Universidad de Guadalajara
Seminario del Módulo de Estadística

Estimaciones de contagios por COVID-19 en México utilizando un modelo ARIMA

Erik Alberto Ríos Mena
Asesora: Mtra. Lizbeth Díaz Caldera

Mayo de 2022

Índice

Índice de cuadros

Índice de figuras

Introducción

La pandemia de COVID-19, causada por el virus SARS-CoV-2 – coloquialmente llamado coronavirus –, comenzó a principios de diciembre del 2019 en Wuhan, China, y llegó a México en febrero del 2020. Más de dos años después aún sufrimos los efectos de esta enfermedad y las restricciones que impone en la vida diaria. Una de las principales razones por las que el virus sigue afectando nuestras vidas después de tanto tiempo es su alta tasa de contagios, debido parcialmente a la capacidad de transmisión del virus SARS-CoV-2 incluso si una persona enferma no padece síntomas de COVID-19, así como sus mutaciones genéticas: la variante B.1.1.7, también denominada "Alfa", un poco más infecciosa y fatal[7]; la variante B.1.617.2, de nombre "Delta", igual de infecciosa pero más mortal[1]; y la variante denominada B.1.1.529, coloquialmente conocida como "Omicrón", menos severa en cuanto a síntomas y fatalidades, pero muchísimo más contagiosa[8]; entre muchas otras, las cuales no han sido denominadas "variantes preocupantes".

El propósito de este proyecto es, en base a los casos de COVID-19 en México, encontrar un modelo que nos permita predecir cómo avanzaría la pandemia a partir del mes de mayo del año 2022, utilizando métodos inferenciales para aplicarle un modelo a los casos diarios. Para ello, se considerarán los datos de casos confirmados recolectados por el *CONACyT*[4], cuya recolección inició el 26 de febrero del 2020 y continúa al día de hoy; en el contexto de este trabajo, se considerarán datos obtenidos hasta el 30 de abril del 2022. Utilizando métodos de series de tiempo, más precisamente, un modelo de autorregresión y media movible integrado – también llamado modelo ARIMA – con lo cual se busca realizar un ajuste de modelo a los datos obtenidos, vistos como una serie de tiempo, así como generar predicciones tras haber elegido los parámetros de modelo que se ajusten mejor a los datos. Para lograr esto, se probaron 1000 modelos, ajustados individualmente a la serie considerando sólo ciertos periodos de tiempo, y se escogió aquel que minimizase el error, de cierta manera.

1. Marco contextual

El modelado para ajustar a y pronosticar contagios y/o defunciones causados por diversas enfermedades es un tema de aplicaciones frecuente y la pandemia de SARS-CoV-2 no es la excepción. En marzo del 2020, algunos meses después de que se iniciara cuarentena en muchas regiones del mundo, Dehesh et al.[5] aplicaron una variedad de modelos ARIMA a los casos diarios en China, Italia, Corea del Sur, Irán y Tailandia. En abril, BBVA comisionó un estudio de Ng y Serrano[13] donde se aplicó un modelo a los casos diarios que había en México. Ese mismo mes, Tandon et al.[14] ajustaron un modelo ARIMA a los casos diarios en la India. En octubre de ese mismo año, Kumar y Susan[10] utilizaron dos métodos de series de tiempo, incluyendo un modelo ARIMA, y los comparan en los casos de diez países: Estados Unidos, España, Italia, Francia, Alemania, Rusia, Irán, el Reino Unido, Turquía e India, con el propósito de verificar cuál método da mejor ajuste a los datos. Más recientemente, a principios del 2022, Gowrisankar et al.[8] aplicaron un modelo de media movible – un caso especial del modelo ARIMA – a los casos positivos diarios en Dinamarca, Alemania, India, Países Bajos, Sudáfrica y el Reino Unido para intentar pronosticar cómo la variante "Omicrón", que en ese momento era más reciente, podría afectar a la cantidad de casos registrados en dichos países.

Claramente, como los ejemplos muestran arriba, aplicar modelos de autorregresión y media movible integrados a los datos relacionados a la pandemia – ya sea de casos diarios y/o defunciones – vistos como series de tiempo es muy popular y práctico. Sin embargo, no son por los únicos modelos. Por ejemplo, en un artículo publicado en 2021, Zumaya[15] aplicó un modelo de vector de corrección de error a los casos diarios de SARS-CoV-2 en México, vistos como una serie de tiempo. En mayo de 2020, Malki et al.[12] usaron una extensión del modelo ARIMA con los datos de casos diarios mundiales para intentar predecir cuándo podría terminar la pandemia y un riesgo de una segunda ola – en aquel momento –, y se llegó a que su modelo esperaba que el número de casos confirmados más altos sería entre diciembre de 2020 y abril de 2021. Claro, esto era antes de que la variante Omicrón apareciera, pero se acercó bastante considerando lo poco que se sabía entonces.

2. Fundamentos estadísticos

La gran parte de resultados de esta sección fueron obtenidos de [9].

2.1. Conceptos básicos de estadística

Antes de hacer inferencia sobre un conjunto de datos necesitamos entender cómo describir los datos que tenemos. Para ello, este capítulo introducirá brevemente algunos de estos conceptos.

Definición 2.1 (Valor esperado)**.** Sea X una variable aleatoria. La *esperanza* o *valor esperado* de X se denota $\mathrm{E}[X]$ y se define de la siguiente manera:

1. Si X tiene una distribución discreta,
$$\mathrm{E}[X] = \sum_x x f_X(x).$$
2. Si X tiene una distribución continua,
$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x.$$

En ocasiones, a la esperanza se le llama media. En esos casos, suele denotarse como μ.

Proposición 2.2. *Sean X, Y variables aleatorias con la misma distribución y $a, b \in \mathbb{R}$. Entonces*
$$\mathrm{E}[aX + bY] = a\,\mathrm{E}[X] + b\,\mathrm{E}[Y].$$

Demostración. Ambos casos son similares. Veamos que
$$\mathrm{E}[aX + bY] = \sum_x (ax + bx) f_X(x) = a \sum_x x f_X(x) + b \sum_y y f_Y(y) = a\,\mathrm{E}[X] + b\,\mathrm{E}[Y],$$
$$\mathrm{E}[aX + bY] = \int_{-\infty}^{\infty} (ax + bx) f_X(x)\,\mathrm{d}x = a \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x + b \int_{-\infty}^{\infty} y f_Y(y)\,\mathrm{d}y = a\,\mathrm{E}[X] + b\,\mathrm{E}[Y],$$
como queríamos. ∎

Definición 2.3 (Varianza)**.** Sea X una variable aleatoria y $\mu = \mathrm{E}[X]$. La *varianza* de X se denota $\mathrm{Var}[X]$ y se define como
$$\mathrm{Var}[X] := \mathrm{E}\left[(X - \mu)^2\right].$$

Proposición 2.4. *Sea X una variable aleatoria y $\mu = \mathrm{E}[X]$. Se cumple que*
$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2.$$

Demostración. Tenemos que
$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\ &= \mathrm{E}\left[X^2 - 2X\,\mathrm{E}[X] + \mathrm{E}[X]^2\right] \\ &= \mathrm{E}\left[X^2\right] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + \mathrm{E}[X]^2 \\ &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2,\end{aligned}$$
como queríamos. ∎

Definición 2.5 (Muestra aleatoria)**.** Sean $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, donde $\overset{iid}{\sim}$ indica que las variables son independientes (es decir, $\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]$) e idénticamente distribuidas (en este caso, todas tienen función de distribución f). En este caso, decimos que las variables aleatorias constituyen un *muestra aleatoria* de tamaño n con distribución f.

El concepto de muestra es fundamental a la inferencia. Ahora podemos introducir algunas características de la muestra relacionadas a la esperanza y varianza.

Definición 2.6 (Media muestral)**.** Sean $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$. La *media muestral* de las X_i se denota $\overline{X}$ y se define como

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i,$$

es decir, es la media aritmética de nuestras variables aleatorias.

Definición 2.7 (Varianza muestral)**.** Sean $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$. La *varianza muestral* de las X_i se denota s^2 y se define como

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Definición 2.8 (Covarianza muestral)**.** Sean $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$. La *covarianza muestral* de las X_i se denota s^2_{jk} y se define como

$$s^2_{jk} := \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X})(X_{ik} - \overline{X}).$$

Observación: el uso de $\frac{1}{n-1}$ en vez de $\frac{1}{n}$ es comúnmente llamado *corrección de Bessel*.

Finalmente, tenemos dos propiedades importantes relacionadas a sucesiones de variables, y un teorema que no será demostrado.

Definición 2.9 (Convergencia en distribución)**.** Sea $\{X_n\}$ una sucesión de variables aleatorias. Decimos que $\{X_n\}$ *converge en distribución* a X si

$$\lim_{n\to\infty} F_n(x) = F(x),$$

donde F_n es la función de distribución acumulada de X_n. Esto se denota como $X_n \xrightarrow{d} X$.

Definición 2.10 (Convergencia en probabilidad)**.** Sea $\{X_n\}$ una sucesión de variables aleatorias. Decimos que $\{X_n\}$ *converge en probabilidad* a X si

$$\lim_{n\to\infty} \mathrm{P}(|X_n - X| > \epsilon) = 0.$$

Esto se denota como $X_n \xrightarrow{p} X$.

Teorema 2.11 (Central del límite)**.** *Sean $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, con $\mathrm{E}[X_i] = \mu$ y $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Conforma $n \to \infty$, la sucesión de variables aleatorias*

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma}$$

convergen en distribución a una variable aleatoria normal estándar, esto es,

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \xrightarrow[n\to\infty]{d} N(0,1).$$

2.2. Estimadores y estimaciones

2.2.1. Introducción y propiedades

La teoría de estimación es una parte fundamental de la inferencia estadística, ya que la mayoría, si no es que todos los métodos utilizados al realizar hipótesis y pruebas inferenciales requieren en la práctica que estimemos ya sean los parámetros de alguna distribución, o la misma función de distribución o de densidad que toma una muestra de valores conocidos, los cuales suponemos vienen – formalmente, son una *realización* – de una muestra aleatoria.

Definición 2.12 (Estadístico)**.** Sean $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$. Un *estadístico* es simplemente una función de la muestra de variables aleatorias $T = T(X_1, \ldots, X_n)$.

Cuando $x_1, \ldots, x_n$ son realizaciones de una muestra aleatoria $X_1, \ldots, X_n$, denotamos a $T(x_1, \ldots, x_n)$ como t y lo llamamos *realización* del estadístico.

Definición 2.13 (Estimador puntual)**.** Sean $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$ y T un estadístico de estas variables. Denotamos a T como $\hat{\theta}$ y decimos que es un *estimador* o *estimador puntual* de θ si usamos a $\hat{\theta}$ para inferir sobre el valor verdadero de θ.
A la realización del estimador la llamamos *estimado*.

Existen varias propiedades de los estimadores que, de ser cumplidas, son un buen indicador de su calidad.

Definición 2.14 (Sesgo)**.** Sea $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$ y $\hat{\theta}$ un estimador de θ. El *sesgo* de $\hat{\theta}$ se define como

$$\operatorname{Sesgo}(\hat{\theta}) := \mathrm{E}[\hat{\theta}] - \theta.$$

Si $\operatorname{Sesgo}(\hat{\theta}) = 0$, decimos que $\hat{\theta}$ es un estimador *insesgado* de θ.

En términos intuitivos, el sesgo es una manera de ver cuánto esperamos que se aleje o acerque el valor de nuestro estimador comparado al parámetro real.

Definición 2.15 (Consistencia)**.** Sean $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$ y θ_n un estadístico de estas variables. Decimos que θ_n es un estimador *consistente* de θ si

$$\theta_n \xrightarrow[n\to\infty]{p} \theta.$$

La consistencia indica que, al aumentar nuestro tamaño de muestra, crece la probabilidad de que el estadístico que tomamos sea o se acerque el verdadero parámetro poblacional.
Para la siguiente propiedad necesitamos unos resultados preeliminares.

Definición 2.16 (Información de Fisher)**.** Sean $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$. Definimos a la *información de Fisher* de nuestra muestra como

$$I_n(\theta) = n\,\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right].$$

Cuando $n = 1$ denotamos a I_n simplemente como I.

La siguiente definición enumera una serie de condiciones que cumplen la mayoría de distribuciones tradicionales y que, más aún, ayudan a simplificar el cálculo de muchos resultados.

Definición 2.17 (Condiciones de regularidad)**.** Sean $X_1, \ldots, X_n \overset{iid}{\sim} f$. A las siguientes tres condiciones se les llaman *condiciones de regularidad*:

- Si $\theta \neq \theta'$, entonces las funciones de densidad son distintas; es decir, $F(x_i; \theta) \neq F(x_i; \theta')$.
- Para cualquier valor de θ la función de distribución $f(x_i; \theta)$ tiene el mismo soporte[1].
- Si θ_0 es el valor verdadero de θ, entonces θ_0 vive en el interior del espacio muestral Ω.

Observación: bajo condiciones de regularidad, la información de Fisher es equivalente a

$$I_n(\theta) = -n\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right]_\theta.$$

El recíproco de la información de Fisher nos da la mínima varianza que puede tener un estimador. Esto lo podemos formalizar por medio del siguiente resultado.

Teorema 2.18 (Cota de Cramér-Rao)**.** *Sean $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$ y $\hat{\theta}$ un estimador de θ. Si $\hat{\theta}$ es insesgado, la siguiente desigualdad se cumple:*

$$\mathrm{Var}[\hat{\theta}] \geq \frac{1}{I_n(\theta)}.$$

Demostración. Basta demostrar la cota para una variable aleatoria $X \sim f(x;\theta)$. Definamos a S como el *score*:

$$S = \frac{\partial}{\partial\theta}\log f(x;\theta) = \frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}.$$

Observemos que

$$E[S] = \int_{-\infty}^{\infty}\frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}f(x;\theta)\,\mathrm{d}\theta = \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x;\theta)\,\mathrm{d}\theta = \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}f(x;\theta)\,\mathrm{d}\theta = \frac{\partial}{\partial\theta}1 = 0,$$

así que $\mathrm{Cov}(S,\hat{\theta}) = E[S\hat{\theta}]$. Se sigue que

$$\begin{aligned}
\mathrm{Cov}(S,\hat{\theta}) &= E[S\hat{\theta}] \\
&= E\left[\hat{\theta}\frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}\right] \\
&= \int_{-\infty}^{\infty}\hat{\theta}\frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}f(x;\theta)\,\mathrm{d}\theta \\
&= \int_{-\infty}^{\infty}\hat{\theta}\frac{\partial}{\partial\theta}f(x;\theta)\,\mathrm{d}\theta \\
&= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\theta f(x;\theta)\,\mathrm{d}\theta \\
&= \frac{\partial}{\partial\theta}E[\hat{\theta}] \\
&= \frac{\partial}{\partial\theta}\theta \\
&= 1.
\end{aligned}$$

[1] En teoría de la medida, el *soporte* de una variable aleatoria es el conjunto cerrado $R_X \subset \mathbb{R}$ más pequeño donde $\Pr(X \in R_X) = 1$.

Para terminar, veamos que la desigualdad de Cauchy-Schwartz establece que

$$|\langle u, v\rangle|^2 \leq \langle u, u\rangle \cdot \langle v, v\rangle .$$

Definamos el producto interior de dos variables aleatorias X, Y como la esperanza de su producto[2]:

$$\langle X, Y\rangle = \mathrm{E}[XY].$$

Observemos que

$$\begin{aligned} |\operatorname{Cov}(S, \hat{\theta})|^2 &= \left|\mathrm{E}\left[(S - \mathrm{E}[S])(\hat{\theta} - \mathrm{E}[\hat{\theta}])\right]\right|^2 \\ &\leq \mathrm{E}\left[(S - \mathrm{E}[S])^2\right] \cdot \mathrm{E}\left[(\hat{\theta} - \mathrm{E}[\hat{\theta}])^2\right] \\ &= \operatorname{Var}[S] \operatorname{Var}[\hat{\theta}]. \end{aligned}$$

Como $\operatorname{Cov}(S, \hat{\theta}) = 1$ llegamos a que $\operatorname{Var}[S] \operatorname{Var}[\hat{\theta}] \geq 1$. Pero

$$\operatorname{Var}[S] = \mathrm{E}\left[(S - \mathrm{E}[S])^2\right] = \mathrm{E}[S^2] = \mathrm{E}\left[\left(\frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)}\right)^2\right] = I(\theta).$$

Dividiendo por $I(\theta)$ obtenemos lo que queríamos. ■

Definición 2.19 (Eficiencia)**.** Sean $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$ y $\hat{\theta}$ un estimador insesgado de θ. Decimos que $\hat{\theta}$ es *eficiente* si $\operatorname{Var}[\hat{\theta}] = \frac{1}{I_n(\theta)}$ (es decir, se alcanza la cota de Cramér-Rao).
La *eficiencia* de $\hat{\theta}$ se denota $e(\hat{\theta})$ y se define como

$$e(\hat{\theta}) = \frac{\frac{1}{I_n(\theta)}}{\operatorname{Var}[\hat{\theta}]}$$

Por la cota de Cramér-Rao se sigue que $e(\hat{\theta}) \leq 1$. Valores cercanos a 1 indican un mejor estimador. Podemos definir un concepto análogo a la eficiencia para comparar dos estimadores utilizando sus varianzas.

Definición 2.20 (Eficiencia relativa)**.** Sean $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$ y T_1, T_2 estimadores insesgados de θ. La *eficiencia relativa* de T_1 y T_2 se denota $\operatorname{ER}(T_1, T_2)$ y se define como

$$\operatorname{ER}(T_1, T_2) = \frac{\operatorname{Var}[T_2]}{\operatorname{Var}[T_1]}.$$

En caso de que la eficiencia relativa no dependa de θ, lo cual sucede en muchas ocasiones, diremos que T_1 es preferible a T_2 si $\operatorname{ER}(T_1, T_2) > 1$, con T_1 siendo preferible si la eficiencia relativa es menor a 1. En caso de igualdad, necesitaríamos considerar criterios distintos. El siguiente podría ser de utilidad, pues no requiere que los estimadores sean insesgados, así que es un modo de verificar el error de un estimador con más generalidad.

Definición 2.21 (Error cuadrático medio)**.** Sean $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$ y $\hat{\theta}$ un estimador de θ. El *error cuadrático medio* (comúnmente abreviado ECM) de $\hat{\theta}$ se denota $\operatorname{ECM}(\hat{\theta})$ y se define como

$$\operatorname{ECM}(\hat{\theta}) = \mathrm{E}\left[(\hat{\theta} - \theta)^2\right].$$

[2] Es fácil ver que esto es un producto interior sobre $\mathbb{R}$ pues, como las variables aleatorias conmutan y distribuyen, $\langle X, Y\rangle = \mathrm{E}[XY] = \mathrm{E}[YX] = \langle Y, X\rangle$; $\langle aX + bY, Z\rangle = \mathrm{E}[(aX + bY)Z] = \mathrm{E}[aXZ + bYZ] = a\,\mathrm{E}[XZ] + b\,\mathrm{E}[YZ] = a\langle X, Z\rangle + b\langle Y, Z\rangle$; y como $X^2 \geq 0$, se sigue que $\langle X, X\rangle = E[X^2] \geq 0$.

El error cuadrático medio de un estimador puede ser expresado en términos de su varianza y sesgo de la siguiente manera.

Proposición 2.22. *Sean $X_1, \dots, X_n \overset{iid}{\sim} f(x;\theta)$ y $\hat{\theta}$ un estimador de θ. La siguiente igualdad se cumple:*

$$\mathrm{ECM}(\hat{\theta}) = \mathrm{Var}[\hat{\theta}] + \mathrm{Sesgo}(\hat{\theta})^2.$$

Demostración. Por definición del ECM,

$$\begin{aligned}\mathrm{ECM}(\hat{\theta}) &= \mathrm{E}\left[(\hat{\theta}-\theta)^2\right] \\ &= \mathrm{E}[\hat{\theta}^2] + \mathrm{E}[\theta^2] - 2\theta\,\mathrm{E}[\hat{\theta}] \\ &= \mathrm{Var}[\hat{\theta}] + \mathrm{E}[\hat{\theta}]^2 + \theta^2 - 2\theta\,\mathrm{E}[\hat{\theta}] \\ &= \mathrm{Var}[\hat{\theta}] + (\mathrm{E}[\hat{\theta}] - \theta)^2 \\ &= \mathrm{Var}[\hat{\theta}] + \mathrm{Sesgo}(\hat{\theta})^2,\end{aligned}$$

como queríamos. ∎

Se sigue un corolario cuya demostración es trivial.

Corolario 2.23. *Si $\hat{\theta}$ un estimador insesgado de θ entonces* $\mathrm{ECM}(\hat{\theta}) = \mathrm{Var}[\hat{\theta}]$.

