Modelos de Series de Tiempo para Predecir el Número de Casos de Variantes Dominantes del SARS-COV-2 Durante las Olas Epidémicas en Chile

Barría-Sandoval, Claudia [1,2]; **Salas, Patricio** [3]; **Ferreira, Guillermo** [3,*]

[1] *Universidad de las Américas, Escuela de Enfermería, Concepción, Chile*
[2] *Universidad de Concepción, Facultad de Enfermería, Concepción, Chile*
[3] *Universidad de Concepción, Departamento de Estadística, Concepción, Chile*

Resumen: El COVID-19 y sus variantes han creado una pandemia a nivel global. En Chile, hasta el 28 de febrero del 2022, ya se han infectado más de 3 millones de personas y han muerto más de 42 mil personas. En este artículo, se realiza un estudio comparativo de diferentes modelos matemáticos utilizados para modelar y predecir el número de casos diarios confirmados de COVID-19 en Chile. Esta investigación considera los registros diarios de casos confirmados desde el inicio de la pandemia y por lo tanto incluye los contagiados por las distintas variantes del virus (Delta, Gamma y Omicron), estas variantes han dominado la evolución de los contagios diarios en Chile, siendo la variante Omicron la que ha demostrado tener una mayor tasa de contagios a nivel nacional. El objetivo de este estudio es brindar información relevante sobre la evolución de la pandemia por COVID-19 en Chile mediante modelos de series de tiempo que han sido validados en distintas investigaciones y evaluar su precisión frente a la variante Omicron del virus SARS-CoV-2.

Palabras clave: COVID-19; SARS-COV-2; Delta; Gamma; Omicron; Modelo de Series de Tiempo

Time Series Models for Forecasting the Number of Cases of SARS-COV-2 Dominant Variants During the Epidemic Waves in Chile

Abstract: COVID-19 and its variants have created a global pandemic. In Chile, as of February 28 2022, more than 3 million people have been infected and more than 42 thousand people have died. In this article, a comparative study of different mathematical models used to model and predict the number of daily confirmed cases of COVID-19 in Chile is carried out. This research considers the daily records of confirmed cases since the beginning of the pandemic and therefore, includes those infected by the different variants of the virus (Delta, Gamma and Omicron), these variants have dominated the evolution of daily infections in Chile, being the Omicron variant the one that has shown to have a higher rate of infection at national level. The objective of this study is to provide relevant information on the evolution of the COVID-19 pandemic in Chile through time series models that have been validated in different investigations and to assess their validity with the appearance of the Omicron variant of the SARS-CoV-2 virus.

Keywords: COVID-19; SARS-COV-2; Delta; Gamma; Omicron; Time Series Models

1. Introducción

En la ciudad China de Wuhan durante el mes de diciembre del año 2019, comenzó la propagación de un nuevo virus SARS-CoV-2 que causa la enfermedad infectocontagiosa por coronavirus, y desde ese momento se expandió rápidamente por el mundo. El 11 de marzo del año 2020 la OMS, Organización Mundial de la Salud (2020) declaró al COVID-19 como pandemia. El COVID-19 ha tenido un fuerte impacto en la salud a nivel mundial. Dado que este virus ha infectado un gran número de personas causando complicaciones severas, secuelas a largo plazo, defunciones e incremento de mortalidad tanto en Chile como en todo el mundo. En este sentido, los gobiernos y los sistemas de salud han enfocado todo su trabajo y recursos financieros en controlar la propagación de la pandemia por COVID-19. Lo anterior refleja la necesidad de estudiar el comportamiento de la velocidad de propagación y de esta manera brindar información oportuna para tomar decisiones informadas y aplicar las medidas de control pertinentes.

Desde el comienzo de la pandemia, se han propuesto diferentes

*gferreir@udec.cl
Recibido: 10/03/2022
Aceptado: 01/07/2022
Publicado en línea: 23/12/2022
10.33333/rp.vol50n3.02

metodologías para modelar el comportamiento del COVID-19. Entre los modelos más utilizados se encuentran el enfoque econométrico basado en modelos de series de tiempo, los modelos Susceptible-Exposed-Infectious-Removed (SEIR) y el enfoque de aprendizaje automático. Diversos estudios han propuesto modelos del tipo Autoregresivo Integrado de Media Moviles (ARIMA) para predecir el comportamiento del número de contagios por COVID-19. Por ejemplo, Ibrahim et al. (2020) propusieron un modelo del tipo ARIMA de orden $(1,1,0)$ para modelar y realizar predicciones de la propagación del COVID-19 en Nigeria. Talkhi et al. (2021) han realizado un estudio comparativo de diferentes técnicas de series de tiempo y concluyeron que el modelo más adecuado para los datos de casos confirmados por COVID-19 en Irán, fue el Multilayer Perceptron (MLP) y el modelo Holt-Winter para predecir los casos de muerte por COVID-19. En la misma línea, Yonar et al. (2020) propusieron un estudio para predecir y modelar el número de casos de COVID-19 utilizando dos metodologías: ARIMA y Métodos de Suavizado Exponencial. En este trabajo, los autores mencionan que, para los diferentes países en estudio no existe un único modelo para describir el comportamiento del número de casos, pero según las características de los datos, ambos métodos son efectivos para describir las curvas de propagación del virus. Como una alternativa a los métodos econométricos tradicionales están los métodos provenientes del campo de aprendizaje automático (AA). Ghafouri-Fard et al. (2021) realizan una exhaustiva revisión de trabajos de AA que se enfocan en la predicción de la tendencia de propagación del COVID-19. En esta revisión, destacan el beneficio de usar el modelo *Long short-term memory* (LSTM) propuesto por Hochreiter et al. (1997), el cual pertenece a la familia de las redes neuronales recurrentes. Finalmente, Roda et al. (2020) señalaron que las predicciones de la pandemia de COVID-19 que utilizan modelos más complejos pueden no ser más confiables que el uso de un modelo más simple. Se remite al lector a los siguientes autores y sus referencias: Perone (2020), Sarkar (2020), Tran et al. (2020) para complementar la revisión de otros modelos utilizados en las predicciones de COVID-19.

