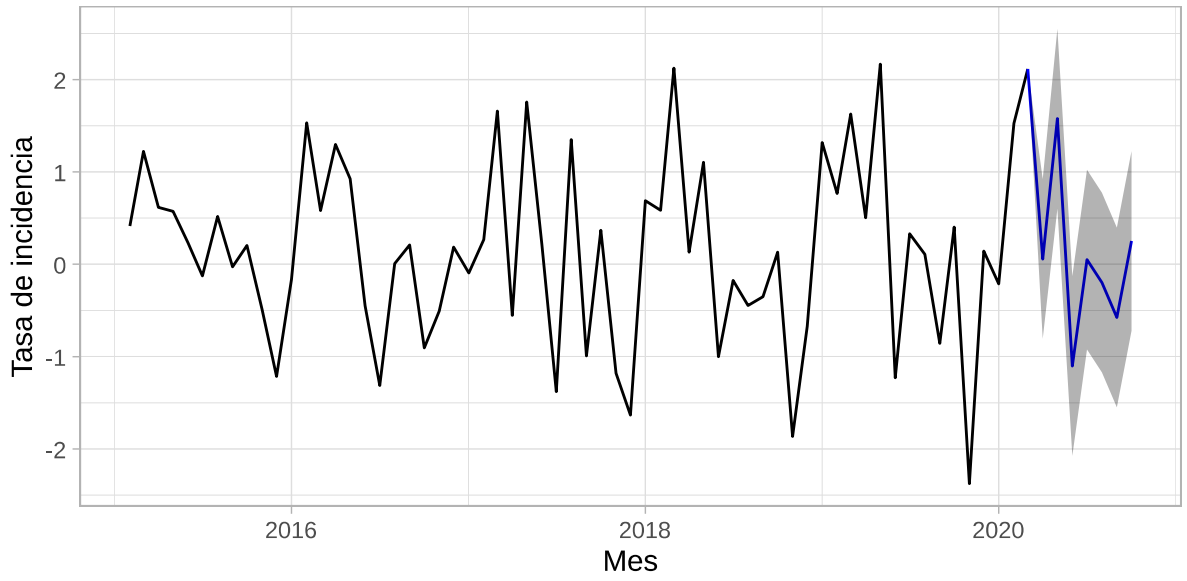

Figura 11. En azul, el prónostico para los siguientes siete meses del modelo ARIMA(0, 0, 1) para la serie de tiempo diferenciada de la tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes; en gris, los intervalos de confianza del pronóstico; y en negro, serie de tiempo original diferenciada

Si bien el pronóstico es correcto, este debe ser transformado dado que la serie de tiempo utilizada para el modelo fue diferenciada. En la Figura 12, se puede visualizar la serie de tiempo original de enero del 2020 a octubre del 2020 junto con el pronóstico transformado de marzo del 2020 a octubre del 2020.

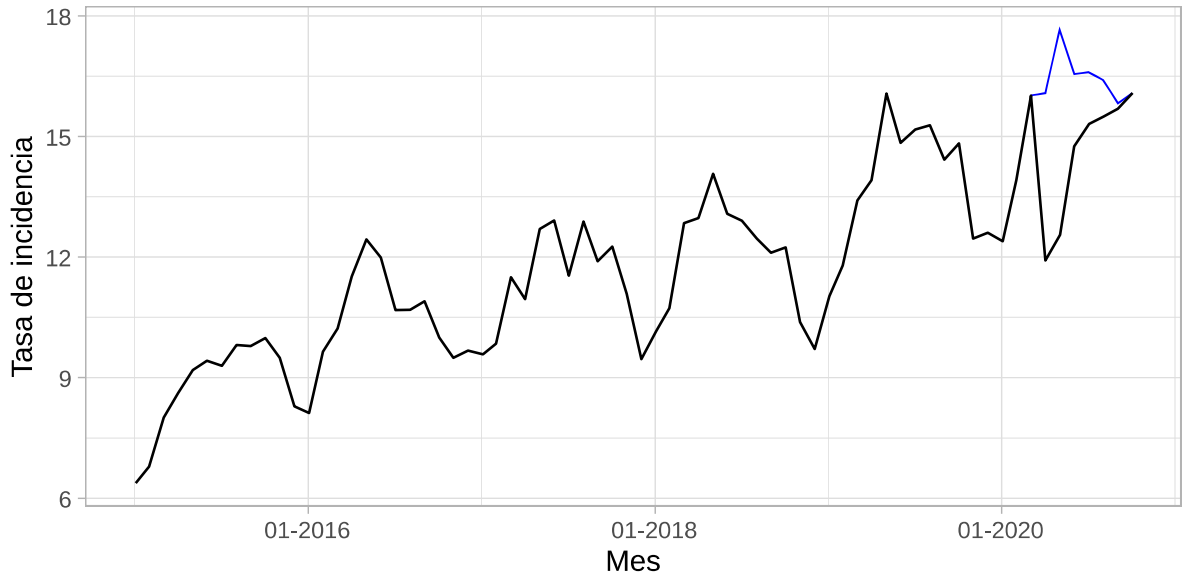

Figura 12. En azul, el prónostico para los siguientes siete meses del modelo ARIMA(0, 0, 1) para la serie de tiempo de la tasa de incidencia de violencia familiar mensual en México por cada 1,000 habitantes; en negro, la serie de tiempo original

Resultados

Como se puede observar enl la Tabla 7, de abril a octubre del 2020, el valor pronosticado supera al valor original de la serie de tiempo, y es hasta el mes de octubre cuando el valor original de la serie de tiempo coincide por menos de 0.1 puntos porcentuales con el valor pronosticado.