2.2.2. Estimadores de máxima verosimilitud

En la práctica, la mayoría de los casos no conocemos el valor verdadero θ_0 de un parámetro θ a estimar. Entonces, nuestro objetivo es encontrar un estimación la cual se asemeje lo más posible a θ_0. Un método para encontrar a dicho estimador se basa en la función de verosimilitud, introducida a continuación.

Definición 2.24 (Función de verosimilitud)**.** Sean $X_1, \dots, X_n \sim f(x;\theta)$. La *función de verosimilitud* de θ, denotada $L(\theta)$, es la distribución conjunta de las variables aleatorias X_i vistas como funciones de θ. Si estas son independientes podemos expresar a L como

$$L(\theta) := \prod_{i=1}^{n} f(X_i;\theta).$$

La *log-verosimilitud* de θ se denota como $\ell(\theta)$ y es simplemente $\log L(\theta)$. En el caso anterior:

$$\ell(\theta) := \sum_{i=1}^{n} \log f(X_i;\theta).$$

En casi todos los casos es preferible trabajar con la log-verosimilitud pues es más sencilla, como se verá a continuación.
Ahora bien, buscamos estimar a θ en base a L (ó ℓ). Para esto, buscamos el valor donde esta alcance su máximo, pues esto es donde existe más parecido con θ_0, dado que sólo conocemos la muestra.

Definición 2.25 (Estimador de máxima verosimilitud)**.** Sean $X_1, \dots, X_n \sim f(x;\theta)$. El *estimador de máxima verosimilitud* de θ (comúnmente abreviado EMV o MLE) se denota $\hat{\theta}_n$ y es el punto donde la función de verosimilitud se maximiza, es decir,

$$\hat{\theta}_n := \arg\max_{\theta} L(\theta).$$

Cuando conocemos a f, y esta es diferenciable, podemos maximizar a L (ó a ℓ) resolviendo el sistema de ecuaciones $\frac{\partial L}{\partial \theta} = 0$. En muchos casos no conocemos a f, o el sistema no tiene una forma en funciones elementales, así que se acostumbra a maximizar la función utilizando métodos numéricos como descenso de gradiente.

A continuación introducimos algunas propiedades de los estimadores de máxima verosimilitud.

Teorema 2.26. *Sean $X_1, \dots, X_n \sim f(x;\theta)$ y $\hat{\theta}_n$ el estimador de máxima verosimilitud de θ. Bajo condiciones de regularidad las siguientes propiedades se cumplen:*

1. *Invarianza: si $\alpha = g(\theta)$ entonces $\hat{\alpha_n} = g(\hat{\theta}_n)$.*
2. *Consistencia: el EMV de θ converge en probabilidad al valor verdadero, i.e. $\hat{\theta}_n \xrightarrow[n\to\infty]{p} \theta_0$.*
3. *Normalidad asintótica: la sucesión converge en distribución a una normal estándar:*

$$\frac{\sqrt{n}\left(\hat{\theta}_n - \theta\right)}{\sqrt{\frac{1}{I_n(\theta)}}} \xrightarrow[n\to\infty]{d} N(0,1).$$

Demostración. Tenemos lo siguiente.

1. Tomemos a la preimagen de $g(\alpha)$, para cualesquier α:

   $$g^{-1}(\alpha) := \{\theta : g(\theta) = \alpha\}.$$

   El máximo de esta ocurre en $\hat{\theta}_n$. Como el dominio de g es el espacio parametral, tenemos que $\hat{\theta}_n \in g^{-1}(\alpha)$. Tomamos a $\hat{\alpha_n}$ tal que $g^{-1}(\hat{\alpha_n})$ sea la única preimagen que contenga a $\hat{\theta}_n$. Así $\hat{\alpha_n} = g(\hat{\theta}_n)$.

2. Sean $\epsilon > 0$ y

   $$S_n := \{\boldsymbol{X} : \ell(\theta_0; \boldsymbol{X}) > \ell(\theta_0 - \epsilon; \boldsymbol{X})\} \cap \{\boldsymbol{X} : \ell(\theta_0; \boldsymbol{X}) > \ell(\theta_0 + \epsilon; \boldsymbol{X})\}$$

   un evento. En el emv de θ restringido a $(\theta_0 - \epsilon, \theta_0 + \epsilon)$ tenemos que $\ell'(\hat{\theta}_n) = 0$, así que

   $$S_n \subset \left\{\boldsymbol{X} : \left|\hat{\theta}_n - \theta\right| \le \epsilon\right\} \cap \left\{\boldsymbol{X} : \ell'\left(\hat{\theta}_n\right) = 0\right\}.$$

   Luego

   $$1 = \lim_{n\to\infty} S_n \le \limsup_{n\to\infty} \mathrm{P}\left(\left\{\boldsymbol{X} : \left|\hat{\theta}_n - \theta\right| \le \epsilon\right\} \cap \left\{\boldsymbol{X} : \ell'\left(\hat{\theta}_n\right) = 0\right\}\right) \le 1,$$

   y por el teorema del sándwich la probabilidad es 1 cuando $n \to \infty$, i.e.,

   $$\mathrm{P}\left(\left\{\boldsymbol{X} : \left|\hat{\theta}_n - \theta\right| \le \epsilon\right\}\right) = 1 \Longrightarrow \mathrm{P}\left(\left\{\boldsymbol{X} : \left|\hat{\theta}_n - \theta\right| > \epsilon\right\}\right) = 0,$$

   que es la definición de consistente.

3. Consecuencia del teorema central del límite.

■

2.3. Intervalos de confianza

En ocasiones prácticas necesitamos más que sólo una estimación de un valor para el parámetro. Existe más incertidumbre y por ende nos gustaría decir con una confianza cuáles valores podría tomar un parámetro de algún modelo. Es por ello que introduciremos un intervalo o región donde podemos hablar de más valores, pero primero necesitamos unos conceptos auxiliares.

Definición 2.27 (Cantidad pivotal)**.** Sean $X_1, \ldots, X_n \sim f(x;\theta)$ y $Q = Q(X_1, \ldots, X_n; \theta)$ una función de la muestra y el parámetro θ. Si la distribución de Q no depende de θ, es decir, $Q\ g(x;\alpha)$ para algún $\alpha \neq \theta$, llamamos a Q una *cantidad pivotal*.

Las cantidades pivotales nos permiten expresar a θ como una variable aleatoria la cuál no depende de sí misma, cosa que las hace indispensables para los intervalos de confianza pues podemos asignar probabilidades a θ sin saber su valor en primer lugar. Ahora sí podemos introducir los intervalos de confianza.

Definición 2.28 (Intervalo de confianza)**.** Sean $X_1, \ldots, X_n \sim f(x;\theta)$, $\alpha \in (0,1)$ y $L = L(X_1, \ldots, X_n)$ y $U = U(X_1, \ldots, X_n)$ dos estadísticos de nuestra muestra. Llamamos *intervalo de confianza* de $100(1-\alpha)$ % al intervalo (L, U) si

$$\Pr(\theta \in (L, U)) = 1 - \alpha.$$

Definición 2.29 (Intervalo de confianza con cantidad pivotal)**.** Sean $X_1, \ldots, X_n \sim f(x;\theta)$, $\alpha \in (0,1)$ y Q una cantidad pivotal para θ. Llamamos *intervalo de confianza* de $100(1-\alpha)$ % al intervalo (l, u), $l, u \in \mathbb{R}$, si

$$\Pr(a < Q < b) = 1 - \alpha.$$

2.4. Pruebas de hipótesis

Ya introdujimos la estimación puntual y los intervalos de confianza como métodos para realizar inferencia sobre un parámetro. Otra clase de pruebas es similar en un cierto sentido al método aplicado en ciencias naturales donde, dada una hipótesis, realizamos alguna prueba o experimento para llegar a una conclusión. En estadística, llamamos una *prueba de hipótesis* a esta misma idea. Para formalizarla, se introducen los siguientes conceptos.

Definición 2.30. Sea $X \sim f(x;\theta)$. Sea $\{\omega_1, \omega_2\}$ una partición[3] del espacio muestral Ω; entonces $\theta \in \omega_1$ ó $\theta \in \omega_2$. Denotemos H_0 a lo primero y H_1 a lo segundo.

- A $H_0 : \theta \in \omega_1$ se le llama *hipótesis nula*.
- A $H_1 : \theta \in \omega_2$ se le llama *hipótesis alternativa*.

Definición 2.31 (Prueba de hipótesis)**.** Sean $H_0 : \theta \in \omega_1$, $H_1 : \theta \in \omega_2$ las hipótesis nula y alternativa respectivamente. Una *prueba de hipótesis* para θ es un regla de decisión con la cual elegir si se rechaza H_0 en favor de H_1 o no.

Definición 2.32. Sea $\mathcal{D} = \{(x_1, \ldots, x_n) : (x_1, \ldots, x_n) \text{ es un posible valor de } (X_1, \ldots, X_n)\}$, donde $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$ es una muestra aleatoria. Consideremos la prueba de hipótesis siguiente:

$$\text{Rechazar } H_0 : \theta \in \omega_0 \text{ si } (X_1, \ldots, X_n) \in C,\ C \subset \mathcal{D}.$$

Al subconjunto C se le llama *región crítica*. Si T es un estadístico y $C = \{(X_1, \ldots, X_n) : T > c\}$, a T se le llama *estadístico de prueba* y a c se le llama *valor crítico*.

[3] Es decir, $\omega_1 \cup \omega_2 = \Omega$ y $\omega_1 \cap \omega_2 = \emptyset$.

La tabla 1 muestra los resultados de realizar la prueba de hipótesis anterior comparado con lo que realmente pasa en la realidad. Aunque en la práctica es imposible evitar completamente los errores, uno busca la manera de minimizarlos lo más que sea posible al tomar una decisión.

	H_0 es verdadera	H_1 es verdadera
Rechazar H_0	Error tipo I	Correcto
No rechazar H_0	Correcto	Error tipo II

Cuadro 1: Tabla de decisión para una prueba de hipótesis.

Definición 2.33. Sea $H : \theta \in \omega_1$.

- Decimos que H es una *hipótesis simple* si es de la forma $H : \theta = \theta_0$.
- Decimos que H es una *hipótesis compuesta* si se conforma de varias hipótesis simples; por ejemplo, $H : \theta < \theta_0$ o $H : \theta \neq \theta_0$.

Definición 2.34 (Tamaño de una región crítica)**.** Decimos que una región crítica $C \subset \mathcal{D}$ es de *tamaño* α si

$$\max_{\theta \in \omega_1} \Pr((X_1, \ldots, X_n) \in C|\theta).$$

A α también se le conoce como *nivel de significación*.

Una manera intuitiva de ver a α es como el máximo de la probabilidad de cometer un error de tipo I.
En la práctica para no obtener resultados o conclusiones distintas dadas dos hipótesis similares, observar los mismos datos y aplicar la misma prueba, los estadísticos buscan ajustar el tamaño de la prueba α para no ser tan aleatorios.

Definición 2.35 (Valor p)**.** Sean $H_0 : \theta \in \omega_1$, $H_1 : \theta \in \omega_2$ las hipótesis nula y alternativa respectivamente. El *valor* p es una probabilidad condicional de obtener un resultado como alguno ya observado o más extremo, dado que la hipótesis nula es verdad. Si el valor de p es menor que algún nivel de significación $\alpha \in (0, 1)$, establecido previamente, se rechaza la hipótesis nula.

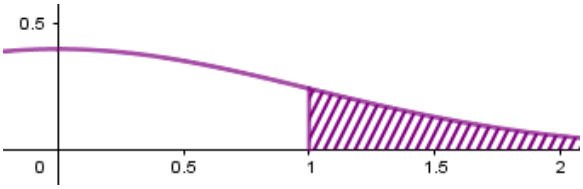

Figura 1: Gráficamente, el valor p (el valor del área sombreada) es la probabilidad del resultado ya observado o uno más extremo, suponiendo que la hipótesis nula es verdad.

Hay que tener cuidado al utilizar los valores de p: no rechazar la hipótesis nula *no implica* aceptarla, o que si se cumple siempre con error α. Las pruebas de hipótesis deben ser complementadas con intervalos o regiones de confianza, también para el mismo α, y estimaciones de los parámetros los cuales queramos probar. Existen otros métodos de contrastar hipótesis, como bondad de ajuste y funciones potencia, que no serán vistas aquí pues no se utilizan. Dicho esto, no se dependerá exclusivamente de los valores de p a la hora de probar hipótesis, sino que se verificarán luego con otras pruebas y/o supuestos en la sección de análisis de datos.

2.5. Regresión lineal

Esta sección sólo tratará con regresión lineal pues es una introducción suficiente para los modelos de autorregresión y media movible vistos más adelante. Para introducir un modelo lineal, consideremos el problema de verificar la relación entre dos variables, para ver cómo se espera cambiar una en función de la otra. Digamos, tenemos un conjunto de datos $\{y_i, x_{i1}, \ldots, x_{ip}\}$, $i = 1, \ldots, n$. Vamos a suponer que existe una relación lineal entre los valores $y_1, \ldots, y_p$ y los $x_{i1}, \ldots, x_{ip}$ es lineal, es decir,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \ i = 1, \ldots, n,$$

donde $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Buscamos encontrar los coeficientes β_i que mejor se ajusten a nuestros datos. Vamos a formalizar este concepto.

Definición 2.36 (Modelo lineal)**.** Sea $\{y_i, x_{i1}, \ldots, x_{ip}\}$, $i = 1, \ldots, n$ un conjunto de datos. Un *modelo de regresión lineal* para y_i con relación a x_i está dado por el sistema de ecuaciones

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \ i = 1, \ldots, n,$$

donde $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. A los β_i se les llama *parámetros* del modelo, y a y_i se le conoce como *variable dependiente*.

En el caso que $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ puede ser demostrado que la estimación de máxima verosimilitud de los parámetros $\hat{\beta}_i$ es análoga a su estimación por el método de *menores cuadrados ordinarios*. Para realizar esta estimación, reescribimos a nuestro modelo en forma matricial:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i,$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\begin{aligned} Y &= X\beta \\ X^T Y &= (X^T X)\beta \\ \beta &= (X^T X)^{-1} X^T Y, \end{aligned}$$

siempre que $(X^T X)^{-1} X^T$ exista. Ahora bien, para que un modelo de regresión lineal sea válido, se espera que cumpla las siguientes propiedades.

- *Linealidad*: es decir, que la esperanza de la variable dependiente sea sólo una combinación lineal de los parámetros y las variables dependientes. Aquí, la linealidad se refiere a los coeficientes del modelo y no de las variables.
- *Varianza constante*: la varianza de los errores no depende de los valores de las variables dependientes. Se sigue que la variabilidad de estas, para valores fijos de las variables independientes, es igual sin importar qué valores tomen las variables dependientes. Para revisar

este supuesto, se puede examinar una gráfica de residuos con las variables dependientes para verificar que no haya un "embudo" (i.e. un patrón vertical que crece o decrece a medida que se va de izquierda a derecha).

- *Errores independientes*: no hay correlación en los errores de las variables dependientes.

Los modelos de regresión lineal forman la base de la predicción con series de tiempo. A pesar de ser un modelo muy simple, y a veces con deficiencias al momento de querer pronosticar, pueden ser adaptados para crear modelos más avanzados que se usan hoy en día con aplicaciones que van desde finanza hasta investigación.

3. Métodos estadísticos aplicados

3.1. Series de tiempo

Primero comenzamos introduciendo el concepto de serie de tiempo y un modelo para estas. Las definiciones para este capítulo fueron obtenidas de [2].

Definición 3.1 (Serie de tiempo)**.** Una *serie de tiempo* o una *serie temporal* es un conjunto de datos $\{x_t\}$ registrados en un tiempo específico $t \in T_o$.

- Una *serie de tiempo discreta* es aquella en la cual el conjunto de valores para el tiempo T_o consiste en observaciones discretas. Por ejemplo, T_o podría ser un conjunto finito o $\mathbb{N}$.
- Una *serie de tiempo continua* es aquella cuyo conjunto de valores para el tiempo T_o es un intervalo continuo, e.g. de la forma $T_0 = [a, b]$, $a, b \in \mathbb{R}$.

En el contexto de este proyecto se trabajará con une serie de tiempo discreta, teniendo la serie de los casos diarios de COVID-19 en México. Ahora bien, en este caso nuestro conjunto de observaciones proviene de un suceso el cuál es influenciado por un comportamiento aleatorio (en este contexto, no sabemos cuándo el virus SARS-CoV-2 podría mutar, volviéndose más contagioso y/o letal, así como los comportamientos de la población, las restricciones impuestas por un gobierno, etcétera). Para poder lograr nuestro objetivo de intentar predecir a partir de los datos observados, necesitamos darle más forma a nuestros datos.

Definición 3.2 (Modelo de serie de tiempo)**.** Sea $\{x_t\}$ una serie de tiempo. Un *modelo* para nuestra serie es una especificación de las distribuciones conjuntas, o en algunos casos, sólo de las medias y covarianzas, de una sucesión de variables aleatorias $\{X_t\}$, suponiendo que las observaciones en nuestra serie $\{x_t\}$ son una realización de dichas variables.
Comúnmente, se le llama simplemente *serie de tiempo* a las observaciones $\{x_t\}$ junto con dicha realización.

En la práctica se tiende a sólo especificar los primeros y segundos momentos de las distribuciones conjuntas $\mathrm{E}[X_t]$ y $\mathrm{E}[X_t X_{t+h}]$, $t = 1, \ldots, n$, $h = 0, \ldots, n$, esto debido a que intentar especificar la distribución conjunta de todas las variables aleatorias $\{X_t\}$ – o de sus probabilidades – requerirá, en la mayoría de los casos, de muchos parámetros a estimar, lo cual no es práctico.
A continuación se introducirán dos modelos los cuales son comúnmente utilizados.

Definición 3.3 (Modelo con tendencia)**.** Sea $\{x_t\}$ una serie de tiempo que corresponde a una sucesión de variables aleatorias $\{X_t\}$. Un *modelo con tendencia* para $\{x_t\}$ está definido por la fórmula

$$X_t = m_t + Y_t,$$

donde Y_t es una variable aleatoria con media cero y m_t es una función de t llamada *tendencia*.

La función de tendencia normalmente describirá un fenómeno que constantemente está en incremento o decremento. Usualmente esta tendencia será de la forma $m_t = a_0 + a_1 t + a_2 t^2$, dependiendo de como sean nuestros datos, y la podremos aproximar a nuestros datos utilizando la técnica de mínimos cuadrados.

Definición 3.4 (Modelo con temporalidad)**.** Sea $\{x_t\}$ una serie de tiempo que corresponde a una sucesión de variables aleatorias $\{X_t\}$. Un *modelo con temporalidad* para $\{x_t\}$ está definido por la fórmula

$$X_t = s_t + Y_t,$$

donde Y_t es una variable aleatoria con media cero y s_t es una función de t, con periodo d, conocida como *temporada*.

La función de temporada usualmente describirá un fenómeno el cual varía de la misma forma cada cierto tiempo d – por ejemplo, en las estaciones del año, o cada ciertos meses. En la práctica una función de temporada suele estar dada como

$$s_t = \sum_{i=1}^{k} \left[a_i \sin(\lambda_i t) + b_i \cos(\lambda_i t) \right], \quad \lambda_i = \frac{2\pi k_i}{d},\ k_i \in \mathbb{Z}.$$

Esto es debido a que las funciones trigonométricas, con su forma de onda, en la mayoría de los casos son una estimación apropiada para un conjunto de datos con temporalidad. En este caso buscamos aproximar los coeficientes a_i y b_i por medio de regresión harmónica.
En general, un modelo para una serie de tiempo estará descrito tanto por una función de tendencia como una de temporada, más una variable aleatoria de media cero. A continuación introducimos las funciones utilizadas para detectar si una serie tiene algunos de estos componentes.

Definición 3.5. Sea $\{X_t\}$ una serie de tiempo (en este caso aleatoria) tal que $\mathrm{E}\left[X^2\right] < \infty$.

- La *función de media* de $\{X_t\}$, denotada μ_X, está dada por la esperanza

$$\mu_X(t) := \mathrm{E}[X_t].$$

- La *función de covarianza* de $\{X_t\}$, denotada γ_X, está dada por la covarianza

$$\gamma_X(r, s) := \mathrm{Cov}(X_r, X_s).$$

  Cuando sólo se especifica un parámetro h, la función de covarianza se define como

$$\gamma_X(h) := \gamma_X(h, 0) = \gamma_X(t + h, t) = \mathrm{Cov}(X_{t+h}, X_t).$$

Definición 3.6 (Serie de tiempo estacionaria)**.** Sea $\{X_t\}$ una serie de tiempo (en este caso aleatoria) tal que $\mathrm{E}\left[X^2\right] < \infty$. Decimos que $\{X_t\}$ es *estacionaria*[4] si cumple las siguientes condiciones:

- $\mu_X(t)$ no depende de t,
- $\gamma_X(t + h, h)$ no depende de t, para toda h.

Definición 3.7. Sea $\{X_t\}$ una serie de tiempo (en este caso aleatoria) estacionaria.