Existen diferentes estudios que han analizado la propagación del COVID-19 en Chile. Por ejemplo, Vicuña et al. (2020) han especificado un modelo de crecimiento logístico generalizado multiparámetro. Entre sus conclusiones mencionan que la política de confinamiento implementada en la mayor parte del país ha demostrado ser eficaz para detener la propagación, y las políticas de confinamiento-relajación, aunque graduales, parecen haber provocado un quiebre al alza en la tendencia. Tariq et al. (2021) han empleado ecuaciones de renovación para estimar el número de reproducción (R) para la fase ascendente temprana de la epidemia de COVID-19 y sus resultados indican una transmisión sostenida temprana de SARS-CoV-2. Por otro lado, Freire-Flores et al. (2021) han utilizado un modelo SEIRD multigrupo para estudiar la propagación de COVID-19 en las diferentes regiones de Chile. Barría-Sandoval et al. (2021) proponen un estudio de diferentes técnicas de series de tiempo para predecir tanto el número de casos confirmados como el número de muertes por COVID-19 en Chile. Barría-Sandoval et al. (2022) proponen un estudio de diferentes modelos de datos de panel para analizar el efecto de la pandemia del COVID-19 sobre las muertes por enfermedades respiratorias en las regiones de Chile.

A diferencia de los estudios antes mencionados, este artículo considera las diferentes variantes del virus que han sido dominantes en Chile, las cuales han sido declaradas como variantes de preocupación (VOC, siglas en inglés de Variant of Concern) por la Organización Mundial de la Salud (2021). De acuerdo a esta clasificación y a los informes del Instituto de Salud Pública (2020) de Chile, las variantes dominantes que han circulado y causado peak de la pandemia en Chile son: "Delta", "Gamma" y "Omicron". A la fecha de la realización de este estudio, la variante Omicron es la dominante y desde su detección el 04 de diciembre de 2021 el número de contagios ha aumentado exponencialmente llegando a cifras de contagios históricas en lo que lleva la pandemia a nivel nacional, pasando de 2060 casos a 22845 casos el 15 de febrero de 2022 (Secretaría de Comunicaciones-Gobierno de Chile, 2021); lo que representa un incremento del 1009 % de los casos confirmados de COVID-19 en los últimos dos meses. De lo anterior, surge el cuestionamiento sobre la efectividad de los modelos de series de tiempo estudiados hasta este momento para predecir la propagación de la pandemia por COVID-19: ¿Son efectivos estos modelos ante el quiebre de tendencia provocado por la variante Omicron?. Esta es la interrogante que esta investigación busca responder.

2. Materiales y Métodos

2.1 Conjunto de Datos y Definición de Variable

En este estudio, se analizó la cantidad de casos diarios confirmados por COVID-19 en Chile desde el 2 de marzo de 2020 al 15 de febrero de 2022. Los datos fueron obtenidos del Ministerio de Ciencia y Tecnología, Conocimiento e Innovación disponibles en el sitio web http://www.minciencia.gob.cl/covid19. La Figura 1 señala el número de casos confirmados de COVID-19 en Chile durante el inicio de la pandemia hasta la fecha de término de este estudio. En esta figura, se observan tres olas de contagios que han producido 3 peaks, los cuales están resaltados por las bandas de color gris claro. La primera ola tiene un peak el 15 de junio del 2020 (bloque gris (a)) con variante dominante Delta. Luego la segunda ola ocurre entre los meses de abril y junio del 2021 con dos peak, uno el 09 de abril y el otro el 05 de junio respectivamente (bloque gris (b)), en este caso la variante dominante fue Gamma. Finalmente, el bloque gris (c) representa el aumento significativo de la Variante Omicron, la cual se inició a principios de enero del 2022.

2.2 Modelos Empíricos

Se realizó una revisión de las metodologías estadísticas utilizadas para modelar los datos registrados sobre el número de contagios diarios confirmados por COVID-19 en Chile a lo largo del tiempo. Y entre ellas se proponen diferentes modelos que permiten capturar la dependencia temporal entre observaciones, incluyendo un modelo proveniente del campo de Machine Learning. A continuación, se describen cada uno de estos modelos:

- **Modelo 1**: ARIMA(p,d,q)

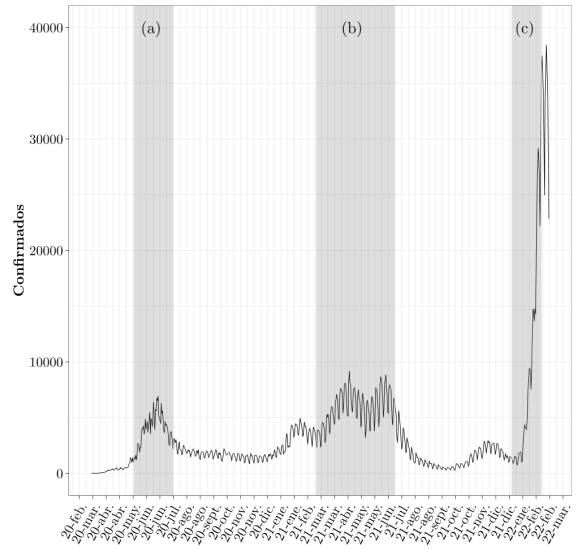

Figura 1. *COVID-19 en Chile:* Casos Confirmados del 03-marzo-2020 al 15-febrero-2022. Paneles (a) Variante Delta; (b) Variante Gamma; (c) Variante Omicron

El modelo autoregresivo integrado de medias móvil (ARIMA) es un modelo del tipo Box-Jenkins. Este tipo de modelo ha sido utilizado en diferentes áreas de la ciencia para estudiar el comportamiento de datos registrados secuencialmente en un periodo de tiempo y fue propuesto por Box et al. (2015) en su libro seminal de 1970 "Time Series Analysis: Forecasting and Control".

Definition 2.1 *Si d es un entero no negativos, entonces* $\{X_t\}$ *es un proceso ARIMA*(p,d,q) *si la serie diferenciada* $(1-B)^d X_t$ *es un proceso causal ARMA. Este proceso puede ser escrito como:*

$$\phi(B)\,(1-B)^d X_t = \theta(B)\,\varepsilon_t \text{ con, } \{\varepsilon_t\} \sim RB(0,\sigma^2)$$

donde $\phi(z) = 1-\phi_1 z-\cdots-\phi_p z^p$, *y* $\theta(z) = 1+\theta_1 z+\cdots+\theta_p z^q$ *corresponden a los polinomios autorregresivo y de medias móviles respectivamente.*

El proceso denotado por *RB* es un proceso de ruido blanco no-correlacionado con media cero y varianza σ^2. El lector puede encontrar una fácil descripción sobre modelos de datos de series temporales en Brockwell et al. (2016).