Ahora bien, para obtener la maginitud del subregistro de casos de violencia familiar en México con base en los pronosticos del modelado desarollado, se calculó el número de casos a partir de la tasa de incidencia con la siguiente fórmula:

Casos de violencia familiar = (Tasa de incidencia/100)(Proyección de población anual/1000)*

El subregistro mensual de casos de violencia familiar se puede consultar en la Tabla 8. El subregstro total fue de 17,110.46.

Mes	Valor original de la serie de tiempo (%)	Valor pronosticado (%)	Valor original de la serie de tiempo - Valor pronosticado (%)
Abril 2020	11.923255	16.101520	-4.178265
Mayo 2020	12.568834	17.678145	-5.109311
Junio 2020	14.775540	16.575592	-1.800052
Julio 2020	15.339737	16.624698	-1.284961
Agosto 2020	15.527541	16.425748	-0.898207
Septiembre 2020	15.719259	15.848989	-0.12973
Octubre 2020	16.112084	16.100827	0.011257

Tabla 7. Valores originales de la serie de tiempo, valores pronosticados y diferencia entre los valores originales de la serie de tiempo y los valores pronosticados por mes

Mes	Casos de violencia familiar registrados por el SESNSP	Casos de violencia familiar pronosticados	Casos de violencia familiar registrados por el SESNSP – Casos de violencia familiar pronosticados
Abril 2020	15,237	20,576.5	-5339.5
Mayo 2020	16,062	22,591.31	-6529.31
Junio 2020	18,882	21,182.33	-2300.33
Julio 2020	19,603	21,245.08	-1642.08
Agosto 2020	19,843	20,990.84	-1147.84
Septiembre 2020	20,088	20,253.79	-165.79
Octubre 2020	20,590	20,575.61	14.39

Tabla 8. Casos de violencia familiar registrados por el SESNSP, casos de violencia familiar pronosticados y diferencia entre los casos de violencia familiar registrados por el SESNSP y los casos de violencia familiar pronosticados

Conclusiones

A pesar de que los modelos autorregresivos tipo ARIMA fueron diseñados para el estudio de fenómenos económicos, fue posible la modelación de la tasa de incidencia de violencia familiar en México por cada 1,000 habitantes con uno de tipo ARIMA(0, 0, 1). No obstante, para conseguirlo se la serie de tiempo se diferenció y se validó su estacionariedad. Así mismo, se validó que los residuales del modelo cumplieran con el supuesto de normalidad.

El modelo permitió pronosticar la tasa de incidencia de violencia familiar en México de abril a octubre del 2020, meses en los cuales, dados los efectos intrínsecos de las medidas de confinamiento del Gobierno de México, se presentó un subregistro en los casos reportados por la SESNSP de violencia familiar en su reporte estatal mensual. Se puede concluir

entonces que los modelos autorregresivos tipo ARIMA pueden resultar como una herramienta apropiada para la estimación de datos de la índole en escenarios de subregistro.

Ya sea que se presente un subregistro evidente o no, tener pronósticos de tasas de incidencia delictivas resulta esencial para la validación de indicadores tan importantes como el estudiado en el presente trabajo. En este caso, tener un pronóstico de la violencia familiar durante la pandemia por covid-19 en México permite conocer el crecimiento base de este fenómeno y permite tener una estimación del número mínimo de víctimas. De esta forma, los planes de respuesta y de recuperación del covid-19 nacionales y subnacionales pueden reforzarse a favor de las víctimas con mayor riesgo de sufrir violencia familiar: los niños, las niñas y las mujeres.

Recordemos siempre que las decisiones que se basan en datos precisos y que incluyen una perspectiva de género tienen más probabilidades de ser eficaces.

Referencias

Falta ponerlas en formato correcto de bibliografía.

1. Subdirección de Informática Jurídica, Dirección General de Tecnologías de Información y Comunicaciones, Comisión Nacional de los Derechos Humanos. (2009) *Norma Oficial Mexicana NOM-046-SSA2-2005. Violencia familiar, sexual y contra las mujeres. Criterios para la prevención y atención.* https://www.cndh.org.mx/DocTR/2016/JUR/A70/01/JUR-20170331-NOR19.pdf Accesado el 9 de diciembre del 2020.
2. UN WOMEN. (2020) *COVID-19: Emerging gender data and why it matters* https://data.unwomen.org/resources/covid-19-emerging-gender-data-and-why-it-matters#vaw Accesado el 9 de diciembre del 2020.
3. Intersecta, EQUIS Justicia para las mujeres, La red Nacional de Refugios, A.C. (2020) *Las dos pandemias: Violencia contra las muejres en México en el contextod e COVID-19*. https://equis.org.mx/wp-content/uploads/2020/08/informe-dospandemiasmexico.pdf
4. Hyndman, R.J., & Athanasopoulos, G. (2019) *Forecasting: principles and practice*, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3 Accesado el 9 de diciembre del 2020.
5. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed). Hoboken, New Jersey: John Wiley & Sons.

Notas adicionales

Todo el código en R empleado para el preprocesamiento, análisis, modelación y visualización de datos en el presente trabajo, así como las bases de datos utiizadas y construídas, y las visualizaciones en alta definición, se enceuntran alojadas en el siguiente repositorio: https://github.com/sarahiaguilar/violencia-intrafamiliar