- La *función de autocovarianza* (abreviada FACV) de $\{X_t\}$ con *desfase* o *retraso h* se define como $\gamma_X(h) := \mathrm{Cov}(X_{t+h}, X_t)$.
- La *función de autocorrelación* (abreviada FAC) de $\{X_t\}$ con *desfase* o *retraso h* se define como

$$\rho_X(h) := \frac{\gamma_X(h)}{\gamma_X(0)} = \mathrm{Cor}(X_{t+h}, X_t).$$

[4] En un sentido estricto decimos que $\{X_t\}$ es *débilmente estacionaria*. Una serie de tiempo estacionaria usualmente requeriría conocer las distribuciones de toda X_t. En la práctica esto usualmente no es posible.

Proposición 3.8. *Las funciones de autocovarianza $\gamma_X(h)$ y de autocorrelación $\rho_X(h)$ son pares.*

Demostración. En efecto,

$$\gamma_X(h) = \mathrm{Cov}(X_{t+h}, X_t) = \mathrm{Cov}(X_t, X_{t+h}) = \gamma_X(-h),$$

mostrando que $\gamma_X(h)$ es par. Para $\rho_X(h)$ se sigue inmediatamente de este hecho, por definición. ■

En la práctica la mayoría de las veces no conocemos el modelo de nuestra serie, sino que tenemos un conjunto de datos $\{x_t\}$. De manera similar a los métodos de estimación de parámetros, usamos funciones de la muestra independientes de cualquier distribución.

Definición 3.9. Sea $\{x_t\}$ una serie de tiempo que corresponde a una sucesión de variables aleatorias $\{X_t\}$.

- La *función de media muestral* de $\{x_t\}$ es simplemente la media muestral

$$\overline{x} := \frac{1}{n}\sum_{t=1}^{n} x_t.$$

- La *función de autocovarianza muestral* de $\{x_t\}$ con *desfase* o *retraso h* se define como

$$\hat{\gamma}(h) := \frac{1}{n}\sum_{t=1}^{n-|h|} (x_{t+|h|} - \overline{x})(x_t - \overline{x}),\ -n < h < n.$$

- La *función de autocorrelación muestral* de $\{x_t\}$ con *desfase* o *retraso h* se define como

$$\hat{\rho}(h) := \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)},\ -n < h < n.$$

Observemos que la función de autocovarianza muestral $\hat{\gamma}$ como está definida arriba no incluye la corrección de Bessel. Esto es para que la matriz de covarianzas muestrales sea positiva semidefinida. Más adelante veremos que esto es importante, para definir la función de autocorrelación parcial. Además, a diferencia de la ACF y la ACVF, estas funciones pueden ser aplicadas a una serie de tiempo que no es estacionaria. De hecho, graficarlas es muy útil para ver si se tiene una serie que es estacionaria o no. Por ejemplo, series con tendencia m_t resultarán en $\hat{\rho}$ incrementando o decayendo en un periodo de tiempo, y series con temporada s_t tendrán un claro comportamiento periódico en $\hat{\rho}$.

3.2. Autorregresión y media movible

Antes de introducir un modelo de autorregresión básico, necesitamos de las siguiente definiciones y resultados.

Definición 3.10 (Ruido blanco)**.** Sean $X_0, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$ tal que $\mathrm{E}[X_t] = 0$, $t = 0, \ldots, n$. A la sucesión $\{X_t\}$ se le llama *ruido iid*. Si, más aún, $\mathrm{Cor}(X_t, X_{t+h}) = 0$, $t = 0, \ldots, n$, $h = 1, \ldots, n$, decimos que $\{X_t\}$ es *ruido blanco*.

Definición 3.11 (Serie de tiempo Gaussiana)**.** Sea $\{X_t\}$ una serie de tiempo. Si, para cualesquiera $i_1, \ldots, i_n$, el vector $(X_{i_1}, \ldots, X_{i_n})^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (es decir, todo vector aleatorio conformado por algunas $X_i \in \{X_t\}$ es normal multivariado), entonces a $\{X_t\}$ se le llama una *serie de tiempo Gaussiana*.

Recordemos que en general podemos describir a una serie de tiempo como

$$X_t = m_t + s_t + Y_t,$$

donde Y_t tiene media cero. Entonces, si toda Y_t es idénticamente ruido iid, podemos ver que $\{X_t\}$ será Gaussiana. Esto será importante al momento de intentar encontrar los coeficientes de nuestros modelos.

3.2.1. Introducción a los modelos de autorregresión

Recordemos que la forma de un modelo de regresión lineal estaba dado por

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i,\ i = 1, \ldots, n,$$

con $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. ¿Qué tal si, en vez de considerar un conjunto de variables independientes y_i, corriéramos este modelo sobre nuestros mismos valores dependientes x_i? Esto nos permitiría considerar, incluso, valores de i que yacen más allá de los n datos que nosotros tenemos, pudiendo así *predecir* valores que no sabemos, ya sea porque no han ocurrido aún, o debido a que no conocemos esa información (e.g. está muy en el pasado). Ahí radica la utilidad de considerar esto. Para ello, hay que formalizar este pensamiento.

Definición 3.12 (Modelo autorregresivo)**.** Sea $\{X_t\}$ una serie de tiempo estacionaria o una realización de esta. Se le conoce como *modelo autorregresivo* o *modelo AR* de orden p, y se denota $\mathrm{AR}(p)$, a la relación dada por el sistema de ecuaciones

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t,$$

donde ε_t es ruido blanco y c una constante. A los φ_i se les llama *parámetros* del modelo.

De manera análoga a un modelo lineal, podemos expresar a nuestro sistema en forma matricial al considerar la esperanza. Por conveniencia tomemos a $c = 0$. Si tenemos N valores conocidos en nuestra serie $\{X_t\}$:

$$\begin{aligned}
X_t &= c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t, \\
\begin{pmatrix} X_p \\ X_{p+1} \\ \vdots \\ X_N \end{pmatrix} &= \begin{pmatrix} X_0 & X_1 & \cdots & X_{p-1} \\ X_1 & X_2 & \cdots & X_p \\ \vdots & \vdots & \ddots & \vdots \\ X_{N-p} & X_{N-p+1} & \cdots & X_{N-1} \end{pmatrix} \begin{pmatrix} \varphi_p \\ \varphi_{p-1} \\ \vdots \\ \varphi_1 \end{pmatrix} \\
Y &= X\varphi \\
X^T Y &= (X^T X)\varphi \\
\varphi &= (X^T X)^{-1} X^T Y,
\end{aligned}$$

siempre que la pseudoinversa exista.

Sin embargo, existe un método distinto para calcular los coeficientes de nuestro modelo la cuál no tenemos en uno lineal. Una vez más consideremos $c = 0$ por simplicidad. Si más aún suponemos

que nuestra serie es estacionaria, realizando una variedad de operaciones, podemos transformar nuestro sistema de ecuaciones a lo siguiente.

$$\begin{aligned}
X_t &= c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \\
&= \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + \varepsilon_t \\
X_t X_{t-h} &= \varphi_1 X_{t-1} X_{t-h} + \varphi_2 X_{t-2} X_{t-h} + \cdots + \varphi_p X_{t-p} X_{t-h} + \varepsilon_t X_{t-h} \\
\mathrm{E}[X_t X_{t-h}] &= \mathrm{E}[\varphi_1 X_{t-1} X_{t-h} + \varphi_2 X_{t-2} X_{t-h} + \cdots + \varphi_p X_{t-p} X_{t-h} + \varepsilon_t X_{t-h}] \\
&= \varphi_1 \,\mathrm{E}[X_{t-1} X_{t-h}] + \varphi_2 \,\mathrm{E}[X_{t-2} X_{t-h}] + \cdots + \varphi_p \,\mathrm{E}[X_{t-p} X_{t-h}] + \mathrm{E}[\varepsilon_t X_{t-h}] \\
\gamma_X(h) &= \varphi_1 \gamma_X(h-1) + \varphi_2 \gamma_X(h-2) + \cdots + \varphi_p \gamma_X(h-p) \\
\rho_X(h) &= \varphi_1 \rho_X(h-1) + \varphi_2 \rho_X(h-2) + \cdots + \varphi_p \rho_X(h-p).
\end{aligned}$$

La suposición de estacionariedad es necesaria para pasar de las esperanzas a las autocovarianzas. Dividiendo entre $\gamma_X(0)$ se obtiene la última ecuación. Usando el hecho que ρ_X es par (proposición 3.8) y $\rho_X(0) = 1$, para $h = 1, \ldots, p$, obtenemos el siguiente sistema de ecuaciones.

Definición 3.13 (Ecuaciones de Yule-Walker)**.** Sea

$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

un modelo $\mathrm{AR}(p)$. A las ecuaciones

$$\begin{aligned}
\rho_X(1) &= \varphi_1 + \varphi_2 \rho_X(1) + \cdots + \varphi_p \rho_X(p-1) \\
\rho_X(2) &= \varphi_1 \rho_X(1) + \varphi_2 + \cdots + \varphi_p \rho_X(p-2) \\
&\vdots \\
\rho_X(p) &= \varphi_1 \rho_X(p-1) + \varphi_2 \rho_X(p-2) + \cdots + \varphi_p
\end{aligned}$$

se les llaman *ecuaciones de Yule-Walker*.

Si reemplazamos los valores teóricos de ρ_X con sus estimaciones muestrales $\hat{\rho}_X$, podremos encontrar los coeficientes φ_i como la solución del sistema

$$\begin{pmatrix} 1 & \hat{\rho}(1) & \cdots & \hat{\rho}(p-1) \\ \hat{\rho}(1) & 1 & \cdots & \hat{\rho}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}(p-1) & \hat{\rho}(p-2) & \cdots & 1 \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_p \end{pmatrix} = \begin{pmatrix} \hat{\rho}(1) \\ \hat{\rho}(2) \\ \vdots \\ \hat{\rho}(p) \end{pmatrix}.$$

Recordemos que definimos a $\hat{\gamma}$ sin corrección de Bessel para asegurar que la matriz de autocovarianzas sea positiva semidefinida. De hecho, la matriz de autocorrelaciones muestrales $\hat{R}_p$ también es positiva semidefinida. Aunque no es una condición tan fuerte como ser positiva definida – lo que garantiza invertibilidad y, por lo tanto, solución a las ecuaciones de Yule-Walker – en la mayoría de los casos sí tendremos una solución.

3.2.2. Introducción a los modelos de media movible

En la sección anterior consideramos un modelo para nuestros datos utilizando los mismos para formar predicciones sobre datos futuros. Sin embargo, con cada predicción adelantada el error crece. En contraste, un modelo de media movible no considera hacer regresión sobre sí mismo, como veremos a continuación. En sí, un modelo de media movible es una generalización de otro concepto.

Definición 3.14 (Modelo lineal general)**.** Sea $\{X_t\}$ una serie de tiempo o una realización de esta. Se le conoce como *modelo lineal general* a la relación dada por el sistema de ecuaciones

$$X_t = \mu_X + \varepsilon_t + \sum_{i=1}^{\infty} \beta_i \varepsilon_{t-i},$$

donde ε_t es ruido blanco y μ_X la función de media de la serie. A los β_i se les llama *parámetros* del modelo.

Observemos que el modelo anterior es una serie, y no una combinación lineal de coeficientes y variables. De ahí viene el nombre *general*. Otra característica que diferencia a este modelo de los otros es que, en vez de considerar alguna variable dependiente conocida o a los términos del pasado de la serie, aquí se considera a un *error* totalmente aleatorio pasado. Entonces, lo que considera el modelo lineal general es cómo cambia una variable futura basándonos sólo en cómo se tiene el error en el pasado. Una última cosa que considerar es que, al ser infinita, debemos tener cuidado en que la serie converja. De otro modo, no habría un modelo ni nada pues todo tendería al infinito conforme aumenta el tiempo.
Vamos a considerar un modelo lineal general pero tal que los β_i, a partir de cierto i, son idénticamente cero.

Definición 3.15 (Modelo de media movible)**.** Sea $\{X_t\}$ una serie de tiempo o una realización de esta. Se le conoce como *modelo de media movible* o *modelo MA* de orden q, y se denota $\mathrm{MA}(q)$, a la relación dada por el sistema de ecuaciones

$$X_t = \mu_X + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i},$$

donde ε_t es ruido blanco y μ_X la función de media de la serie. A los θ_i se les llama *parámetros* del modelo.

De ahí el nombre media movible. Visualizamos a los valores futuros en una serie de tiempo de la forma que tengan una tendencia (la función de media), y cuya variable aleatoria de media cero es el ruido blanco no observable. Nuestro trabajo entonces es obtener los coeficientes θ_i. Desafortunadamente, como los ε_i no son observables (al ser ruido aleatorio iid), no podemos expresar a nuestro sistema matricialmente tal que lo podamos resolver con mínimos cuadrados, ni existe un método equivalente a las ecuaciones de Yule-Walker. Tendríamos que usar algún método iterativo que se ajuste a nuestras necesidades. También existen varios algoritmos con los cuales estimar los coeficientes, cuya idea principal es *invertir* un modelo $\mathrm{MA}(q)$ a uno $\mathrm{AR}(\infty)$, tal que ya podamos aplicar algún método más sencillo para encontrar a los θ_i. Un algoritmo será descrito más adelante.

3.2.3. Modelo ARMA y ARIMA

Ya conocemos los modelos de autorregresión y de media movible. Vamos a considerar un nuevo modelo que nos permita tener las ventajas de ambos.

Definición 3.16 (Modelo ARMA)**.** Sea $\{X_t\}$ una serie de tiempo estacionaria o una realización de esta. Se le conoce como *modelo autorregresivo de media movible* o *modelo ARMA* de orden (p, q), y se denota $\mathrm{ARMA}(p, q)$, a la relación dada por el sistema de ecuaciones

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i},$$

donde cada ε es ruido blanco y c una constante. A los φ_i y θ_i se les llama *parámetros* del modelo.

Un modelo ARMA pone en consideración la regresión sobre sí mismo, para ayudarnos a un pronóstico que no sólo tienda hacia la media; y a la regresión sobre los errores, que permiten precisar mejor nuestras predicción al contar con un factor no tan predecible. Claramente si $p = 0$ ó $q = 0$ entonces el modelo decae a un $\mathrm{AR}(p)$ o a un $\mathrm{MA}(q)$, respectivamente. Si $p = q = 0$, estamos asumiendo que nuestra serie puede ser explicada por una tendencia constante y ruido blanco de media 0, lo cual rara vez – si no es que nunca – va a suceder en la práctica.

Como se tiene un modelo de media movible, una vez más los coeficientes se tendrán que estimar por medios distintos a menos cuadrados o ecuaciones de Yule-Walker.

Un último método y algo más general involucra al concepto de diferenciar la serie, lo cuál consiste en sustraer valores pasados a la serie original. Con base a este concepto se cambia un poco el modelo ARMA, no en definición, sino en aplicación. Esto está formalizado en la siguiente definición.

Definición 3.17 (Modelo ARIMA)**.** Sea $\{X_t\}$ una serie de tiempo estacionaria o una realización de esta. Se le conoce como *modelo autorregresivo de media movible* o *modelo ARIMA* de orden (p, d, q), y se denota $\mathrm{ARIMA}(p, d, q)$, a la relación dada por el sistema de ecuaciones

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i},$$

tal que $\{X_t\}$ halla sido diferenciada d veces. Cada ε es ruido blanco y c una constante. A los φ_i y θ_i se les llama *parámetros* del modelo.

En el siguiente capítulo se justificará por qué diferenciar nuestra serie ayuda. Veamos, como nota final, que el modelo ARIMA no agrega ni quita coeficientes, así que una vez más tenemos que estimar los coeficientes con un método iterativo. Lo que sí cambia es la misma serie a la que se aplica el modelo, porque se está diferenciando. Una vez ajustado el modelo a la serie diferenciada, podemos aplicar el proceso inverso a esta para ajustar el modelo a la serie original.

3.3. Estimación de coeficientes en un modelo ARIMA

Afortunadamente, sólo necesitamos verificar cómo estimar los coeficientes de un modelo ARMA, debido a que el modelo ARIMA utiliza los mismos coeficientes, con lo único que cambia siendo la serie, que es diferenciada un número de veces. En este caso, basta aplicar la operación inversa a estas diferenciaciones y el modelo ajustado servirá. Recordemos que en los modelos de media movible no podíamos aplicar las técnicas de mínimos cuadrados para estimar los coeficientes, debido a que se trabaja con un error aleatorio en vez de una variable cuya esperanza podríamos conocer. Si nuestro propósito al ajustar el modelo a una serie de tiempo estacionaria $\{X_t\}$, tal que su media μ y FACV γ sean conocidas, y donde $t = 1, \ldots, n$, entonces buscamos la mejor combinación

lineal de $\{1, X_n, X_{n-1}, \ldots, X_1\}$, que prediga el valor X_{n+h}, $h > 0$ con el menor error cuadrado medio. Es decir, definimos al *mejor predictor lineal*

$$P_n X_{n+h} := a_0 + a_1 X_n \cdots + a_n X_1.$$

El algoritmo de innovaciones requiere más aún que $\{X_t\}$ tenga media cero, que $\mathrm{E}[|X_t|^2] < \infty$ – ambas condiciones las cumple una serie estacionaria on algunas modificaciones menores – y que $\mathrm{E}[X_i X_j] = \mathrm{Cov}(X_i, X_j)$. Supongamos que $\{X_t\}$ cumple las tres condiciones. Definamos a

$$\hat{X_n} := \begin{cases} 0, & \text{si } n = 1, \\ P_{n-1} X_n, & \text{si } n = 2, 3, \ldots \end{cases}$$

como el *mejor predictor lineal* a un paso, y denotemos su ECM como $v_n = \mathrm{E}\left[(X_{n+1} - P_n X_{n+1})^2\right]$. Ahora, introducimos las *innovaciones*, o errores de predicción a un paso, denotados U_n y definidos $U_n := X_n - \hat{X_n}$. Dados los n valores conocidos de nuestra serie, y recordando la definición del mejor predictor lineal, podemos expresar a las n innovaciones como un sistema de ecuaciones, pues

$$\begin{aligned} U_1 &= X_1 - \hat{X_1} \\ &= 1 \cdot X_1, \\ U_n &= X_n - \hat{X_n} \\ &= X_n - P_{n-1} X_n \\ &= -a_{n-1} X_1 - \cdots - a_1 X_{n-1} + X_n; \\ \boldsymbol{U}_n &= A_n \boldsymbol{X}_n \\ \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix} &= \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -a_1 & 1 & 0 & \cdots & 0 \\ -a_2 & -a_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_{n-1} & -a_{n-2} & -a_{n-3} & \cdots & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}. \end{aligned}$$

Claramente $\det A_n = 1$, así que esta es invertible, digamos

$$A_n^{-1} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \theta_{11} & 1 & 0 & \cdots & 0 \\ \theta_{22} & \theta_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \cdots & 1 \end{pmatrix}.$$

Recordando la definición de las innovaciones, expresamos a los predictores a un paso de forma vectorial como

$$\begin{aligned} \hat{\boldsymbol{X}_n} &= \boldsymbol{X}_n - \boldsymbol{U}_n \\ &= \left(A_n^{-1} - I_n\right) \boldsymbol{U}_n \\ &= \Theta_n \boldsymbol{U}_n \\ &= \Theta_n (\boldsymbol{X}_n - \hat{\boldsymbol{X}_n}), \end{aligned}$$

donde

$$\Theta_n = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ \theta_{11} & 0 & 0 & \cdots & 0 \\ \theta_{22} & \theta_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \cdots & 0 \end{pmatrix}.$$

Así, en notación de sumatoria, reescribimos a $\hat{X_{n+1}}$ como

$$\hat{X_{n+1}} := \begin{cases} 0, & \text{si } n = 0, \\ \sum_{i=1}^{n} \theta_{ni} \left(X_{n+1-i} - \hat{X_{n+1-i}} \right), & \text{si } n = 1, 2, \ldots \end{cases}$$

Con ello, finalmente, introducimos el algoritmo para encontrar a los θ_{ij}.