- **Modelo 2**: Método de Holt-Winters con tendencia amortiguada

Otro modelo para trabajar con datos de series de tiempo es el método de suavizamiento exponencial. Los autores Holt (2004) y Winters (1960) han propuesto una clase más general de métodos para capturar tanto la tendencia como el nivel de la serie de tiempo, la cual es conocida como método de Holt-Winters. Uno de los modelos más simples de este tipo es el suavizamiento exponencial definido por:

$$\begin{aligned} \widehat{X}_{t+h|t} &= \mu_t + hT_t, \\ \mu_t &= \alpha X_t + (1-\alpha)(\mu_{t-1}+T_{t-1}), \qquad \text{Model 2a} \\ T_t &= \beta(\mu_t-\mu_{t-1}) + (1-\beta)T_{t-1}, \end{aligned}$$

donde $0 \le \alpha \le 1$ es el parámetro de suavizamiento y $0 \le \beta \le 1$ parámetro de suavizamiento para la tendencia. Por otro lado, Gardner et al. (1985) desarrollaron un modelo de suavizado exponencial diseñado para amortiguar las tendencias erráticas definido de la siguiente manera:

$$\begin{aligned} \widehat{X}_{t+h|t} &= \mu_t + (\lambda+\lambda^2+\ldots+\lambda^{h-1})T_t, \\ \mu_t &= \alpha X_t + (1-\alpha)(\mu_{t-1}+\lambda T_{t-1}), \qquad \text{Model 2b} \\ T_t &= \beta(\mu_t-\mu_{t-1}) + (1-\beta)\lambda T_{t-1}, \end{aligned}$$

donde $0 < \lambda < 1$ es el parámetro de amortiguación. Los pronósticos de h paso $\widehat{X}_{t+h|t}$ se calculan usando las ecuaciones de suavizado para el nivel μ_t y la tendencia T_t.

- **Modelo 3**: Red Neuronal Recurrente

Una red neuronal recurrente (RNN) es una familia de redes neuronales que permiten procesar datos secuenciales $(x_1,\ldots,x_T)$. Los detalles de la arquitectura de las RNN se pueden encontrar en Goodfellow et al. (2016) y Aggarwal (2018). Una RNN consiste en estados ocultos que se distribuyen de forma temporal y son capaces de predecir los eventos futuros con más precisión que los métodos tradicionales de suavización exponencial Shastri et al. (2020). Los estados ocultos de una RNN son definidos de la siguiente manera:

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{x}_t; \theta),$$

donde $f(\cdot)$ es una función de activación, x_t son los input y θ son los pesos de la red. En este estudio, se utilizó el modelo Long Short Term Memory (LSTM) desarrollado por Hochreiter et al. (1997), el cual es un tipo especial de RNN capaz de aprender de rezagos o dependencias temporales a largo plazo en los datos. A diferencia de las redes neuronales estándar, este modelo garantiza la conexión de retroalimentación como propone Ghafouri-Fard et al. (2021). Además de procesar puntos de datos individuales, el LSTM puede procesar secuencias completas de datos como lo señala Hochreiter et al. (1997). En la Figura 2, se muestra la arquitectura general de un modelo LSTM.

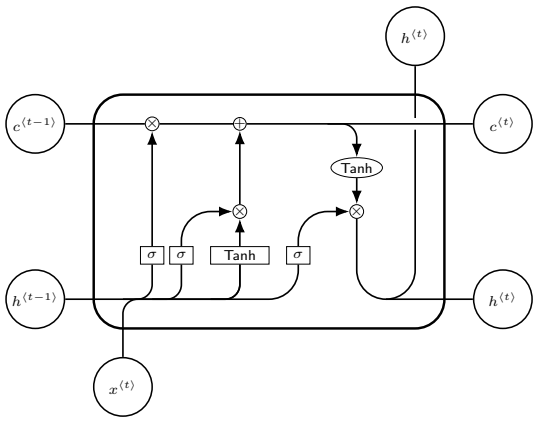

Figura 2. LSTM: Arquitectura del modelo LSTM

El papel central de un modelo LSTM lo desempeña una célula de memoria conocida como "estado de célula" ($c^{\langle t \rangle}$) que mantiene su estado a lo largo del tiempo. El estado de la célula es la línea horizontal que atraviesa la parte superior de la Figura 2. Puede visualizarse como una cinta transportadora a través de la cual la información simplemente fluye, sin cambios.

La información puede añadirse o eliminarse del estado de la célula en la LSTM y se regula mediante puertas □. Estas puertas permiten opcionalmente que la información fluya dentro y fuera de la célula. Por ejemplo, la información se incorpora si la función sigmoide $\sigma(\cdot)$ le asigna un valor 1, se incorpora parcialmente si la $\sigma(\cdot)$ entrega un valor entre 0 y 1 y se descarta si $\sigma(\cdot)$ entrega un valor 0. Luego se procede a actualizar el estado de la célula multiplicando la salida de la puerta $\boxed{\sigma}$ por el estado anterior ($c^{\langle t-1 \rangle}$) mediante la operación de multiplicación puntual ($\otimes$). El proceso continua para actualizar el estado oculto ($h^{\langle t \rangle}$), multiplicando los resultados de la puerta de salida por el resultado que entrega la función de activación $\boxed{\text{Tanh}}$ que tiene como argumento de entrada el estado de la célula actualizado $c^{\langle t \rangle}$. Finalmente, el último estado de celda ($c^{\langle t \rangle}$) y el estado oculto ($h^{\langle t \rangle}$) regresan a la unidad recurrente y el proceso se repite en el paso de tiempo $t+1$. El ciclo continúa hasta que se llega al final de la secuencia.