Teorema 3.18 (Algoritmo de innovaciones)**.** *Sea $\{X_t\}$ una serie de tiempo con media cero, tal que $\mathrm{Cov}(X_i, X_j) = \mathrm{E}[X_i X_j]$, y si $[\mathrm{Cov}(X_i, X_j)]_{i,j=1}^{n}$ es no singular, entonces los predictores $\hat{X_{n+1}}$, $n \geq 0$, estarán dados por*

$$\hat{X_{n+1}} := \begin{cases} 0, & \text{si } n = 0, \\ \sum_{i=1}^{n} \theta_{ni} \left(X_{n+1-i} - \hat{X_{n+1-i}} \right), & \text{si } n = 1, 2, \ldots, \end{cases}$$

y los errores cuadrados medios v_i y coeficientes θ_{ij} se calculan como

$$\begin{aligned} v_0 &= \mathrm{Cov}(X_1, X_1), \\ \theta_{n,n-j} &= v_j^{-1} \left[\mathrm{Cov}(X_{n+1}, X_{j+1}) - \sum_{i=0}^{j-1} \theta_{j,j-i} \theta_{n,n-i} v_j \right], \ j = 0, 1, \ldots, n-1, \\ v_n &= \mathrm{Cov}(X_{n+1}, X_{n+1}) - \sum_{i=0}^{n-1} \theta_{n,n-i}^2 v_j. \end{aligned}$$

Como v_0 tiene una derivación más fácil, podemos calcular θ_{11} a partir de v_0, y con éste calcular v_1, y así sucesivamente. La demostración usa álgebra lineal y análisis funcional. Previo a ver la demostración, se introducirán algunas definiciones y resultados relacionados a esos temas, sin demostración.[5]

Definición 3.19 (Conjunto generado cerrado)**.** Si $X = \{x_i : i \in I\}$[6] es un subconjunto de un espacio de Hilbert H, el *conjunto generado cerrado* $\overline{\mathrm{gen}}(X)$ se define como el subespacio más pequeño de H que contiene a todo x_i, $i \in I$.

Proposición 3.20. *Sea $X \subset H$ donde H es un espacio de Hilbert.*

1. $\overline{\mathrm{gen}}(X) = \overline{\mathrm{gen}(X)}$,
2. *Si $X = \{x_1, \ldots, x_n\}$ (i.e. X es finito), entonces $\overline{\mathrm{gen}}\{x_1, \ldots, x_n\}$ es el conjunto de combinaciones lineales de los elementos de X.*

Teorema 3.21 (Proyección)**.** *Si M es un subespacio cerrado de un espacio de Hilbert H, y dado $x \in H$, entonces*

[5] Obtenidas de [3], pp. 50–52.

[6] I es un conjunto de índices posiblemente no numerable.

1. *Existe un único $\hat{x} \in M$ tal que*
$$||x - \hat{x}|| = \inf_{y \in M} ||x - y||,$$
2. *$\hat{x} \in M$ y $||x - \hat{x}|| = \inf_{y \in M} ||x - y||$ sí y sólo si $\hat{x} \in M$ y $(x - \hat{x}) \in M^{\perp}$, donde $M^{\perp}$ es el complemento ortogonal de M ($x \in M^{\perp}$ sí y sólo si $||x, y|| = 0$ para toda $y \in M$).*

Corolario 3.22 (Mapa de proyección)**.** *Si M es un subespacio cerrado de un espacio de Hilbert H e I representa el mapa identidad en H, existe y es único el mapa de proyección $P_M : H \to M$ tal que $(I - P_M)(H) = M^{\perp}$. A P_M se le llama el mapa de proyección de H en M.*

Lema 3.23. *El mejor predictor de X, $\hat{X_n}$, es un mapa de proyección.*

Demostración del algoritmo de innovaciones 3.18 (Brockwell y Davis, 2009)[3]. Sea $\mathcal{X}_n = \overline{\text{gen}}\{X_1, \ldots, X_n\}$. Definamos al producto interior de dos variables aleatorias X, Y como la esperanza de su producto (ver la demostración de la cota de Cramér-Rao 2.18) $\langle X_i, X_j \rangle := \text{Cov}(X_i, X_j) = \text{E}[X_i X_j]$. El conjunto $\{X_1 - \hat{X_1}, \ldots, X_n - \hat{X_n}\}$ es ortogonal pues, para $i < j$, $X_i - \hat{X_i} \in \mathcal{X}_{j-1}$, y por otro lado $(X_j - \hat{X_j}) \perp \mathcal{X}_{j-1}$ por definición de $\hat{X_j}$. Entonces

$$\left\langle \hat{X_{n+1}}, X_{j+1} - \hat{X_{j+1}} \right\rangle = \left\langle \hat{X_{n+1}}, X_{j+1} \right\rangle - \left\langle \hat{X_{n+1}}, \hat{X_{j+1}} \right\rangle = \theta_{n,n-j} v_j.$$

Por lo ya establecido, $(X_{n+1} - \hat{X_{n+1}}) \perp (X_{j+1} - \hat{X_{j+1}})$, y así

$$\theta_{n,n-j} = v_j^{-1} \left\langle \hat{X_{n+1}}, X_{j+1} - \hat{X_{j+1}} \right\rangle.$$

Tomando a j en el mejor predictor a un paso, se sigue que

$$\begin{aligned} \theta_{n,n-j} &= v^{-1} j \left[\text{Cov}(X_{n+1}, X_{j+1}) - \sum_{i=0}^{j-1} \theta_{j,j-i} \left\langle \hat{X_{n+1}}, X_{i+1} - \hat{X_{i+1}} \right\rangle \right] \\ &= v^{-1} j \left[\text{Cov}(X_{n+1}, X_{j+1}) - \sum_{i=0}^{j-1} \theta_{j,j-i} \theta_{n,n-i} v_i \right], \end{aligned}$$

como queríamos para $\theta_{n,n-j}$. Finalmente, por la ortogonalidad, utilizando el teorema de proyección 3.21 y el lema 3.23, concluimos que

$$\begin{aligned} v_n &= \left\| X_{n+1} - \hat{X_{n+1}} \right\|^2 \\ &= ||X_{n+1}||^2 - \left\| \hat{X_{n+1}} \right\|^2 \\ &= \text{Cov}(X_{n+1}, X_{n+1}) - \sum_{i=0}^{n-1} \theta_{n,n-i}^2 v_j. \end{aligned}$$

Así, obtenemos todos las formas que requería el algoritmo. ∎

El cálculo manual de los coeficientes es claramente muy tedioso, por lo que existen implementaciones numéricas que fueron utilizadas aquí para ajustar el modelo.

3.4. Pruebas de hipótesis para verificar estacionariedad

En las secciones anteriores definimos cuándo una serie de tiempo es estacionaria. La importancia de este tipo de series radica en que la mayoría de las pruebas o procedimientos que se aplican para la predicción de datos (incluidos los utilizados aquí) requieren que una serie sea, por lo menos, débilmente estacionaria. Aquí se introducirán dos pruebas de hipótesis con las cuales podemos verificar si una serie de tiempo es estacionaria y, en caso de no ser así, un método con el cual podremos convertir una serie de tiempo que no es estacionaria a una que sí lo es.

Primero necesitamos introducir algunos conceptos nuevos.

Definición 3.24 (Polinomio característico)**.** Sea $\{X_t\}$ una serie de tiempo estacionaria o una realización de esta. Dado un modelo $\mathrm{AR}(p)$ ajustado a nuestra serie

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t,$$

el *polinomio característico* del modelo se define como

$$p(m) := m^p - \varphi_1 m^{p-1} - \varphi_2 m^{p-2} - \cdots - \varphi_p.$$

Vamos a investigar el polinomio característico y sus raíces. De particular interés son aquellas que yacen dentro de y en la frontera del círculo unitario $|m| \leq 1$, debido a que su presencia causa que los modelos se "atoren" después de cierto periodo, causando problemas con las predicciones y un mal ajuste del modelo. Es decir, el modelo no es *estacionario*.

Definición 3.25 (Raíz unitaria)**.** Sea $\{X_t\}$ una serie de tiempo estacionaria o una realización de esta. Consideremos el polinomio característico de un modelo $\mathrm{AR}(p)$ ajustado a nuestra serie

$$p(m) := m^p - \varphi_1 m^{p-1} - \varphi_2 m^{p-2} - \cdots - \varphi_p.$$

Decimos que el modelo tiene *raíz unitaria* si $m = 1$ es una raíz de p.

Con esto en mente, existen varias pruebas de hipótesis que nos permiten comprobar si existe o no raíz unitaria en nuestro modelo. Estos consideran la realización de una serie de tiempo y no el modelo al que ajustamos, lo que nos permite aplicarlas antes de ajustar un modelo. Dos de las más populares son las siguientes.

Definición 3.26 (Prueba ADF)**.** Sea $\{x_t\}$ una serie de tiempo. La *prueba de Dickey-Fuller aumentada*, comúnmente abreviada *prueba ADF*, es una prueba de hipótesis para verificar si $\{x_t\}$ es estacionaria. Se prueban las siguiente hipótesis:

- *Hipótesis nula* H_0: el polinomio característico de la serie de tiempo $\{x_t\}$ tiene raíz unitaria.
- *Hipótesis alternativa* H_1: la serie de tiempo $\{x_t\}$ es estacionaria.

Si $\{x_t\}$ es una realización de una sucesión de variables aleatorias $\{X_t\}$, la prueba ADF supone que podemos expresar a cada X_t como

$$(X_t - X_{t-1}) = \alpha + \beta t + (\varphi - 1)X_{t-1} + \varepsilon_t,$$

con $\alpha + \beta t$ la tendencia y ε_t ruido blanco. La hipótesis nula entonces es equivalente a que $\varphi = 1$.

Originalmente Dickey y Fuller introdujeron esta prueba en 1979[6]. Podemos ver la intuición tras esta prueba de la siguiente manera: si $\{x_t\}$ is estacionaria (o estacionaria con tendencia), entonces ésta tiende a volver a una media constante (o determinada por la tendencia). Por lo tanto, los valores grandes tienden a ser seguidos por valores pequeños, y viceversa. Así, esto será un predictivo significativo de cómo cambiará la serie en un próximo periodo, y tendrá un coeficiente negativo. Por otro lado, si la serie está diferenciada, los cambios positivos o negativos suceden con una probabilidad que no depende de la serie, si no del ruido blanco.

La segunda prueba tiene las hipótesis opuestas.

Definición 3.27 (Prueba KPSS)**.** Sea $\{x_t\}$ una serie de tiempo. La *prueba Kwiatowski-Phillips-Schmidt-Shin*, comúnmente abreviada prueba KPSS, es una prueba de hipótesis para verificar si $\{x_t\}$ es estacionaria. Se prueban las siguiente hipótesis:

- *Hipótesis nula* H_0: la serie de tiempo $\{x_t\}$ es estacionaria.
- *Hipótesis alternativa* H_1: el polinomio característico de la serie de tiempo $\{x_t\}$ tiene raíz unitaria.

La prueba KPSS se basa en las pruebas de score, también llamadas del multiplicador de Lagrange. Si $\{x_t\}$ es una realización de una sucesión de variables aleatorias $\{X_t\}$, la prueba KPSS supone que podemos expresar a cada X_t como

$$X_t = m_t + S_t + \varepsilon_t,$$

con m_t la tendencia, S_t un *camino aleatorio* y ε_t un error (usualmente ruido blanco). Entonces la prueba de hipótesis es una prueba de score donde la hipótesis nula es equivalente a que $\mathrm{Var}[S_t] = 0$.

La prueba KPSS fue introducida en 1991[11] por los autores que ahora le dan su nombre. Estas pruebas no son mutualmente exclusivas simplemente por tener las hipótesis en orden opuesto; es decir, puede que el valor p obtenido en ambas pruebas no nos de evidencia para rechazar ninguna de las hipótesis. Más aún, es buena práctica utilizar ambas, por esta misma razón. Existen muchas pruebas más, con distintos estadísticos. Sin embargo estas dos son adecuadas al tener hipótesis opuestas y aplicaciones en los modelos ARIMA.

Para finalizar esta sección vamos a formar el concepto de diferenciar una serie y por qué lo aplicamos. Ya lo mencionamos brevemente al introducir el modelo ARIMA, pero sin embargo, sólo dimos un breve descriptivo de lo que era y no entramos en detalles de su funcionamiento.

Definición 3.28 (Diferencia)**.** Sea $\{X_t\}$ una serie de tiempo. El operador de *diferencia* se denota como ∇ y se define como $\nabla X_t := X_t - X_{t-1}$.

Definición 3.29 (Diferenciación de una serie de tiempo)**.** Si $\{X_t\}$ es una serie de tiempo, una *diferenciación* de dicha serie consiste en una aplicación del operador de diferencia para formar una nueva serie $\{\nabla X_t\} := \{X_t - X_{t-1}\}$. Recursivamente, múltiples diferenciaciones de la misma serie de tiempo implican diferenciar la diferencia anterior, es decir, $\nabla^d X_t := \nabla(\nabla^{d-1} X_t)$.

Con diferenciar una serie de tiempo lo que queremos es darle la propiedad de *estacionariedad*. Estamos aplicando una transformación a una serie no estacionaria con el propósito de hacerla *estacionaria en el sentido de su media* (es decir, quitar alguna tendencia no constante), sin modificar el hecho de que su varianza o autocovarianza no son estacionarias.

Finalmente, podemos redefinir al modelo ARIMA al haber formalizado el concepto de diferenciación.

Definición 3.30 (Operador de retraso)**.** Sea $\{X_t\}$ una serie de tiempo. El operador de *retraso* se denota como L y se define como $LX_t := X_{t-1}$.

Veamos que, debido a que $\nabla X_t = X_t - X_{t-1}$, podemos reescribir esto mediante un abuso de notación como $\nabla X_t = X_t - LX_t = (1-L)X_t$. En general, podemos expresar a la d-ésima diferencia en términos del operador de retraso como $\nabla^d X_t := (1-L)^d X_t$.

Definición 3.31 (Modelo ARIMA)**.** Sea $\{X_t\}$ una serie de tiempo estacionaria o una realización de esta. Se le conoce como *modelo autorregresivo de media movible* o *modelo ARIMA* de orden (p,d,q), y se denota $\mathrm{ARIMA}(p,d,q)$, si la serie $(1-L)^d X_t = \nabla^d X_t$ está descrita por un modelo $\mathrm{ARMA}(p,q)$.

Si vamos todavía un paso adelante en nuestro abuso de notación y vemos a $(1-L)^d$ como un polinomio de grado d, podemos ver por qué hace sentido diferenciar a la serie d veces: porque tenemos una raíz unitaria de multiplicidad d. Al diferenciar, la estamos "factorizando" por así decirlo[7] del modelo para obtener una serie que sí es estacionaria y que por ende le podamos aplicar un modelo ARMA.

Un último comentario antes de pasar a cómo elegir el modelo óptimo. Si, al diferenciar una vez nuestra serie de tiempo, aplicamos nuestras pruebas de hipótesis ADF y KPSS y concluimos que podemos rechazar y no rechazar, respectivamente, la hipótesis nula (que implica no necesitamos diferenciar más), no es *necesario* seguir diferenciando, *pero* en ocasiones esto nos ayudará a ajustar un modelo aún mejor. Recordemos, no estamos aceptando que la serie sea estrictamente estacionaria, ni existe un 100 % de confianza a menos que hagamos infinitas operaciones. Las pruebas de hipótesis nos dan un indicador de la estacionariedad de nuestra serie, pero no son una panacea que mágicamente nos ayuda a usar un modelo ARMA o a lo máximo un modelo ARIMA con $d = 1$.

3.5. Eligiendo un modelo ARIMA

En esta última sección consideraremos cuál modelo $\mathrm{ARIMA}(p,d,q)$ es adecuado para cierta serie. Claro, como cada serie es distinta en su composición, no existe un polvo mágico que nos diga que valores de p, d o q usar, pero podemos ayudarnos de ciertos criterios. El primero que introducimos es una extención de la función ACF que definimos casi al principio. Recordemos que la ACVF está denotada por γ_X y su versión muestral está dada por $\hat{\gamma}$.

Definición 3.32 (Función de autocorrelación parcial)**.** Sea $\{X_t\}$ una serie de tiempo estacionaria o una realización de esta descrita por un modelo $\mathrm{ARMA}(p,q)$. La *función de autocorrelación parcial* de $\{X_t\}$, denotada $\alpha(h)$, se define como

$$\begin{aligned} \alpha(0) &:= 1, \\ \alpha(h) &:= \Phi_{hh},\ h \geq 1, \\ \text{donde } \Phi_{hh} \text{ es el último componente de } \Phi &:= \Gamma_h^{-1}\boldsymbol{\gamma}_h, \\ \Gamma_h &:= [\gamma_X(i-j)]_{i,j=1}^{h}, \\ \text{y } \boldsymbol{\gamma}_h &:= (\gamma_X(1),\ldots,\gamma_X(h))^T. \end{aligned}$$

Cuando $\{x_t\}$ es una realización del modelo de $\{X_t\}$, la *función de autocorrelación parcial muestral* de $\{x_t\}$ se denota $\hat{\alpha}$ y se define de manera análoga, cambiando las funciones exactas por sus versiones de la muestra.

[7] Bueno, siendo muy estrictos, no tiene sentido que un operador lineal sea una raíz, ¿verdad? Pero es una forma intuitiva de visualizar la raíz unitaria. Por eso lo incluyo.

Ya podemos introducir los dos primeros criterios para elegir los parámetros.

Proposición 3.33 (Elegir un valor de p)**.** *Un criterio útil para elegir el valor de p de un modelo* $\mathrm{ARMA}(p,q)$ *es graficando la* función de autocorrelación parcial*, ayudándonos de donde $\hat{\alpha}$ sea idénticamente cero.*

Demostración. En un modelo $\mathrm{AR}(p)$, tenemos que $\alpha(h)$ para $h > p$ es igual a cero pues la covarianza es idénticamente cercana a cero. Entonces, los valores de $\hat{\alpha}$ cercanos a cero indican covarianza muy baja, por lo que el modelo estimado ya tiene una cantidad buena de parámetros. ■

Proposición 3.34 (Elegir un valor de q)**.** *Un criterio útil para elegir el valor de q de un modelo* $\mathrm{ARMA}(p,q)$ *es graficando la* función de autocorrelación*, ayudándonos de donde $\hat{\rho}$ decaiga de manera rápida.*

Demostración. En un modelo $\mathrm{MA}(q)$, tenemos que $\rho(h)$ debe decaer pues los ruidos blancos no deberían estar correlacionados. Entonces, los valores de $\hat{\rho}$ donde la función decae más rápido nos indican una correlación baja, por lo que el modelo estimado ya tiene una cantidad buena de parámetros. ■

No es necesario establecer un criterio para d pues arriba ya se mencionaron dos pruebas de hipótesis que nos indican si la serie requiere diferenciación o no. Por ende, elegir d se apoya en dichas pruebas y en ver cuánta diferenciación da un mejor modelo – claro, sin excedernos. Más aún, estos dos criterios se basan sólo en la evidencia visual. Dos personas podrían tener opiniones distintas de cuántos parámetros ocuparía el modelo en base a lo que ellos vean en las gráficas, y su proceso de razonamiento. Un método más objetivo sería considerar minimizar el error que tenga el modelo. Una métrica que cuantifica esto es la siguiente.

Definición 3.35 (Criterio de información de Akaike)**.** Sea $\{X_t\}$ una serie de tiempo estacionaria o una realización de esta descrita por un modelo $\mathrm{ARIMA}(p,d,q)$[8]. El *criterio de información de Akaike* de nuestro modelo se define como

$$\mathrm{AIC} := 2(p+d+q-1) - 2\log L(\beta, S(\beta)/n),$$

donde L es la función de verosimilitud del modelo y

$$S(\beta) = \sum_{i=1}^{n} (X_j - \hat{X}_j)^2 / \mathrm{ECM}(\hat{X}_j),$$

con $\hat{X}_j$ el valor predicho de $\hat{X}_j$. (Esta predicción es función de β, parámetros de nuestro modelo).

Minimizar el criterio de información de Akaike es una de las maneras recomendadas en [2] para elegir un modelo ARIMA. Como podemos ver, dicho criterio también depende de los parámetros que tiene el modelo, y aunque no tan dominantes comparado con el logaritmo, siguen teniendo que ser tomados en cuenta. Esto muestra que el mejor modelo no es sólo el que ajuste mejor los datos o minimize un sólo valor de tantas métricas existentes, si no el que nos dé el mejor equilibrio entre calidad, tiempo y costo para calcular. Un modelo ARIMA con una cantidad exagerada de parámetros deja de ser práctico porque no se pueden hacer cálculos útiles sobre sus estadísticos, pues al depender de más valores a estimes, somos más propensos a errores de cálculo e incluso nos arriesgamos a no hallar solución (si el sistema no está totalmente determinado, por ejemplo).

[8] En general el criterio se define para cualquier modelo; de hecho, está fundado en la teoría de la información y no en las series de tiempo.

Por último, análogo al intervalo de confianza y su papel en estimar qué valores puede tomar un parámetro, se tiene su análogo (aproximado) para verificar dónde podría caer el valor verdadero de una predicción.

Definición 3.36 (Intervalo de predicción aproximado)**.** Sea $\{X_t\}$ una serie de tiempo estacionaria o una realización de esta descrita por un modelo $\mathrm{ARIMA}(p,d,q)$. Si $\hat{X_{n+m}}$ es un valor predicho (dentro o fuera del conjunto de datos), un intervalo de predicción del $100(1-\alpha)\,\%$ para $\hat{X_{n+m}}$ aproximadamente normal está dado por

$$\left(\hat{X_{n+m}} - z_{1-\alpha/2}\sqrt{\hat{\sigma}^2\sum_{i=0}^{m-1}\theta_i^2}, \hat{X_{n+m}} + z_{1-\alpha/2}\sqrt{\hat{\sigma}^2\sum_{i=0}^{m-1}\theta_i^2}\right),$$

con $\hat{\sigma}^2$ la varianza aproximada del modelo y θ_i los parámetros (incluyendo convirtiendo el modelo $\mathrm{AR}(p)$ a un modelo general lineal, de ser necesario).

Conforme la predicción se aleje, el término $\sum_{i=0}^{m-1}\theta_i^2$ aumenta, por lo que el intervalo de predicción se expande, y a pesar de tener la misma confianza, podría no ser muy útil.

En resumen, uno puede considerar distintas métricas y pruebas visuales o un poco más rigurosas para elegir (p,d,q), pero al final del día el usuario debe optar por el modelo que sea práctico, que se ajuste bien a los datos, y que no sacrifique rendimiento o tiempo. Y, al igual que realizar estimaciones, uno puede crear un intervalo para estas predicciones

4. Análisis de datos

4.1. Visualización de la serie de tiempo

En primer lugar se trabajó con la serie de tiempo de los casos diarios de COVID-19 en México dada por la base de datos del CONACyT[4]. Lo primero hecho fue graficar la serie (Figura 2).

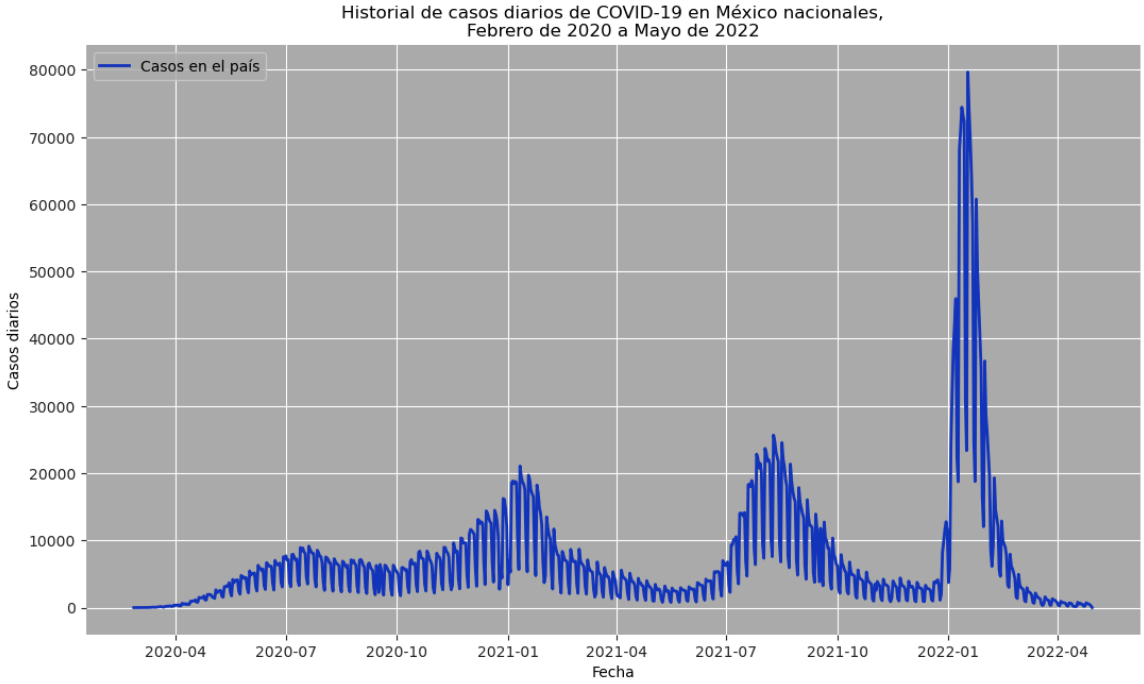

Figura 2: Serie de tiempo con los datos de COVID-19 en todo el país. Se cuenta desde el inicio de la pandemia el 26 de febrero de 2020 hasta el 30 de abril de 2022.

Lo primero que podemos notar son cuatro incrementos: el primero a partir de julio de 2020, correspondiente al periodo típico de las vacaciones de verano, seguido por otro repunte en enero del 2021, explicado por la variante Delta. De aquí sigue una caída significativa hasta el julio del 2021, donde podemos explicar el repunte con una combinación de factores, específicamente el periodo veraniego y la variante Lambda del virus. De esto sigue un periodo valle en casos hasta un pico masivo en enero y febrero del 2022, que corresponde a la variante Omicrón del SARS-CoV-2 y al periodo de invierno asociado con las festividades navideñas. A partir de marzo del 2022 se ve una caída en los casos, en particular siendo la menor cantidad diaria de contagios desde que llegó el virus a México. Esto lo podemos interpretar argumentado que la serie si tiene un componente de temporalidad de alrededor de 6 meses. La serie, sin embargo, no parece tener tendencia alguna, pues nunca hay un aumento o decremento constante o polinómico en los casos. En conclusión, nuestra serie $\{x_t\}$ es una realización de una sucesión de variables $\{X_t\}$ de la forma

$$X_t = s_t + Y_t.$$

4.2. Modelo ARIMA aplicado a marzo y abril del 2022

Para mostrar la viabilidad de aplicar los métodos de series de tiempo a nuestros datos, vamos a considerar un conjunto reducido de estos. Para ello, se tomaron datos de marzo y abril del 2022 (Figura 3). Claramente podemos ver que los casos van a la baja en este intervalo de tiempo.

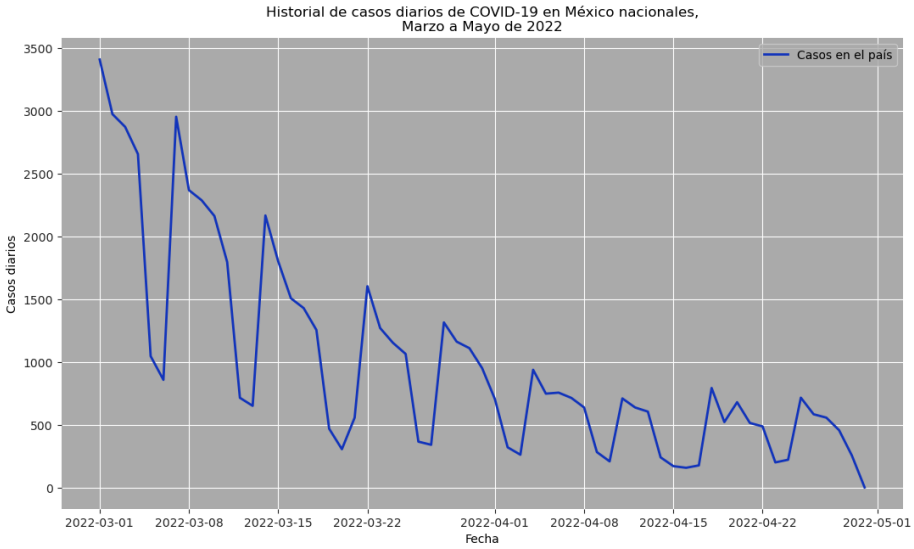

Figura 3: Serie de tiempo con los datos de COVID-19 en todo el país, restringiendo los datos a marzo y abril de 2022.

Ahora bien, vamos a ver si esta parte de la serie es estacionaria o si hay tendencia o temporada. Para ello, utilizaremos las pruebas ADF y KPSS introducidas en las proposiciones 3.26 y 3.27. Esperando un 95 % de confianza (i.e. $\alpha = 0.05$), obtenemos los siguientes valores de p al aplicar las pruebas de hipótesis.

```
Valor p de la prueba DFA:  5.250540088390928e-06
Valor p de la prueba KPSS: 0.1
```

Como $p < 0.05$ en la prueba DFA, podemos rechazar que la serie tiene una raíz unitaria, y como $p > 0.05$ en la prueba KPSS, no podemos rechazar que la serie sea estacionaria, con un 95 % de confianza. Por lo tanto, no necesitamos diferenciar nuestra serie.

Para decidir qué modelo ARIMA se ajusta mejor a la serie reducida, nos ayudaremos con las gráficas de las funciones de autocorrelación y autocorrelación parcial (Figuras 4 y 5).

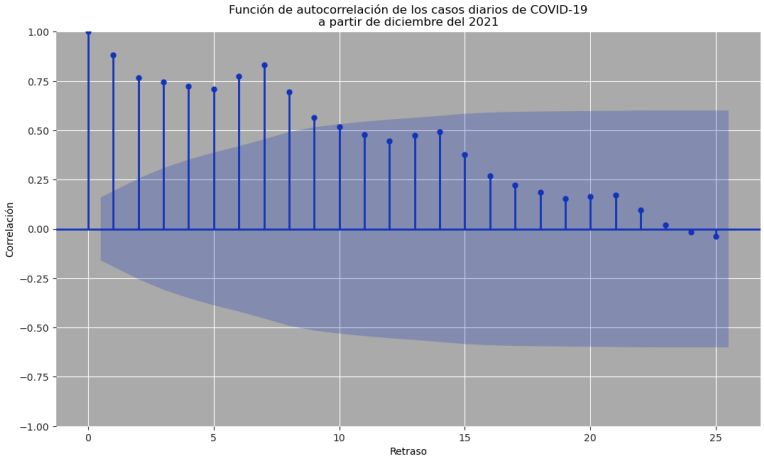

Figura 4: Función de autocorrelación de los datos de COVID-19 en el país, restringiendo los datos a marzo y abril de 2022.

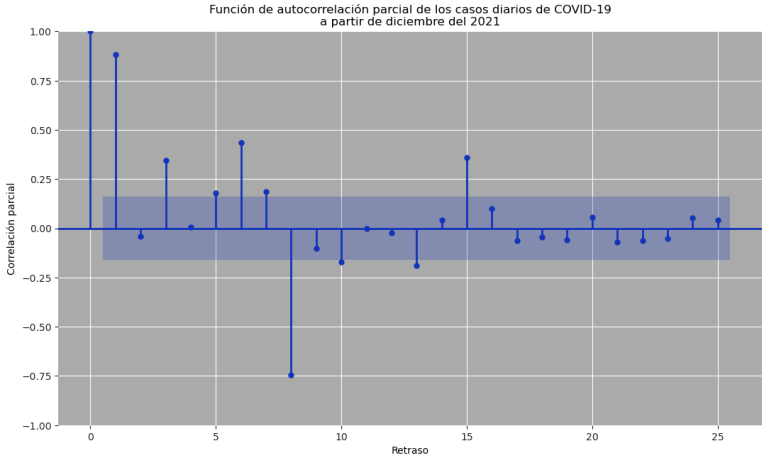

Figura 5: Función de autocorrelación parcial de los datos de COVID-19 en el país, restringiendo los datos a marzo y abril de 2022.

En este caso la evidencia visual nos podría sugerir un valor de $q = 9$, viendo que a partir del décimo retraso la función de correlación parcial se estabiliza alrededor de cero, pero no podemos inferir nada para p o d. Entonces buscaremos los parámetros (p, d, q) los cuales minimizarán el criterio de información de Akaike con el modelo ajustado a nuestros datos. Considerando un máximo de 10 parámetros – i.e. $0 \leq p, d, q \leq 10$ – y probando todas las combinaciones, encontramos que un modelo $\mathrm{ARIMA}(6, 3, 9)$ se ajustaba mejor a los datos. Con dicho modelo tenemos los siguientes estadísticos descriptivos.

SARIMAX Results

Dep. Variable:	Nacional	No. Observations:	61
Model:	ARIMA(6, 3, 9)	Log Likelihood	-327.486
Date:	Sat, 07 May 2022	AIC	686.972
Time:	17:57:22	BIC	716.912
Sample:	03-01-2022	HQIC	698.286
- 04-30-2022			
Covariance Type:	opg		

coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.7546	0.459	-1.643	0.100	-1.655	0.146
ar.L2	-0.7950	0.136	-5.864	0.000	-1.061	-0.529
ar.L3	-0.8529	0.360	-2.366	0.018	-1.559	-0.146
ar.L4	-0.7176	0.234	-3.063	0.002	-1.177	-0.258
ar.L5	-0.8117	0.225	-3.612	0.000	-1.252	-0.371
ar.L6	-0.5961	0.311	-1.916	0.055	-1.206	0.014
ma.L1	-2.2333	6.255	-0.357	0.721	-14.492	10.026
ma.L2	2.0134	11.665	0.173	0.863	-20.850	24.877
ma.L3	-1.8280	8.083	-0.226	0.821	-17.671	14.015
ma.L4	1.1723	6.975	0.168	0.867	-12.498	14.842
ma.L5	0.6461	6.866	0.094	0.925	-12.810	14.102
ma.L6	-1.9291	7.810	-0.247	0.805	-17.237	13.378

```
ma.L7          0.7921     11.868      0.067      0.947     -22.469      24.054
ma.L8          1.2155      5.122      0.237      0.812      -8.823      11.254
ma.L9         -0.8646      6.454     -0.134      0.893     -13.515      11.786
sigma2      1.558e+04   1.25e+05      0.125      0.901   -2.29e+05     2.6e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.07   Jarque-Bera (JB):                 1.72
Prob(Q):                              0.80   Prob(JB):                         0.42
Heteroskedasticity (H):               0.27   Skew:                             0.20
Prob(H) (two-sided):                  0.01   Kurtosis:                         3.83
===================================================================================
```

Con este modelo ajustado a la serie restringida, vamos a verificar su calidad comparando los datos reales de los casos que hubo en abril de 2022 contra los que el modelo esperaría (Figura 6).

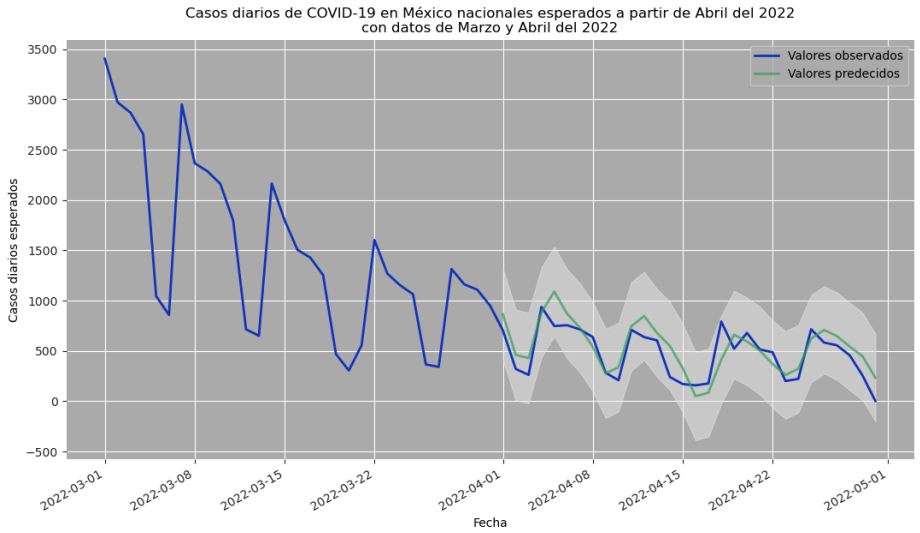

Figura 6: Serie de tiempo con los datos de COVID-19 nacionales de abril y marzo de 2022 (azul), junto con una predicción de los casos de abril a partir de los de marzo (verde) y un intervalo de predicción (gris claro).

Como podemos ver, los valores que predijo el modelo se asemejan bastante a los casos que verdaderamente hubo en México en abril de 2022. Teniendo eso en cuenta, vamos a verificar qué cantidad de casos son esperados por el modelo $\mathrm{ARIMA}(6,3,9)$ para el mes de mayo del 2022 (Figura 7). Podemos ver que el modelo espera que los casos aumenten un poco, lo cual es de esperarse pues los primeros datos de marzo muestran que los casos eran algo altos. Otra cosa más que podemos observar es el intervalo de predicción. Veamos que, conforme aumentan los días, el modelo crece en incertidumbre de dónde caerán los datos verdaderos. Esto lo podemos atribuir, primero a los casos un poco elevados de principios de marzo, y segundo, que usamos muy pocos valores en la serie de tiempo (tan sólo la información de 61 días). Así, aún teniendo un 95 % de confianza, no podemos ser muy específicos en dónde caerán los casos diarios.

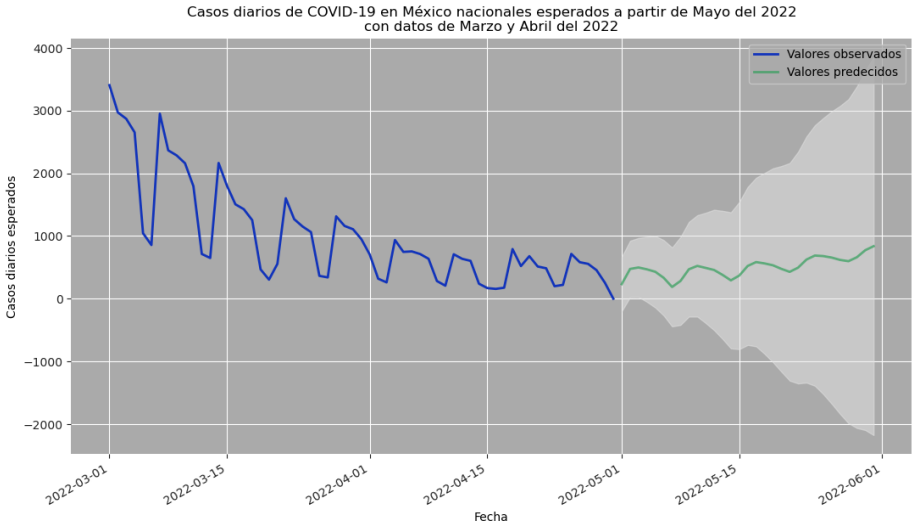

Figura 7: Serie de tiempo con los datos de COVID-19 nacionales de abril y marzo de 2022 (azul), junto con una predicción de los casos de mayo a partir de los de marzo y abril (verde) y un intervalo de predicción (gris claro).

Ahora vamos a verificar la validez de los resultados mediante una gráfica de residuos, así como supuestos de normalidad y las correlaciones. (Figura 8) La gráfica de residuos no muestra patrones ni tendencias. Además, el histograma y la gráfica Q-Q sugieren que los datos sí están normalmente distribuidos, y no has valores atípicos en el correlograma. Por lo tanto, nuestro modelo es válido.

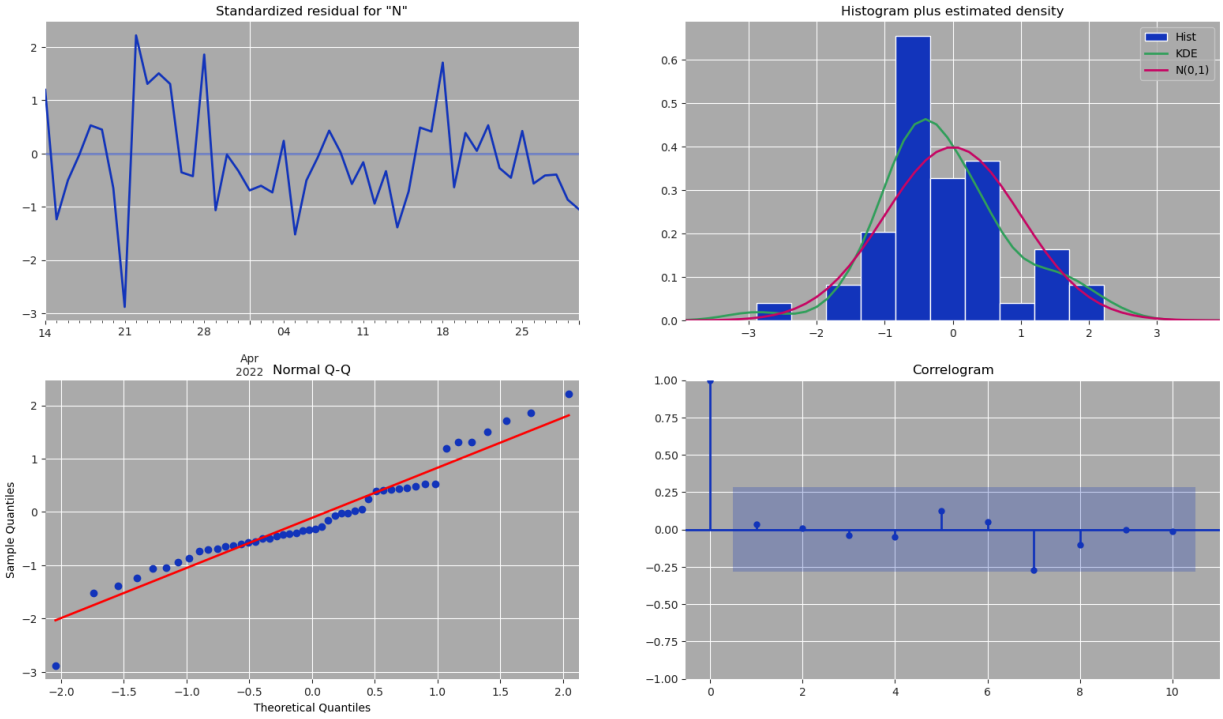

Figura 8: Gráfica de residuos, histograma, gráfica QQ y correlograma del modelo ajustado a la serie de tiempo restringida a marzo y abril de 2022.