- **Modelo 4**: Modelo Híbrido
  Los modelos híbridos permiten interactuar con diferentes métodos de predicción mediante una relación lineal ponderada. En particular, el paquete `forecastHybrid` del software libre `R` Team (2013) proporciona predicciones a h pasos ponderando las predicciones de m modelos individuales como sigue:

$$f(i) = \sum_{m=1}^{n} \omega_m f_m(i), \text{ con } 1 \leq i \leq h,$$

  donde $f(i)$ representa la i-ésima predicción total, ω_m son los pesos y $f_m(i)$ son las predicciones de m componentes individuales de los siguientes modelos:

  - **ARIMA**
  - Modelos de suavizamiento exponencial **ETS**, los cuales ajustan datos con una componente de tendencia (T), componente estacional (S) y un término de error (E).
  - El **modelo Theta** propuesto por Assimakopoulos et al. (2000) permite obtener una línea Theta denotada por $Z(\theta)$ la cual se logra con la solución de la ecuación:

$$\nabla^2 Z_t(\theta) = \theta \nabla^2 X_t, \tag{1}$$

    donde $\{X_t\}$ son las observaciones y ∇ es el operador de diferencias (es decir, $\nabla Y_t = Y_t - Y_{t-1}$). Una solución analítica de (1) es dada por:

$$Z_t(\theta) = \theta X_t + (1-\theta)(a+bt),$$

    donde a y b son constantes obtenidas al minimizar $\sum_{t=1}^{n} [X_t - Z_t(\theta)]^2$.

  - Red Neuronal con valores rezagados de la serie de tiempo como entradas y posiblemente algunas otras entradas exógenas, denotado por **NNETAR** en la autoría de Hyndman et al. (2002).
  - Otros modelos tales como: **STL** (Seasonal Decomposition of Time Series by Loess), **TBATS** (Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components) y **SNAIVE**, se pueden revisar en detalle en De Livera et al. (2012).

  Finalmente, el paquete `forecastHybrid` permite utilizar todos los modelos antes mencionados a partir de los comandos `auto.arima()`, `ets()`, `them()`, `nnetar()`, `stlm()`, `tbats()` y `snaive()`; los cuales se pueden combinar con pesos iguales, pesos que utilizan errores en la muestra, o determinados por validación cruzada según lo mencionado por Shaub et al. (2020). Estas formas de ponderación originan tres modelos diferentes denotados en este estudio como Modelo 4a, Modelo 4b y Modelo 4c, respectivamente.

3. RESULTADOS

Se analizó el desempeño de los Modelos $1-4$ descritos en la Sección 2. Los datos analizados fueron la cantidad de casos con un diagnóstico confirmado de COVID-19 en Chile desde 2 de marzo de 2020 al 15 de febrero de 2022 (ver Figura 1). Desde esta figura, se puede apreciar un aumento significativo de la varianza al final del periodo provocado por la variante Omicron del COVID-19. Para estabilizar la varianza se realizó una transformación logarítmica a las observaciones, es decir, $Y_t = \log X_t$. Tanto las estimaciones como las predicciones fueron obtenidas a partir de la serie $\{Y_t\}$ para luego transformar a su escala original por la utilización de la función inversa, es decir $\widehat{X}_t = \exp(\widehat{Y}_t)$. Esta transformación ha sido aplicada por Feroze (2021) para evaluar la progresión del COVID-19 en Irán. A continuación, se describen los tipos de modelos utilizados para los datos, los comandos y paquetes del software libre `R` para la estimación de parámetros.

Para el Modelo 1, se aplicó un modelo ARIMA$(3,1,3)$ con media constante. Tales estimaciones se obtuvieron usando el comando `Arima` del paquete `forecast`. En el caso de los Modelos 2a, las estimaciones fueron obtenidas a través del comando `holt` del paquete `forecast` y para el Modelo 2b se utilizó el mismo comando y agregó el argumento `damped=TRUE`.

El Modelo 3 fue construido en Google Colaboratory utilizando `Python 3.0` con librerías de código abierto como: `Tensorflow` propuesto por Abadi et al. (2016), `Pandas` elaborada por McKinney (2010), `Numpy` desarrollada por Oliphant (2006) y `Keras` construida por Chollet (2018). La arquitectura considerada para el Modelo 3 fue simple; donde la primera capa de entrada de la red estuvo compuesta por 15 neuronas con función de activación tangente hiperbólica. Posteriormente, la capa de salida estuvo conformada por 1 neurona. Para el entrenamiento de este modelo se utilizó el optimizador ADAM y el cuadrado medio del error como función de pérdida para evaluar

el desempeño del mismo.

Para el Modelo 4, se utilizó la función `hybridModel` del paquete `forecastHybrid`, el argumento `weights` que permitió seleccionar las ponderaciones de los métodos que interactúan en los modelos híbridos. En particular, las ponderaciones del tipo `equal, insample.errors, cv.errors` fueron implementadas para generar las predicciones de los Modelos 4a, 4b y 4c respectivamente.

La Tabla 1 muestra las estimaciones de los parámetros y la desviación estándard (d.e.) estimada de los Modelos $1-2$. Para probar la importancia de las estimaciones de los parámetros se aplicó la estadística $t = \widehat{\theta}/de(\widehat{\theta})$ al Modelo 1 para el conjunto de datos. En la última columna de esta tabla, se puede observar que las estadísticas t son altamente significativas al 5% de nivel de confianza. Por otra parte, la Tabla 2 reporta las ponderaciones de los Modelos 4a, 4b y 4c. La Figura 3 muestra los valores ajustados para cada modelo, donde las líneas discontinuas representan los datos reales, mientras que las líneas continuas representan los valores ajustados y se despliegan en las curvas de colores para los modelos $1-4$. A partir de esta figura, se puede concluir que el mejor método que captura la tendencia y la estructura de dependencia temporal son los modelos del grupo 4.

Tabla 1. Casos Confirmados de la serie COVID-19: Parámetros estimados de los Modelos $1-3$

Modelo	Parámetros	Estimaciones	d.e.	$\lvert t\rvert$
Modelo 1	ϕ_1	$-0,23$	$0,07$	$3,25$
	ϕ_2	$0,19$	$0,06$	$3,22$
	ϕ_3	$-0,29$	$0,05$	$5,67$
	θ_1	$-0,92$	$0,06$	$15,65$
	θ_2	$-0,77$	$0,09$	$8,38$
	θ_3	$0,79$	$0,05$	$16,68$
	σ^2	$0,05$	–	–
Modelo 2a	α	$0,99$	–	–
	β	10^{-4}	–	–
Modelo 2b	α	$0,99$	–	–
	β	10^{-4}	–	–
	λ	$0,98$	–	–

Tabla 2. Casos Confirmados de la serie COVID-19: Ponderaciones del Modelo 4

	ARIMA	ETS	Tthetam	NNETAR	TBATS	SNAIVE
				Pesos		
4a	0,17	0,17	0,17	0,17	0,17	0,17
4b	0,13	0,11	0,11	0,43	0,12	0,11
4c	0,19	0,16	0,16	0,14	0,19	0,16

Para la identificación del mejor modelo, en la siguiente sección se propone un análisis de precisión para predicciones realizadas con datos de entrenamiento y datos de prueba.