4.3. Modelo ARIMA aplicado a partir de diciembre del 2021

Inspirados por el hecho de que el modelo aplicado a una cantidad pequeña de datos nos ayudaba un poco a describir cómo progresarían los casos, vamos hora a aplicar un modero ARIMA a una serie de tiempo que toma datos desde diciembre del 2021, que es cuando llegó la variante Omicrón del SARS-CoV-2 a México (Figura 9).

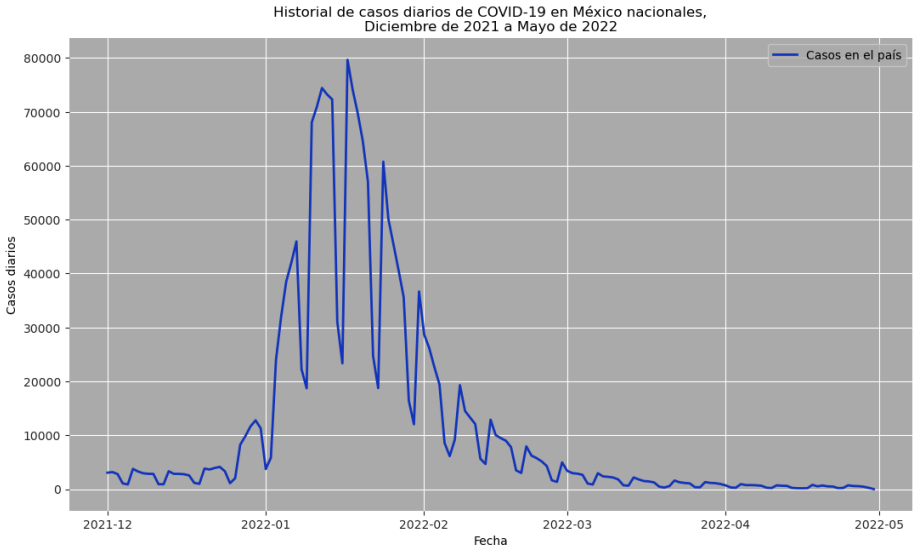

Figura 9: Serie de tiempo con los datos de COVID-19 en todo el país, restringiendo los datos a partir de diciembre del 2021.

De esta gráfica podemos inferir que los casos de COVID-19 iniciaron en aumento a partir de enero del 2022, coincidente con las celebraciones al final de año. Dicho eso, vamos a verificar si la serie es estacionaria.

```
Valor p de la prueba DFA:  0.4029280746681111
Valor p de la prueba KPSS: 0.043701024418397946
```

Como $p > 0.05$ en la prueba DFA, no podemos rechazar que la serie tiene una raíz unitaria, y como $p < 0.05$ en la prueba KPSS, rechazamos que la serie es estacionaria, con un 95 % de confianza. Por lo tanto, debemos de diferenciar nuestra serie (Figura 10). Aplicando estas pruebas de hipótesis una vez más, obtenemos los siguientes valores de p.

```
Valor p de la prueba DFA:  0.003367712376153881
Valor p de la prueba KPSS: 0.1
```

Análogo al caso de la serie restringida a marzo y abril de 2022, podemos concluir que esta serie es estacionaria. Entonces, vamos a encontrar un modelo ARIMA adecuado de manera similar al anterior. Primero verificamos las funciones de autocorrelación y autocorrelación parcial de la serie diferenciada (Figuras 11 y 12). En este caso la evidencia visual no nos ayuda a inferir una suposición para p, q o d. Buscaremos los parámetros (p, d, q) que minimizan el criterio de información de Akaike con el modelo ajustado a nuestros datos. Considerando un máximo de 10 parámetros – i.e. $0 \leq p, d, q \leq 10$ – y probando todas las combinaciones, resultó ser que un modelo $\mathrm{ARIMA}(6, 3, 9)$ es una vez más el mejor para nuestros datos. Con dicho modelo tenemos los siguientes estadísticos descriptivos.

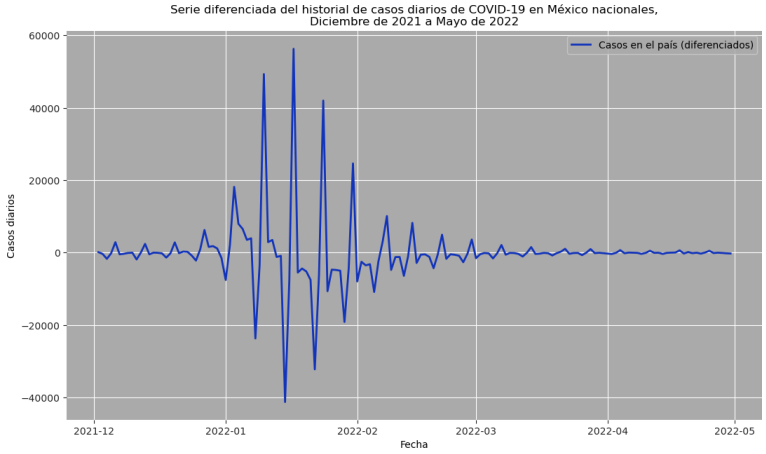

Figura 10: Serie de tiempo con los datos de COVID-19 en todo el país, restringiendo los datos a partir de diciembre del 2021, diferenciada.

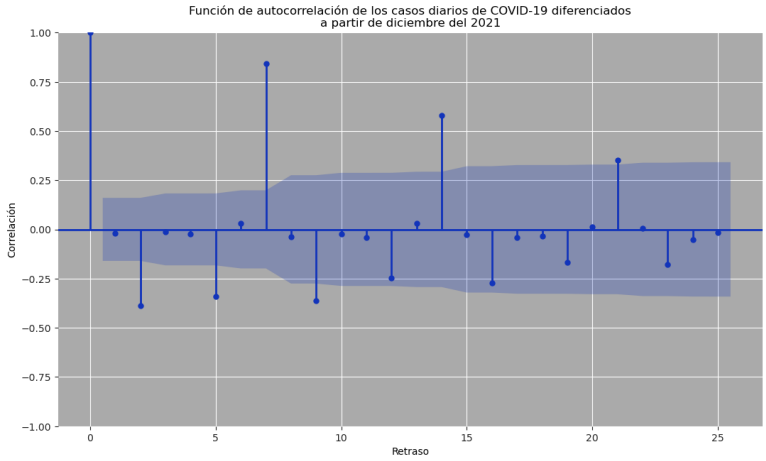

Figura 11: Función de autocorrelación de los datos de COVID-19 en el país, restringiendo los datos a partir de diciembre del 2021, diferenciados.

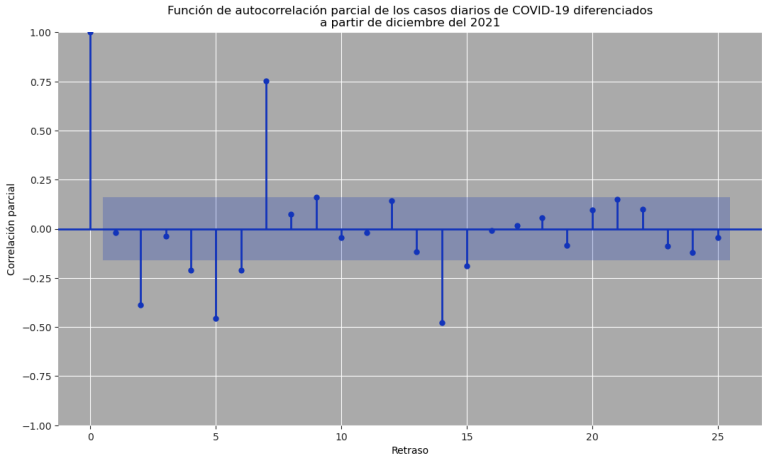

Figura 12: Función de autocorrelación parcial de los datos de COVID-19 en el país, restringiendo los datos a partir de diciembre del 2021, diferenciados.

```
==============================================================================
Dep. Variable:                Nacional   No. Observations:                  150
Model:                 ARIMA(6, 3, 9)   Log Likelihood               -1355.529
Date:                Sat, 07 May 2022   AIC                           2743.059
Time:                        23:02:47   BIC                           2789.778
Sample:                    12-02-2021   HQIC                          2762.044
- 04-30-2022
Covariance Type:                  opg
==============================================================================
coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.9520      0.095    -10.053      0.000      -1.138      -0.766
ar.L2         -1.0050      0.096    -10.480      0.000      -1.193      -0.817
ar.L3         -0.9435      0.114     -8.265      0.000      -1.167      -0.720
ar.L4         -0.9479      0.101     -9.365      0.000      -1.146      -0.750
ar.L5         -0.8924      0.080    -11.211      0.000      -1.048      -0.736
ar.L6         -0.8060      0.084     -9.560      0.000      -0.971      -0.641
ma.L1         -2.1641      0.159    -13.628      0.000      -2.475      -1.853
ma.L2          1.3112      0.375      3.497      0.000       0.576       2.046
ma.L3         -0.0153      0.554     -0.028      0.978      -1.101       1.071
ma.L4         -0.1310      0.630     -0.208      0.835      -1.365       1.103
ma.L5         -0.1325      0.563     -0.235      0.814      -1.236       0.971
ma.L6          0.3987      0.450      0.887      0.375      -0.483       1.280
ma.L7          0.0221      0.402      0.055      0.956      -0.765       0.809
ma.L8         -0.7802      0.332     -2.352      0.019      -1.430      -0.130
ma.L9          0.4933      0.128      3.869      0.000       0.243       0.743
sigma2      3.184e+07   1.64e-08   1.94e+15      0.000    3.18e+07    3.18e+07
===================================================================================
Ljung-Box (L1) (Q):                   0.73   Jarque-Bera (JB):               256.34
Prob(Q):                              0.39   Prob(JB):                         0.00
Heteroskedasticity (H):               0.01   Skew:                            -0.04
Prob(H) (two-sided):                  0.00   Kurtosis:                         9.70
===================================================================================
```

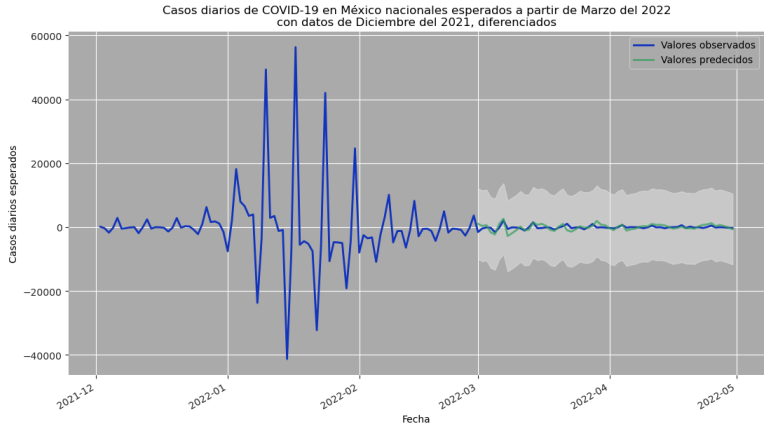

Figura 13: Serie de tiempo con los datos de COVID-19 nacionales de abril y marzo de 2022 (azul), junto con una predicción de los casos de abril a partir de los de marzo (verde) y un intervalo de predicción (gris claro).

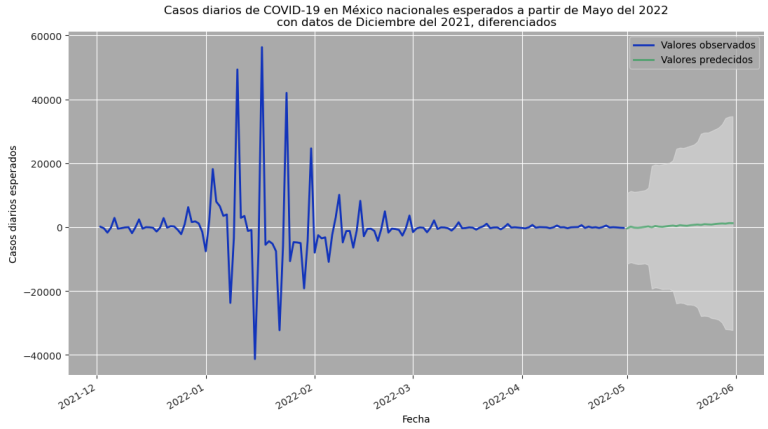

Figura 14: Serie de tiempo con los datos de COVID-19 nacionales de abril y marzo de 2022 (azul), junto con una predicción de los casos de abril a partir de los de marzo (verde) y un intervalo de predicción (gris claro).

Primero aplicamos el modelo a la serie diferenciada (Figuras 13 y 14) antes de invertir las transformaciones y verificar los resultados. Como diferenciamos antes de ajustar al modelo, debemos considerar en este una diferenciación más, es decir, ajustar con un $\mathrm{ARIMA}(6, 4, 9)$.

Una vez transformada, con este modelo ajustado a la serie restringida, vamos a verificar su calidad comparando los datos reales de los casos que hubo en abril de 2022 contra los que el modelo esperaría (Figura 15). De manera similar al ajuste que sólo tomaba a los datos de marzo y abril, los valores que predijo el modelo ajustado a nuestra serie extendida se asemejan bastante a los casos que verdaderamente hubo en México en marzo y abril de 2022. Teniendo eso en cuenta, vamos a verificar qué cantidad de casos son esperados por el modelo $\mathrm{ARIMA}(6, 4, 9)$ para mayo del 2022 (Figuras 16 y 17).

Guiándonos por la figura 17, el modelo espera que los casos diarios de COVID-19 en México tengan un repunte repentino en el mes de mayo. Este aumento se debe a que el modelo toma como referencia el pico de los casos de la variante Omicrón, cuando había muchas infecciones nuevas diarias, como referencia. Más aún, como observamos en la figura 16, el intervalo de predicción es muy extendido, debido a este mismo pico de casos, incluso obviando los valores negativos. Por lo tanto, a pesar de tener un $95\,\%$ de confianza, no hay mucha certidumbre en que los casos se

mantendrán a la alta ni a la baja.

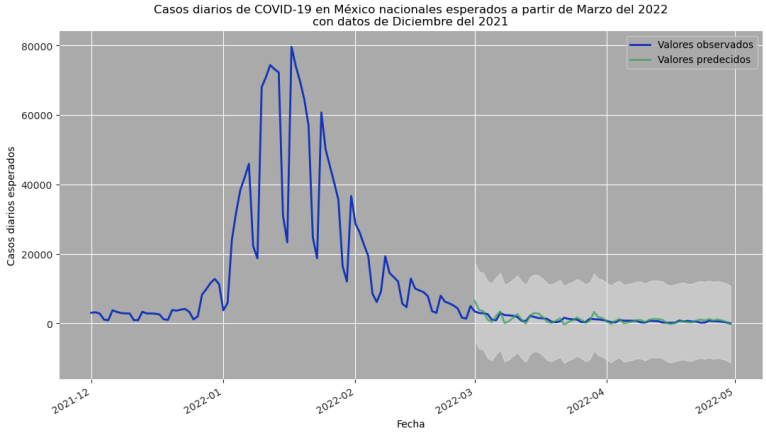

Figura 15: Serie de tiempo con los datos de COVID-19 nacionales de abril y marzo de 2022 (azul), junto con una predicción de los casos de abril a partir de los de marzo (verde) y un intervalo de predicción (gris claro).

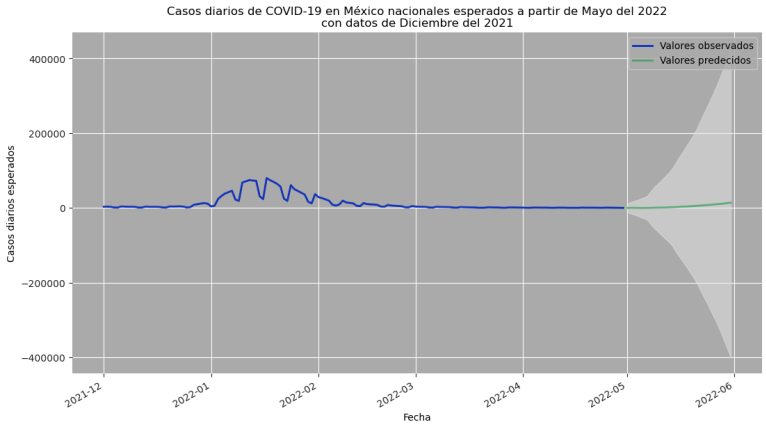

Figura 16: Serie de tiempo con los datos de COVID-19 nacionales de abril y marzo de 2022 (azul), junto con una predicción de los casos de abril a partir de los de marzo (verde) y un intervalo de predicción (gris claro).

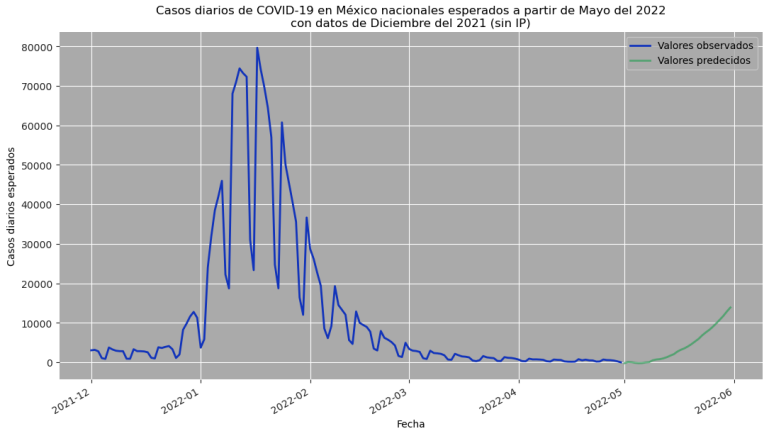

Figura 17: Serie de tiempo con los datos de COVID-19 nacionales de abril y marzo de 2022 (azul), junto con una predicción de los casos de abril a partir de los de marzo (verde).

Ahora vamos a verificar la validez de los resultados mediante una gráfica de residuos, así como supuestos de normalidad y las correlaciones (Figura 18).

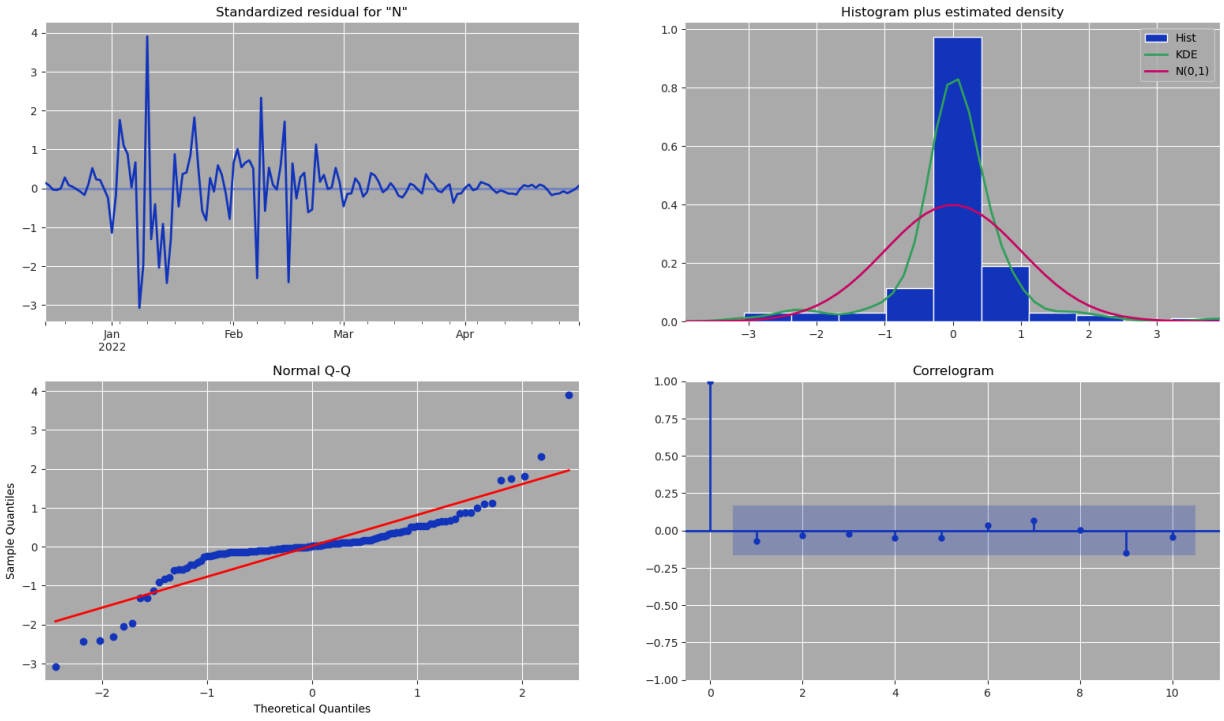

Figura 18: Gráfica de residuos, histograma, gráfica QQ y correlograma del modelo ajustado a la serie de tiempo restringida a diciembre de 2021.

La gráfica de residuos no muestra patrones ni tendencias. Además, el histograma y la gráfica Q-Q sugieren que los datos sí están normalmente distribuidos, y no has valores atípicos en el correlograma. Por lo tanto, nuestro modelo es válido. A pesar de ello, este modelo resultó no ser muy bueno.

4.4. Modelo ARIMA aplicado a toda la serie

Por último vamos a probar formando un modelo de predicción con toda la serie. Recordemos (Figura 2) que la serie de tiempo de los casos diarios totales parecía tener temporada. Aplicando una prueba DFA y KPSS, obtenemos lo siguiente.

```
Valor p de la prueba DFA:  8.502635297854753e-13
Valor p de la prueba KPSS: 0.1
```

Como $p < 0.05$ en la prueba DFA, podemos rechazar que la serie tiene una raíz unitaria, y como $p > 0.05$ en la prueba KPSS, no podemos rechazar que la serie sea estacionaria, con un 95 % de confianza. Por lo tanto, no necesitamos diferenciar nuestra serie. Sin embargo, para asegurarnos que la temporada no influya en el modelo, vamos a diferenciar una vez más (ver Figura 19), y aplicar las pruebas de hipótesis.

```
Valor p de la prueba DFA:  0.005291379924217426
Valor p de la prueba KPSS: 0.08240548911854784
```

Al igual que en el caso anterior, no podemos rechazar que la serie sea estacionaria con un 95 % de confianza.

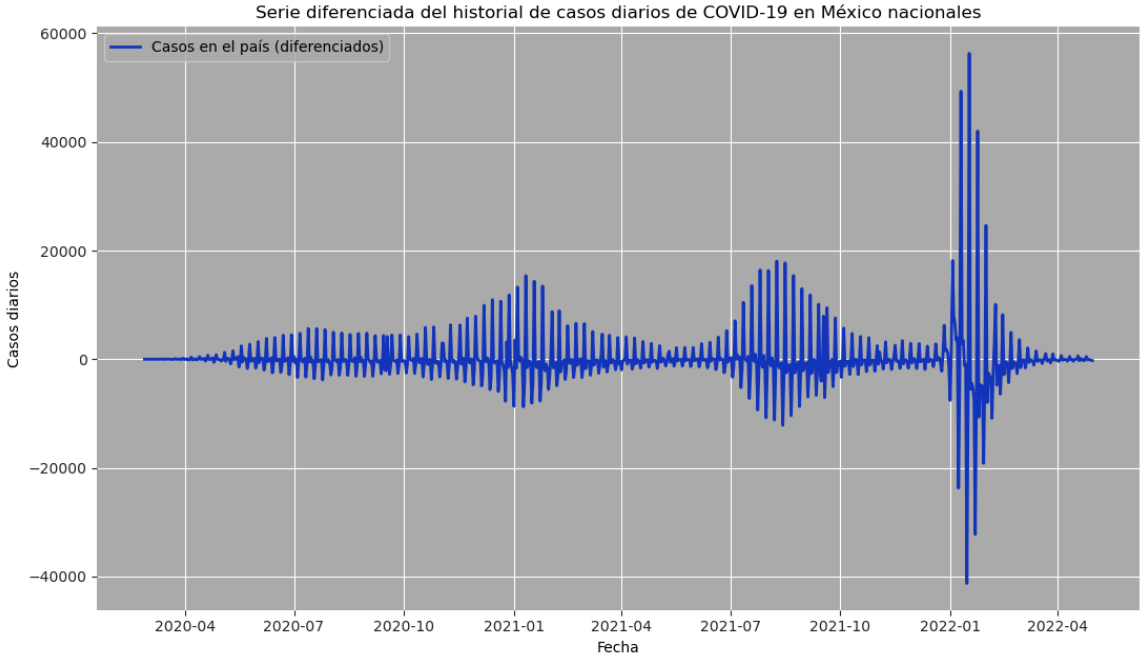

Figura 19: Serie de tiempo con los datos de COVID-19 en todo el país, desde el comienzo de la pandemia, diferenciada.

Entonces, vamos a encontrar un modelo ARIMA adecuado de manera similar al anterior. Primero verificamos las funciones de autocorrelación y autocorrelación parcial de la serie diferenciada (Figuras 20 y 21). En este caso la evidencia visual sugiere un valor alrededor de 10 para p y 8 para q. Buscaremos los parámetros (p, d, q) que minimizan el criterio de información de Akaike con el modelo ajustado a nuestros datos. Considerando un máximo de 10 parámetros – i.e. $0 \leq p, d, q \leq 10$ – y probando todas las combinaciones, resultó ser que un modelo $\text{ARIMA}(8, 1, 9)$ es una vez más el mejor para nuestros datos. Con dicho modelo tenemos los siguientes estadísticos descriptivos.

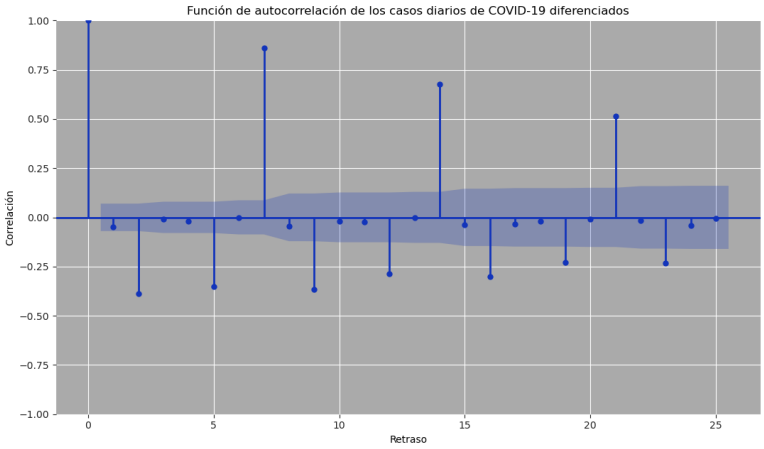

Figura 20: Función de autocorrelación de los datos de COVID-19 en todo el país, desde el comienzo de la pandemia, diferenciados.

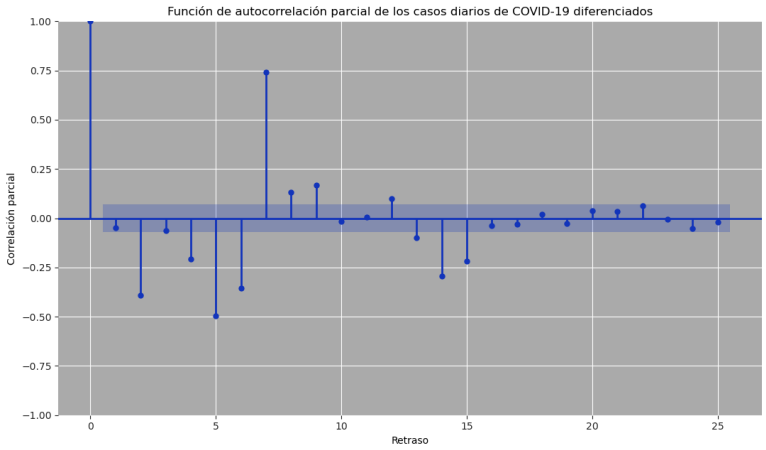

Figura 21: Función de autocorrelación parcial de los datos de COVID-19 en todo el país, desde el comienzo de la pandemia, diferenciados.

```
==============================================================================
Dep. Variable:                Nacional   No. Observations:                  795
Model:                 ARIMA(8, 1, 9)   Log Likelihood               -7173.685
Date:                Sun, 22 May 2022   AIC                          14383.370
Time:                        19:47:59   BIC                          14467.329
Sample:                    02-26-2020   HQIC                         14415.654
- 04-30-2022
Covariance Type:                  opg
==============================================================================
coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.8638      0.166     -5.194      0.000      -1.190      -0.538
ar.L2         -0.2469      0.059     -4.212      0.000      -0.362      -0.132
ar.L3         -0.2731      0.055     -4.993      0.000      -0.380      -0.166
ar.L4         -0.2352      0.048     -4.900      0.000      -0.329      -0.141
ar.L5         -0.2321      0.046     -5.089      0.000      -0.321      -0.143
ar.L6         -0.1912      0.042     -4.583      0.000      -0.273      -0.109
ar.L7          0.6206      0.040     15.567      0.000       0.542       0.699
ar.L8          0.5076      0.116      4.361      0.000       0.280       0.736
ma.L1          0.5368      0.168      3.200      0.001       0.208       0.866
ma.L2         -0.2593      0.050     -5.230      0.000      -0.356      -0.162
ma.L3          0.0981      0.037      2.658      0.008       0.026       0.171
ma.L4          0.1656      0.050      3.330      0.001       0.068       0.263
ma.L5          0.0415      0.037      1.108      0.268      -0.032       0.115
ma.L6          0.2226      0.047      4.769      0.000       0.131       0.314
ma.L7          0.4319      0.054      8.025      0.000       0.326       0.537
ma.L8          0.2263      0.070      3.239      0.001       0.089       0.363
ma.L9         -0.0689      0.043     -1.591      0.112      -0.154       0.016
sigma2      5.366e+06   2.23e-08    2.4e+14      0.000    5.37e+06    5.37e+06
==============================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):             26208.82
Prob(Q):                              0.98   Prob(JB):                         0.00
```

```
Heteroskedasticity (H):                12.96   Skew:                             1.06
Prob(H) (two-sided):                    0.00   Kurtosis:                        31.25
================================================================================
```

Primero aplicamos el modelo a la serie diferenciada (Figuras 22 a 24) antes de invertir las transformaciones y verificar los resultados. Como diferenciamos antes de ajustar al modelo, debemos considerar en este una diferenciación más, es decir, ajustar con un $\mathrm{ARIMA}(8, 2, 9)$.

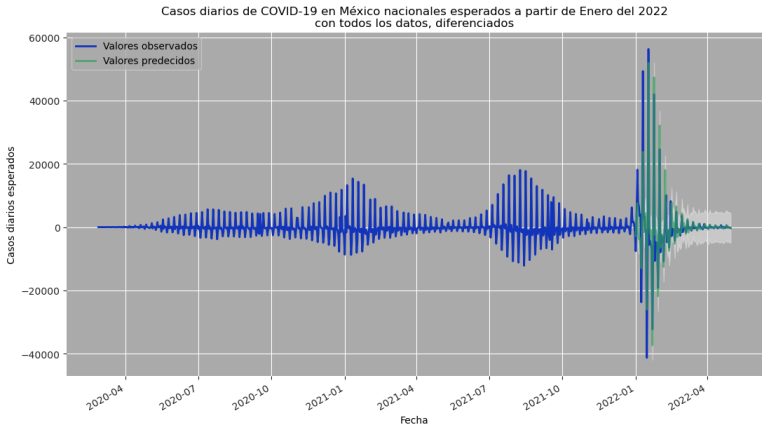

Figura 22: Serie de tiempo con los datos de COVID-19 diferenciados (azul), junto con una predicción de los casos de enero a partir de todos (verde) y un intervalo de predicción (gris claro).

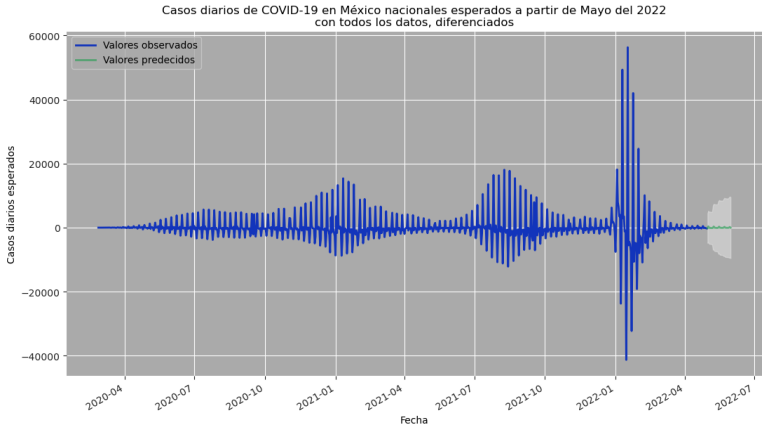

Figura 23: Serie de tiempo con los datos de COVID-19 diferenciados (azul), junto con una predicción de los casos de mayo a partir de todos (verde) y un intervalo de predicción (gris claro).

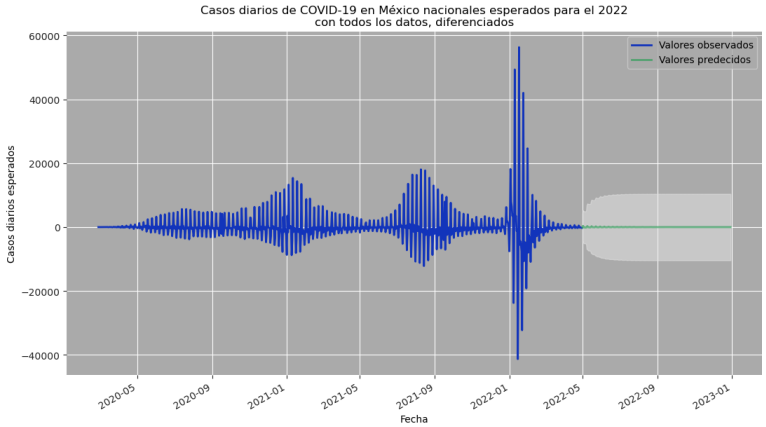

Figura 24: Serie de tiempo con los datos de COVID-19 diferenciados (azul), junto con una predicción de los casos de 2022 a partir de todos (verde) y un intervalo de predicción (gris claro).

Una vez transformada, con el modelo ajustado, vamos a verificar su calidad comparando los datos de los casos que hubo de enero a abril de 2022 con los que el modelo esperaría (Figuras 25 y 26).

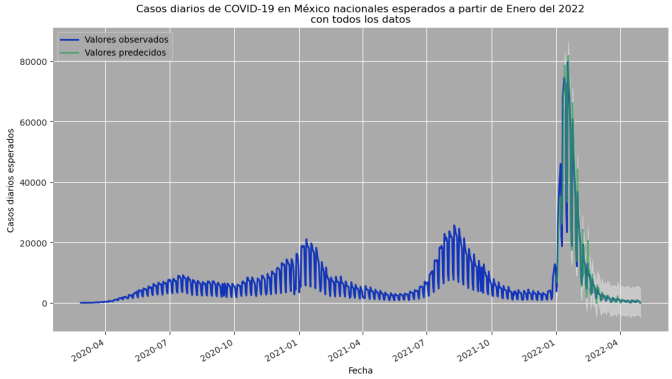

Figura 25: Serie de tiempo con los datos de COVID-19 (azul), junto con una predicción de los casos de enero a partir de todos (verde) y un intervalo de predicción (gris claro).

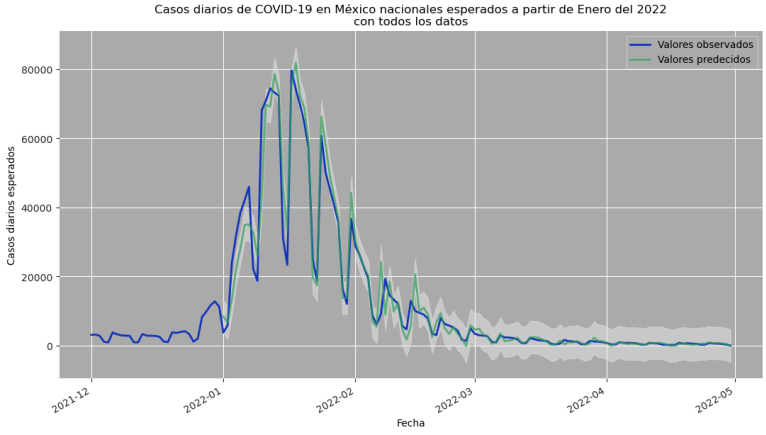

Figura 26: Serie de tiempo con los datos de COVID-19 (azul) visualizados desde diciembre del 2021, junto con una predicción de los casos de enero a partir de todos (verde) y un intervalo de predicción (gris claro).

El ajuste parece bastante bueno. Ahora vamos a comprobar qué cantidad de casos son esperados por el modelo $\mathrm{ARIMA}(8, 2, 9)$ para mayo del 2022 (Figuras 27 y 28).

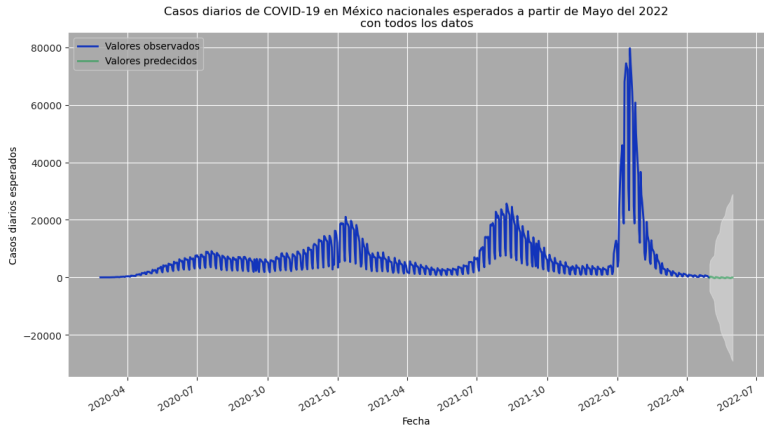

Figura 27: Serie de tiempo con los datos de COVID-19 (azul), junto con una predicción de los casos de mayo (verde) y un intervalo de predicción (gris claro).

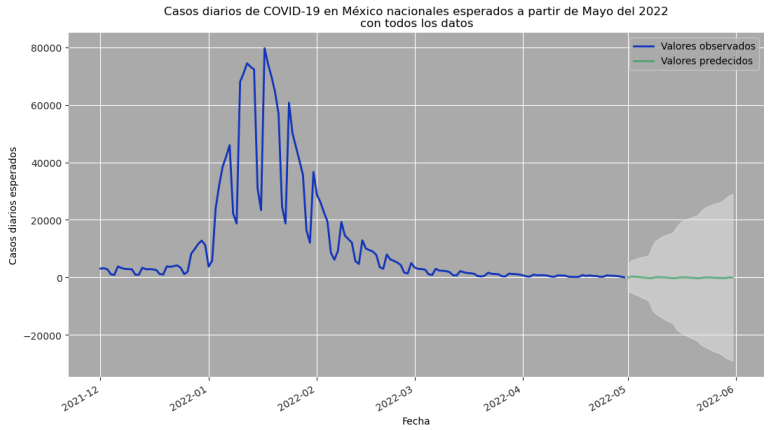

Figura 28: Serie de tiempo con los datos de COVID-19 (azul) visualizados desde diciembre del 2021, junto con una predicción de los casos de mayo (verde) y un intervalo de predicción (gris claro).

La figura 27 nos muestra la incertidumbre con el intervalo de predicción pues tiene una extensión muy grande, aunque en general, esperaría que los casos se mantuvieran a la baja. Esto lo apreciamos con más claridad en la figura 28. Finalmente, vamos a hacer una predicción de todo 2022 con la serie completa, aprovechando que conocemos muchos datos (Figuras 29 y 30). Podemos ver que la incertidumbre de hecho aumenta: el intervalo de predicción es enorme, incluso descartando los datos negativos que obviamente no tienen sentido. Aunque en general el modelo espera que los casos se estabilicen en cero, no puede decirlo con total certeza. Dicho esto, el valor predicho sí indica un decremento total. Eso se puede apreciar mejor en la Figura 31.

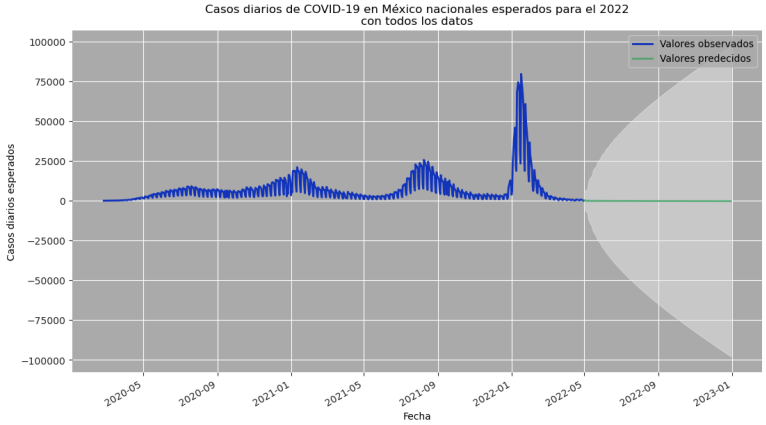

Figura 29: Serie de tiempo con los datos de COVID-19 (azul), junto con una predicción de los casos del 2022 (verde) y un intervalo de predicción (gris claro).