3.1 Análisis de la precisión del pronóstico ex post

Se evaluó la precisión de las predicciones usando un conjunto de entrenamiento y un conjunto de prueba. Se consideró un conjunto de entrenamiento $\{X_t\}$ desde 02 de marzo 2020 al 10 de febrero 2022 (período de estimación) con un total de $N = 709$ observaciones y datos de prueba $\{X_t\}$ del 11 al 15 de febrero 2022 (período validación) que se utilizó para la predicción paso a paso de m con $m = 5$ observaciones diarias.

Los valores $\widehat{X}_{N+1},\ldots,\widehat{X}_{N+m}$ se denominan pronóstico ex post o pronóstico del período. Los pronósticos m-step-ahead se compararon con el período de validación, lo que da lugar a errores de pronóstico ex post, es decir, $X_{N+h} - \widehat{X}_{N+h}$ para el horizonte $h = 1,\ldots,m$. Los errores fueron evaluados por las estadísticas de los residuales, como el error medio (ME), error cuadrático medio (RMSE), error absoluto medio (MAE), error porcentual medio (MPE) y error porcentual absoluto medio (MAPE), definidos a continuación:

$$\text{MAE:} \quad \frac{1}{n}\sum_{t=1}^{n}|X_t - \widehat{X}_t|,$$

$$\text{RMSE:} \quad \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(X_t - \widehat{X}_t\right)^2},$$

$$\text{MPE} = \quad \frac{1}{n}\sum_{t=1}^{n}100(X_t - \widehat{X}_t)/X_t,$$

$$\text{MAPE} = \quad \frac{1}{n}\sum_{t=1}^{n}100|(X_t - \widehat{X}_t)/X_t|$$

Los valores pequeños de estas estadísticas reflejan una mejor bondad de ajuste. La Tabla 3 informa las estadísticas de errores de pronóstico tanto para los datos de entrenamiento como los de prueba. En ambos casos, todos los indicadores favorecieron el modelo híbrido con ponderaciones que utilizaron errores en la muestra, es decir, el Modelo 4b con un MAPE de $0,62\%$ en los datos de prueba. Considerando lo señalado por Lewis (1982) (pág. 40) quien estableció criterios de interpretación de los valores del MAPE argumentando que, en un valor MAPE $<10\%$ la predicción tiene una alta precisión. En nuestro caso, todos los modelos tienen un MAPE menor al 10%; por lo tanto, se puede inferir que a pesar de la llegada de la variante Omicron los modelos de series de tiempo utilizados en este estudio siguen siendo válidos. En particular, se puede señalar que el Modelo 4b posse una mejor performance que el resto de los modelos, demostrando la idoneidad para la predicción del COVID-19.

Los hallazgos anteriores están respaldados por la Figura 4, donde se puede observar que las predicciones (puntos en color) del Modelo 4b son más cercanas a los valores reales denotados por los puntos de color negro.

4. CONCLUSIONES

Este estudio consideró el modelamiento de los datos de casos confirmados por COVID-19 en Chile, con el fin de establecer el mejor modelo para pronosticar el comportamiento de esta enfermedad en presencia de la variante Omicron. Para este propósito se utilizaron distintos modelos para ajustar los datos de COVID-19: SARIMA, Holt-Winter, Holt-Winter Damped Trend, Hybrid, BSTS, TBATS y LSTM.

Los hallazgos de este estudio concluyen que el modelo con menor error de pronóstico en datos de casos confirmados de COVID-19 en Chile es el modelo híbrido (Modelo 4b), el que

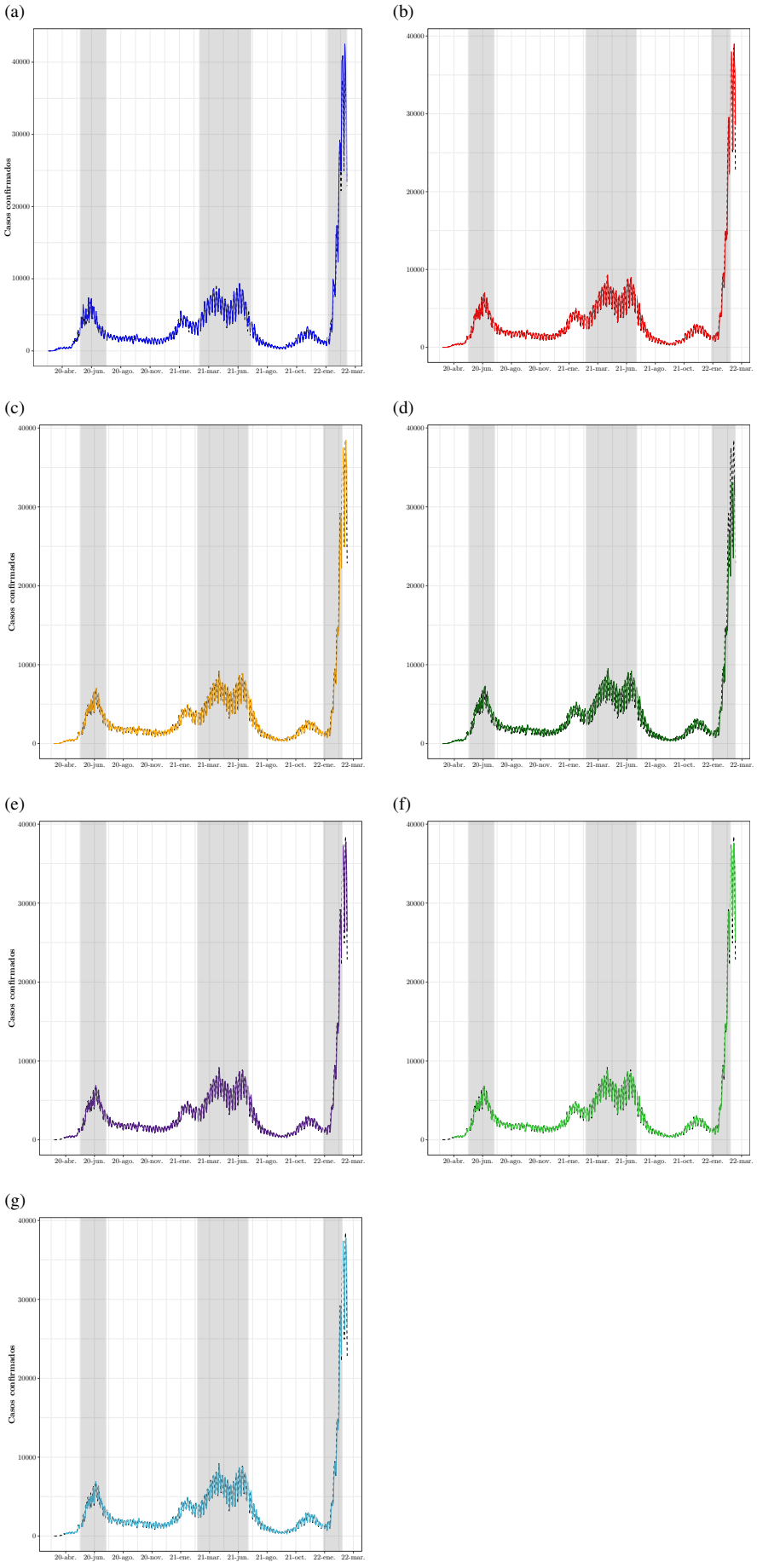

Figura 3. COVID-19 en Chile: Casos confirmados (líneas discontinuas negras) versus valores ajustados (línea continua). (a) Modelo 1. (b) Modelo 2a. (c) Modelo 2b. (d) Modelo 3. (e) Modelo 4a. (f) Modelo 4b. (g) Modelo 4c

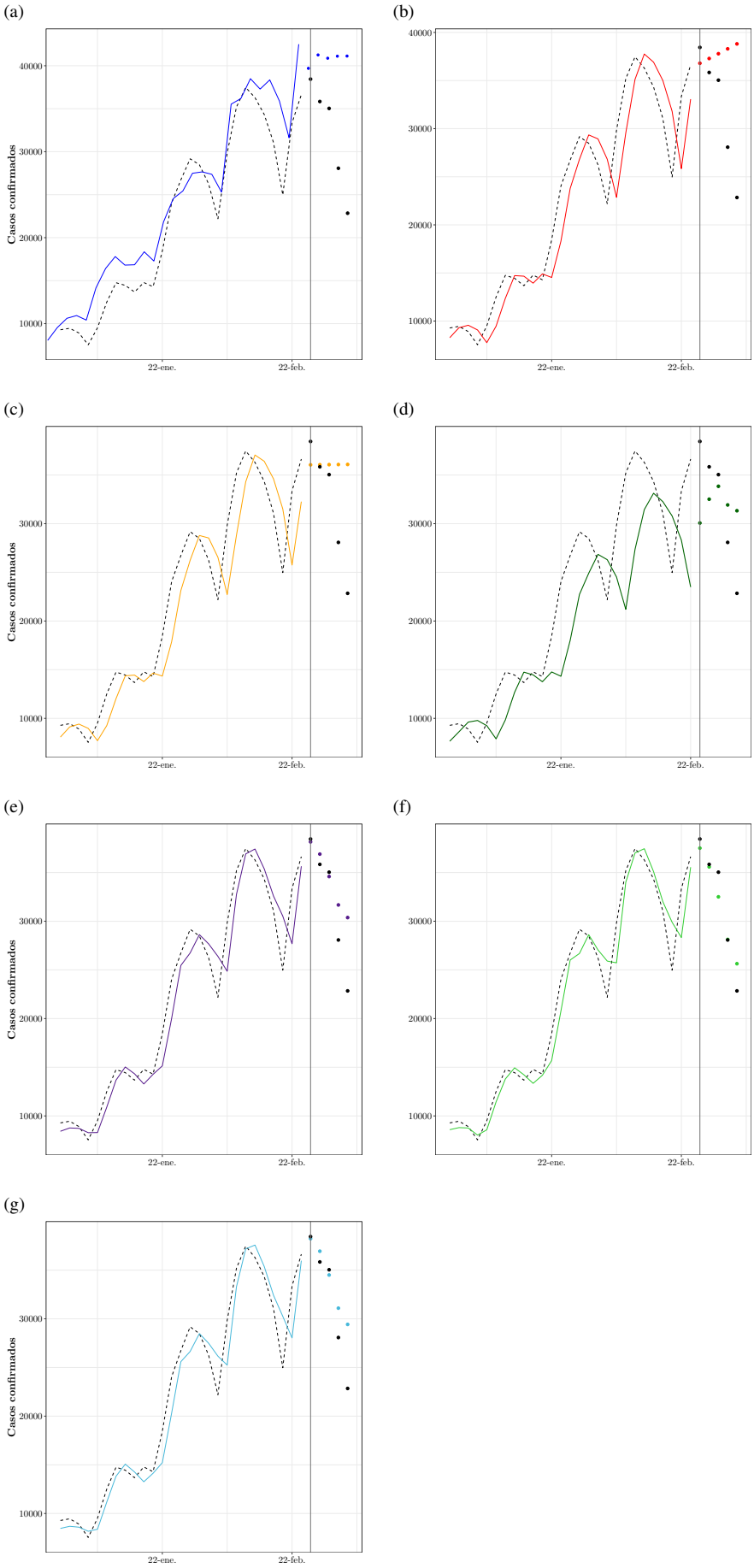

Figura 4. *Predicción para casos confirmados de COVID-19.* La línea continua y los puntos representan valores ajustados y el pronóstico ex post, respectivamente. (a) Modelo 1. (b) Modelo 2a. (c) Modelo 2b. (d) Modelo 3. (e) Modelo 4a. (f) Modelo 4b. (g) Modelo 4c

Tabla 3. Estadísticas descriptivas de los errores de pronóstico ex post

Modelo	ME	RMSE	MAE	MPE	MAPE
		Training Data			
Modelo 1	$-2{,}8\times10^{-3}$	0,21	0,16	-0,11	2,50
Modelo 2a	$5{,}3\times10^{-4}$	0,25	0,18	0,01	2,83
Modelo 2b	$5{,}6\times10^{-3}$	0,25	0,18	-0,11	2,79
Modelo 3	$-3{,}3\times10^{-2}$	0,35	0.28	-0.26	4.38
Modelo 4a	$3{,}6\times10^{-3}$	0,16	0,12	$6{,}8\times10^{-3}$	1,64
Modelo 4b	$\mathbf{2{,}7\times10^{-3}}$	**0,13**	**0.10**	$\mathbf{2{,}1\times10^{-3}}$	**1,37**
Modelo 4c	$3{,}2\times10^{-3}$	0,15	0,12	$3{,}3\times10^{-3}$	1,57
		Testing Data			
Modelo 1	-0,24	0,31	0,24	-2,35	2,35
Modelo 2a	-0,18	0,28	0,20	-1,79	1,97
Modelo 2b	-0,14	0,24	0,16	-1,34	1,59
Modelo 3	-0,01	0,19	0,16	-0,16	1,60
Modelo 4a	$-8{,}9\times10^{-2}$	0,14	0,09	-0,88	0,92
Modelo 4b	$\mathbf{-4.9\times10^{-2}}$	**0,09**	**0,06**	**- 0,49**	**0,62**
Modelo 4c	$-8{,}1\times10^{-2}$	0,13	0,08	-0,80	0,82

permite realizar una predicción con mayor precisión que los otros modelos considerados en este trabajo. En este contexto, podemos sostener que los modelos considerados tienen una alta precisión de predicción, siendo de gran utilidad para analizar el comportamineto dinámico temporal del COVID-19 frente a las distintas variantes presentes en el periodo de estudio. Se debe tener presente que, es difícil encontrar un modelo estadístico que tenga una precisión en la predicción de los datos cercana al 100%, dado que a menudo se exponen a una correlación serial y a estructuras de cambio no-estocásticas. Por lo tanto, no existe un modelo universal capaz de capturar de manera precisa el comportamiento futuro de las olas pandémicas por COVID-19, debido a que en el análisis de series de tiempo se deben considerar componentes de tendencias, ciclos y variables externas como las características biosociodemográficas particulares de cada país.

En particular, en nuestro estudio, los modelos incluidos siguen siendo útiles para capturar la evolución del número de contagios por COVID-19 en presencia de las diferentes variantes dominantes en Chile. A la vez, se destaca que, todos los modelos utilizados han sido validados por otros autores en diversas bases de datos de COVID-19 a nivel mundial. Para complementar lo anterior, se puede leer el trabajo realizado por Chakraborty et al. (2022) quienes han presentado un interesante resumen de los modelos utilizados para predecir el número de casos por COVID-19 en diferentes países e informaron cuales de estos modelos son los más apropiados para capturar la dinámica temporal de la pandemia.

Es relevante mencionar que este estudio, se incorpora al grupo de las escasas investigaciones existentes para analizar modelos de series de tiempo en presencia de la variante Omicron. Según la tendencia de los datos y los resultados de las predicciones, el número de casos confirmados por COVID-19 en Chile, está disminuyendo. Sin embargo, dado que la variante Omicron es más contagiosa se sugiere continuar en alerta y mantener las medidas preventivas indicadas por las autoridades sanitarias de Chile.

REFERENCIAS

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *USENIX Association, OSDI, 16*, 265-283.

Aggarwal, C.C. (2018). *Neural Networks and Deep Learning*. Springer.

Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International journal of forecasting, 16*(4), 521-530. https://doi.org/10.1016/S0169-2070(00)00066-2

Barría-Sandoval, C., Ferreira, G., Benz-Parra, K., & López-Flores, P. (2021). Prediction of confirmed cases of and deaths caused by COVID-19 in Chile through time series techniques: A comparative study. *Plos one, 16*(4), e0245414. https://doi.org/ 10.1371/journal.pone.0245414

Barría-Sandoval, C., Ferreira, G., Méndez, A., & Toffoletto, M. C. (2022). Impact of COVID-19 on deaths from respiratory diseases: Panel data evidence from Chile. *Infection Ecology & Epidemiology, 12*(1), 2023939. https://doi.org/10.1080/20008686.2021.2023939

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control. (5^aed.).JohnWiley&Sons.*

Brockwell, Peter J & Davis, Richard A. (2016). *Introduction to Time Series and Forecasting. (3^aed.)*. Springer.

Chakraborty, Tanujit and Ghosh, Indrajit and Mahajan, Tirna and Arora, Tejasvi. (2022). Nowcasting of COVID-19 confirmed cases: Foundations, trends, and challenges. *Modeling, Control and Drug Development for COVID-19 Outbreak Prevention.* Springer, 1023–1064.

Chollet, F. (2018). Keras: The Python Deep Learning library. Astrophysics Source Code Library. *ASCL Code Record.* https://ascl.net/1806.022

De Livera A. M., Hyndman, R. J., & Snyder, R. D. (2012). Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing. *Journal of the American statistical association, 106*(496), 1513-1527. https://doi.org/10.1198/jasa.2011.tm09771

Freire-Flores, D., Llanovarced-Kawles, N., Sanchez-Daza, A., & Olivera-Nappa, Á. (2021). On the heterogeneous spread of COVID-19 in Chile. *Chaos, Solitons & Fractals, 150*, 111156. https://doi.org/10.1016/j.chaos.2021.111156

Feroze, Navid. (2021). Assessing the future progression of COVID-19 in Iran and its neighbors using Bayesian models. *Infectious Disease Modelling, 6*, 343–350.

Gardner Jr, Everette S and McKenzie, ED. (1985). Forecasting trends in time series. *Management Science, 31*(10), 1237-1246. http://dx.doi.org/10.1287/mnsc.31.10.1237

Ghafouri-Fard, S., Mohammad-Rahimi, H., Motie, P., Minabi, M. A., Taheri, M., & Nateghinia, S. (2021). Application of machine learning in the prediction of covid-19 daily new cases: A scoping review. *Heliyon, 7*(10), e08143. https://doi.org/10.1016/j.heliyon.2021.e08143

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Holt, CC. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting, 20*(1), 5-10. https://doi.org/10.1016/j.ijforecast.2003.09.015

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., ... & Wang, E. (2020). Forecast: Forecasting functions for time series and linear models. *The R Foundation, package Version 8.16*. https://pkg.robjhyndman.com/forecast/

Ibrahim, R. R., & Oladipo, H. O. (2020). Forecasting the spread of COVID-19 in Nigeria using Box-Jenkins modeling procedure. *medRxiv*, 1-15. https://doi.org/10.1101/2020.05.05.20091686

Instituto de Salud Pública Ministerio de Salud Gobierno de Chile (2020). Vigilancia Genómica SARS-Cov2. Obtenido de: https://vigilancia.ispch.gob.cl/app/varcovid. (Marzo, 2020).

Lewis, C.D. (1982). Industrial and business forecasting methods. *London: Butterworths.*

McKinney, W. (2010). Data structures for statistical computing in python. *In Proceedings of the 9th Python in Science Conference, 445*(1), 51-56.

Oliphant, T. E. (2006). A guide to NumPy. *Trelgol Publishing*. https://web.mit.edu/dvp/Public/numpybook.pdf

Organización Mundial de la Salud (2020, Marzo). Alocución de apertura del Director General de la OMS en la rueda de prensa sobre la COVID-19 celebrada el 11 de marzo de 2020. Obtenido de: https://www.who.int/es/director-general/speeches.

Organización Mundial de la Salud (2021, Marzo). Seguimiento de las variantes del SARS-CoV-2. Obtenido de: https://www.who.int/es/activities/tracking-SARS-CoV-2-variants.

Perone, G.(2020). An ARIMA model to forecast the spread and the final size of COVID-2019 epidemic in Italy. *medRxiv* 1-14. https://doi.org/10.1101/2020.04.27.20081539.

Roda, Weston C and Varughese, Marie B and Han, Donglin & Li, Michael Y.(2020).Why is it difficult to accurately predict the COVID-19 epidemic?. *Infectious Disease Modelling.5*: 271–281. https://doi.org/10.1016/j.idm.2020.03.001

Sarkar, D., Biswas M. (2020). COVID 19 Pandemic: A Real-time Forecasts & Prediction of Confirmed Cases, Active Cases using the ARIMA model & Public Health in West Bengal, India. *medRxiv* 1-22. doi: https://doi.org/10.1101/2020.06.06.20124180.

Secretaría de Comunicaciones - MSGG Gobierno de Chile. (2021, Marzo). Cifras Oficiales COVID-19. Obtenido de: https://www.gob.cl/coronavirus/cifrasoficiales/datos.

Shaub, D., & Ellis, P. (2020). forecastHybrid: Convenient Functions for Ensemble Time Series Forecasts. *The R package version 5.0.19*. https://CRAN.R-project.org/package=forecastHybrid.

Shastri, S., Singh, K., Kumar, S., Kour, P., & Mansotra, V. (2020). Time series forecasting of covid-19 using deep learning models: India-usa comparative case study. *Chaos, Solitons & Fractals*, 140:110227. https://doi.org/10.1016/j.chaos.2020.110227

Talkhi, N., Fatemi, N. A., Ataei, Z., & Nooghabi, M. J. (2021). Modeling and forecasting number of confirmed and death caused COVID-19 in IRAN: A comparison of time series forecasting methods. *Biomedical Signal Processing and Control, 66*(102494), 1-8. https://doi.org/10.1016/j.bspc.2021.102494

Tariq, A., Undurraga, E. A., Laborde, C. C., Vogt-Geisse, K., Luo, R., Rothenberg, R., & Chowell, G. (2021). Transmission dynamics and control of COVID-19 in Chile, March-October, 2020. *PLoS neglected tropical diseases, 15*(1), e0009070. https://doi.org/10.1371/journal. pntd.0009070

Team, R.C. (2013). R: A Language and Environment for Statistical Computing. *The R Foundation*. http://www.R-project.org/

Tran, TT and Pham, LT and Ngo, QX.(2020).Forecasting epidemic spread of SARS-CoV-2 using ARIMA model (Case study: Iran). *Global Journal of Environmental Science and Management. 2020;6*(4), 1-10. https://doi.org/10.22034/GJESM.2019.06.SI.01

Vicuña, M. I., Vásquez, C., & Quiroga, B. F. (2021). Forecasting the 2020 COVID-19 epidemic: A multivariate Quasi-Poisson regression to model the evolution of new cases in Chile. *Frontiers in public health, 9* (610479), 1-7. https://doi.org/10.3389/fpubh.2021.610479

Winters, P.R. (1960). Forecasting sales by exponentially weighted moving averages.*Management Science, 6*(3), 231-362. https://doi.org/10.1287/mnsc.6.3.324

Yonar, Harun and Yonar, Aynur and Tekindal, Mustafa Agah & Tekindal, Melike.(2020). Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods.*Eurasian Journal of Medicine and Oncology 4*(2): 160–165. https://ejmo.org/10.14744/ejmo.2020.28273/

BIOGRAFÍAS

Claudia Barría-Sandoval, Académica de Enfermería de la Universidad de las Américas, Concepción, Chile. PhD (c) en Enfermería, Universidad de Concepción, Chile. Mg. en Salud Pública y Gestión Sanitaria, Universidad de Valencia, España. Sus intereses de investigación están orientados a la Salud Pública, Enfermería y Atención Integral en la Comunidad.

Patricio Salas, Profesor Asistente en el Departamento de Estadística de la Universidad de Concepción, Chile. PhD (c) en Ingeniería Industrial de la Universidad de Concepción, Chile. Mg. en Estadística Aplicada, Universidad de Concepción, Chile. Ingeniero Estadístico, Universidad de Concepción, Chile. Sus intereses de investigación incluyen modelos econométricos de elección discreta, machine learning, modelos basados en agentes (ABM), series de tiempo y simulación de eventos discretos.

Guillermo Ferreira, PhD en Estadística de la Pontificia Universidad Católica de Chile. Académico Asociado en el Departamento de Estadística de la Universidad de Concepción, Chile. Sus intereses de investigación incluyen el análisis y predicción de series de tiempo, modelos econométricos, procesos espacio-temporal y modelos epidemiológicos.