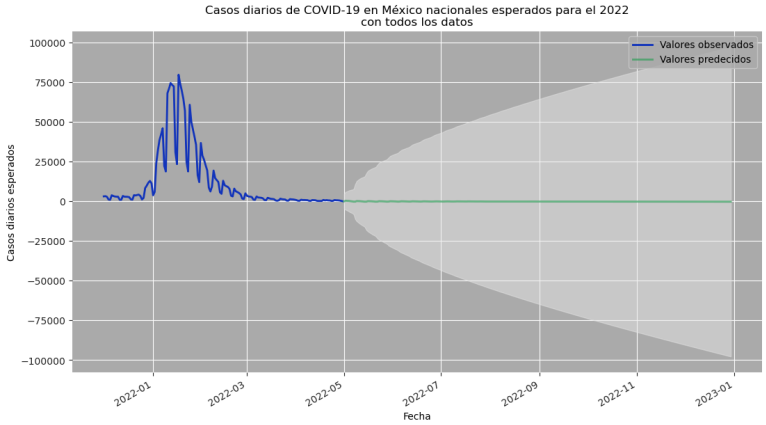

Figura 30: Serie de tiempo con los datos de COVID-19 (azul) visualizados desde diciembre del 2021, junto con una predicción de los casos del 2022 (verde) y un intervalo de predicción (gris claro).

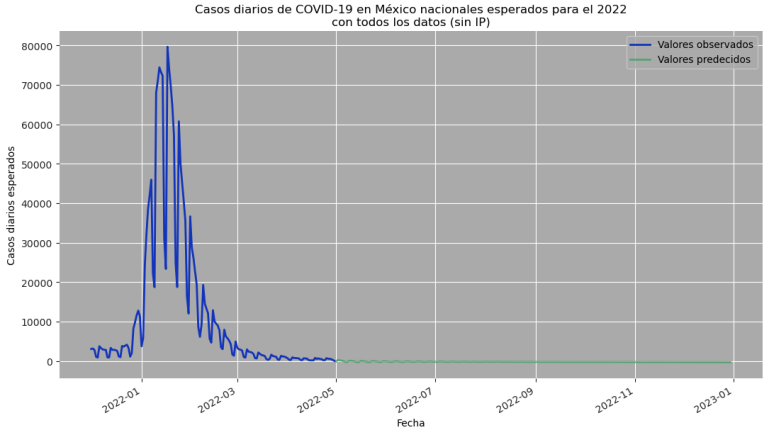

Figura 31: Serie de tiempo con los datos de COVID-19 (azul) visualizados desde diciembre del 2021, junto con una predicción de los casos del 2022 (verde).

Ahora vamos a verificar la validez de los resultados mediante una gráfica de residuos, así como supuestos de normalidad y las correlaciones (Figura 32).

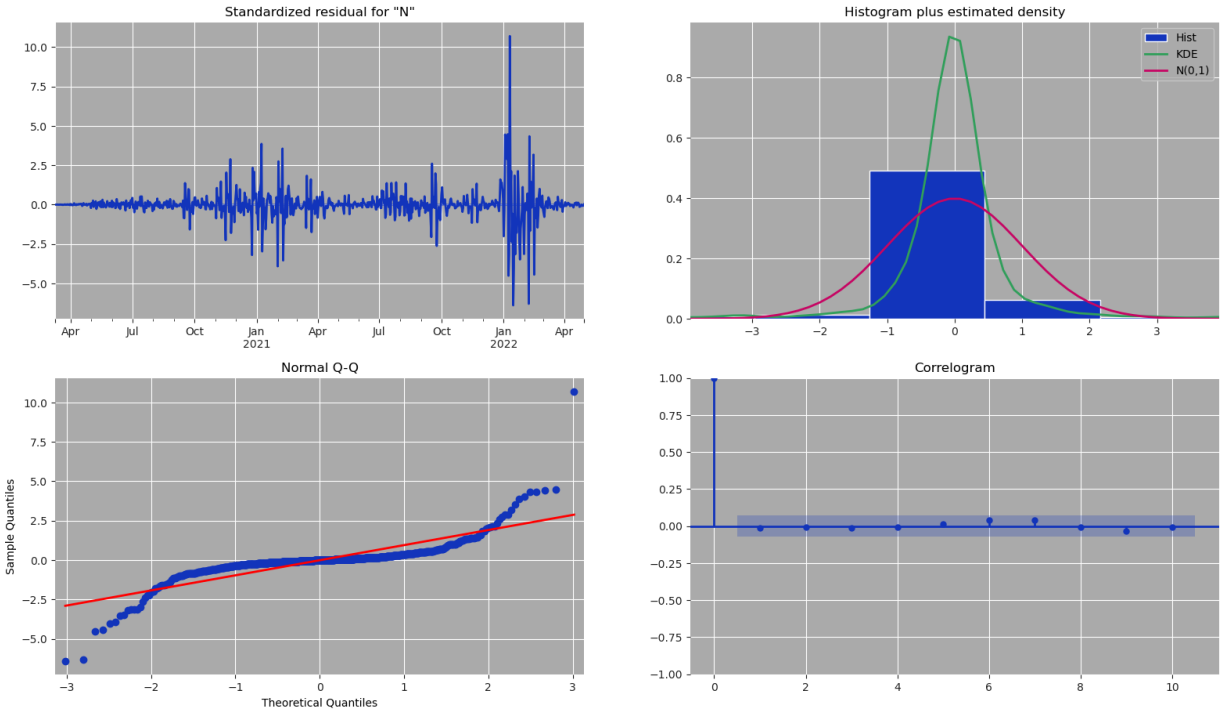

Figura 32: Gráfica de residuos, histograma, gráfica QQ y correlograma del modelo ajustado a la serie de tiempo completa.

La gráfica de residuos no muestra patrones ni tendencias. Además, el histograma y la gráfica Q-Q sugieren que los datos sí están normalmente distribuidos – aunque algunos valores son muy extremos –, y no hay valores atípicos en el correlograma. Por lo tanto, nuestro modelo es válido.

5. Conclusiones

Hemos visto cómo podemos ajustar un modelo a una situación tan variable, como lo es una pandemia de un virus con alta capacidad de mutación, por medio de las técnicas de serie de tiempo – en específico, utilizando un modelo de autorregresión y media movible. Por una parte, los resultados son prometedores; los modelos que ajustamos son muy válidos, y en el tema de cómo predicen valores que ya conocemos con los coeficientes que encuentra el modelo es muy parecida. Por otro lado, y desafortunadamente, la inferencia que podemos realizar sobre éstos datos no es muy útil pues existe mucha incertidumbre en los modelos que ajustamos, lo cual es apreciado en los intervalos de predicción. Más aún, los modelos que encontramos en general usaban muchos coeficientes. Aunque no es mucho problema considerando el poder computacional de hoy en día, también hay que buscar un balance entre precisión y rendimiento. Sin embargo, los resultados que dan los modelos son prometedores.

A final de cuentas, la calidad de predicción de diversos tipos de datos siempre será un problema el cuál se buscará mejorar: hemos conseguido un salto de calidad en los métodos básicos de regresión lineal para lograr describir datos más complejos. Pero también hay que considerar que los modelos ARIMA no son un elixir que mágicamente nos dirán como progresarán las cosas. Existen modelos más complejos, como autorregresión vectorial, generalizaciones de ARIMA que sí consideran temporada (llamados SARIMA), e incluso técnicas de aprendizaje de máquina que, aunque intrigantes, son muy complejas. También se tiene, en el campo de la epidemiología, métodos de modelado que son diseñados específicamente para considerar como podría esparcirse una enfermedad en la población. Hay que cerrar mencionando algo que a pesar de ser muy obvio podríamos olvidar en la búsqueda del análisis de datos, y es que la inferencia, modelado, y predicción, son procesos que se alteran en la subjetividad del estadístico. Pero ahí mismo radica la importancia de su estudio.

Apéndice A Código de Python

El siguiente código de Python fue desarrollado para toda la sección de análisis de datos, con ayuda de las librerías `NumPy`, `pandas`, `statsmodels` y `matplotlib`.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
import os.path
import statsmodels.tsa.ar_model as ar_model
import statsmodels.tsa.arima.model as arima_model
import itertools
from matplotlib import cycler

def graficar(serie,title,xax,yax,etiqueta,filename):
    """
       graficar(serie,titulo,xax,yax,filename): genera una gráfica
       Entrada: serie: datos a graficar
                titulo: título de la gráfica (opcional)
                xax: nombre del eje x (opcional)
                yax: nombre del eje y (opcional)
                filename: nombre para guardar la gráfica
    """
    if title is None:
        title=""
    if xax is None:
        xax="equis"
    if yax is None:
        yax="i griega"
    plt.figure(0,figsize=(12.8,7.2))
    plt.plot(serie,label=etiqueta)
    plt.title(title)
    plt.xlabel(xax)
    plt.ylabel(yax)
    plt.legend()
    plt.savefig(filename,bbox_inches="tight")
    plt.close("all")

def graficar_acf_pacf(serie,lag,acf_title,pacf_title,acf_filename,pacf_filename):
    fig,ax=plt.subplots(figsize=(12.8,7.2))
    sm.graphics.tsa.plot_acf(serie.to_numpy().squeeze(),lags=lag,fft=True,alpha=.05,
                             title=acf_title,ax=ax)
    plt.xlabel("Retraso")
    plt.ylabel("Correlación")
    plt.savefig(acf_filename,bbox_inches="tight")
    fig,ax=plt.subplots(figsize=(12.8,7.2))
    sm.graphics.tsa.plot_pacf(serie.to_numpy().squeeze(),lags=lag,alpha=.05,method="
    ywm",
                              title=pacf_title,ax=ax)
    plt.xlabel("Retraso")
    plt.ylabel("Correlación parcial")
    plt.savefig(pacf_filename,bbox_inches="tight")
    plt.close("all")

def graficar_diag(modelo,filename):
    plt.figure(0)
    modelo.plot_diagnostics(figsize=(19.2,10.8))
    plt.savefig(filename,bbox_inches="tight")
```

```
    plt.close("all")

def graficar_pred(pred,serie,title,xax,yax,filename,ci=True):
    if ci:
        pred_ci = pred.conf_int()
    plt.figure(0,figsize=(12.8,7.2))
    plt.plot(serie,label="Valores observados")
    pred.predicted_mean.plot(label="Valores predecidos", alpha=.7)
    if ci:
        plt.fill_between(pred_ci.index,
                         pred_ci.iloc[:, 0],
                         pred_ci.iloc[:, 1], color="#E6E6E6", alpha=.5)
    plt.xlabel(xax)
    plt.ylabel(yax)
    plt.title(title)
    plt.legend()
    plt.savefig(filename,bbox_inches="tight")
    plt.close("all")

def prueba_adf_kpss(serie):
    test1=sm.tsa.stattools.adfuller(serie)
    test2=sm.tsa.stattools.kpss(serie)
    return test1[1],test2[1]

def min_aic(serie,pdq,aic):
    j=0
    for i in pdq:
        try:
            res=arima_model.ARIMA(serie,order=i,enforce_stationarity=False,
    enforce_invertibility=False).fit()
            print(res.summary())
            aic[j]=res.aic
            print(aic[j])
            j+=1
        except:
            continue
    return pdq[np.argmin(aic)]

if __name__=="__main__":
    alpha=0.05

    # esto se encarga de inicializar las gráficas
    plt.close("all")
    colors = cycler("color",["#1234BB","#319F5B","#C70664",
                             "#C40F0F","#B74B03","#92C101",
                             "#620CC7","#2EA183","#36ACB2"])
    plt.rc("axes", facecolor='#AAAAAA', edgecolor="none",
            axisbelow=True, grid=True, prop_cycle=colors)
    plt.rc("grid", color="w", linestyle="solid")
    plt.rc("xtick", direction="out", color="#222222")
    plt.rc("ytick", direction="out", color="#222222")
    plt.rc("patch", edgecolor="#AAAAAA")
    plt.rc("lines", linewidth=2)

    # lee los casos de los datos diarios y los grafica
    diarios=pd.read_csv("C:/users/esele/Desktop/datos_seminario/diarios.csv",
    low_memory=False,index_col='FECHA',parse_dates=True,infer_datetime_format=True)
    diarios=diarios.loc[:,"Nacional"]
    graficar(diarios,
```

```
            "Historial de casos diarios de COVID-19 en México nacionales,\nFebrero de 2020 a Mayo de 2022",
            "Fecha",
            "Casos diarios",
            "Casos en el país",
            "casos_diarios_total")

    # grafica los datos pero sólo en un rango de fecha
    fecha="3/1/2022"
    diarios_off=diarios.loc[fecha:]
    graficar(diarios_off,
            "Historial de casos diarios de COVID-19 en México nacionales,\nMarzo a Mayo de 2022",
            "Fecha",
            "Casos diarios",
            "Casos en el país",
            "casos_marzo")
    graficar_acf_pacf(diarios_off,25,
                      "Función de autocorrelación de los casos diarios de COVID-19\na partir de marzo del 2022",
                      "Función de autocorrelación parcial de los casos diarios de COVID-19\na partir de marzo del 2022",
                      "acf_marzo","pacf_marzo")
    print("Valor p de la prueba DFA: ",prueba_adf_kpss(diarios_off)[0],"\nValor p de la prueba KPSS:",prueba_adf_kpss(diarios_off)[1])

    diarios_dif=(diarios_off-diarios_off.shift(1)).dropna()
    graficar(diarios_dif,
            "Serie diferenciada del historial de casos diarios de COVID-19 en México nacionales,\nMarzo a Mayo de 2022",
            "Fecha",
            "Casos diarios",
            "Casos en el país (diferenciados)",
            "casos_marzo_dif")
    graficar_acf_pacf(diarios_dif,25,
                      "Función de autocorrelación de los casos diarios de COVID-19 diferenciados\na partir de marzo del 2022",
                      "Función de autocorrelación parcial de los casos diarios de COVID-19 diferenciados\na partir de marzo del 2022",
                      "acf_marzo_dif","pacf_marzo_dif")
    print("Valor p de la prueba DFA: ",prueba_adf_kpss(diarios_dif)[0],"\nValor p de la prueba KPSS:",prueba_adf_kpss(diarios_dif)[1])

    # p=q=range(0,10)
    # pq=list(itertools.product(p,{0},q))
    # aic=np.zeros(100)
    # x=min_aic(diarios_off,pq,aic)
    # print(x)
    # (8,0,9)
    # res=arima_model.ARIMA(diarios_off,order=x,enforce_stationarity=False,enforce_invertibility=False).fit()
    res=arima_model.ARIMA(diarios_off,order=(6,3,9),enforce_stationarity=False,enforce_invertibility=False).fit()
    print(res.summary())
    graficar_diag(res,"modelo_marzo_diag")

    # genera una predicción fuera de muestra
    dias=31
    pred=res.get_forecast(steps=dias)
```

```
    graficar_pred(pred,diarios_off,
                  "Casos diarios de COVID-19 en México nacionales esperados a partir de Mayo del 2022\ncon datos de Marzo y Abril del 2022",
                  "Fecha",
                  "Casos diarios esperados",
                  "pred_mayo_marzo")

    #genera una predicción dentro de muestra
    pred=res.get_prediction(start="4/1/2022")
    graficar_pred(pred,diarios_off,
                  "Casos diarios de COVID-19 en México nacionales esperados a partir de Abril del 2022\ncon datos de Marzo y Abril del 2022",
                  "Fecha",
                  "Casos diarios esperados",
                  "pred_abril_marzo")

    # grafica los datos pero sólo en un rango de fecha
    fecha="12/1/2021"
    diarios_off=diarios.loc[fecha:]
    graficar(diarios_off,
             "Historial de casos diarios de COVID-19 en México nacionales,\nDiciembre de 2021 a Mayo de 2022",
             "Fecha",
             "Casos diarios",
             "Casos en el país",
             "casos_diciembre")
    graficar_acf_pacf(diarios_off,25,
                      "Función de autocorrelación de los casos diarios de COVID-19\na partir de diciembre del 2021",
                      "Función de autocorrelación parcial de los casos diarios de COVID-19\na partir de diciembre del 2021",
                      "acf_marzo","pacf_marzo")
    print("Valor p de la prueba DFA: ",prueba_adf_kpss(diarios_off)[0],"\nValor p de la prueba KPSS:",prueba_adf_kpss(diarios_off)[1])

    diarios_dif=(diarios_off-diarios_off.shift(1)).dropna()
    graficar(diarios_dif,
             "Serie diferenciada del historial de casos diarios de COVID-19 en México nacionales,\nDiciembre de 2021 a Mayo de 2022",
             "Fecha",
             "Casos diarios",
             "Casos en el país (diferenciados)",
             "casos_diciembre_dif")
    graficar_acf_pacf(diarios_dif,25,
                      "Función de autocorrelación de los casos diarios de COVID-19 diferenciados\na partir de diciembre del 2021",
                      "Función de autocorrelación parcial de los casos diarios de COVID-19 diferenciados\na partir de diciembre del 2021",
                      "acf_diciembre_dif","pacf_diciembre_dif")
    print("Valor p de la prueba DFA: ",prueba_adf_kpss(diarios_dif)[0],"\nValor p de la prueba KPSS: ",prueba_adf_kpss(diarios_dif)[1])

    # p=q=range(0,10)
    # pq=list(itertools.product(p,{0},q))
    # aic=np.zeros(100)
    # x=min_aic(diarios_dif,pq,aic)
    # print(x)
    # (6,3,9) (pero una diferenciación aparte)
```

```
207     # res=arima_model.ARIMA(diarios_dif,order=x,enforce_stationarity=False,
        enforce_invertibility=False).fit()
208     res=arima_model.ARIMA(diarios_dif,order=(6,3,9),enforce_stationarity=False,
        enforce_invertibility=False).fit()
209     print(res.summary())
210     graficar_diag(res,"modelo_diciembre_diag")
211
212     # genera una predicción fuera de muestra
213     dias=31
214     pred=res.get_forecast(steps=dias)
215     graficar_pred(pred,diarios_dif,
216                   "Casos diarios de COVID-19 en México nacionales esperados a partir
        de Mayo del 2022\ncon datos de Diciembre del 2021, diferenciados",
217                   "Fecha",
218                   "Casos diarios esperados",
219                   "pred_mayo_diciembre_dif")
220
221     #genera una predicción dentro de muestra
222     pred=res.get_prediction(start="3/1/2022")
223     graficar_pred(pred,diarios_dif,
224                   "Casos diarios de COVID-19 en México nacionales esperados a partir
        de Marzo del 2022\ncon datos de Diciembre del 2021, diferenciados",
225                   "Fecha",
226                   "Casos diarios esperados",
227                   "pred_marzo_diciembre_dif")
228
229     # total!!
230     graficar_acf_pacf(diarios,25,
231                       "Función de autocorrelación de los casos diarios de COVID-19",
232                       "Función de autocorrelación parcial de los casos diarios de
        COVID-19",
233                       "acf_total","pacf_total")
234     print("Valor p de la prueba DFA: ",prueba_adf_kpss(diarios)[0],"\nValor p de la
        prueba KPSS:",prueba_adf_kpss(diarios)[1])
235
236     diarios_dif=(diarios-diarios.shift(1)).dropna()
237     graficar(diarios_dif,
238              "Serie diferenciada del historial de casos diarios de COVID-19 en México
        nacionales",
239              "Fecha",
240              "Casos diarios",
241              "Casos en el país (diferenciados)",
242              "casos_dif")
243     graficar_acf_pacf(diarios_dif,25,
244                       "Función de autocorrelación de los casos diarios de COVID-19
        diferenciados",
245                       "Función de autocorrelación parcial de los casos diarios de
        COVID-19 diferenciados",
246                       "acf_total_dif","pacf_total_dif")
247     print("Valor p de la prueba DFA: ",prueba_adf_kpss(diarios_dif)[0],"\nValor p de
        la prueba KPSS: ",prueba_adf_kpss(diarios_dif)[1])
248
249     # p=d=q=range(0,10)
250     # pdq=list(itertools.product(p,d,q))
251     # aic=np.zeros(1000)
252     # x=min_aic(diarios_dif,pdq,aic)
253     # print(x)
254     # (8,1,9) (pero una diferenciación aparte)
```

```
255     # res=arima_model.ARIMA(diarios_dif,order=x,enforce_stationarity=False,
        enforce_invertibility=False).fit()
256     res=arima_model.ARIMA(diarios,order=(8,2,9),enforce_stationarity=False,
        enforce_invertibility=False).fit()
257     print(res.summary())
258     # graficar_diag(res,"modelo_total_diag")
259
260     # genera una predicción fuera de muestra
261     dias=31
262     pred=res.get_forecast(steps=dias)
263     graficar_pred(pred,diarios,
264                   "Casos diarios de COVID-19 en México nacionales esperados a partir
        de Mayo del 2022\ncon todos los datos (sin IP)",
265                   "Fecha",
266                   "Casos diarios esperados",
267                   "pred_mayo_total_noci",ci=False)
268     dias=244
269     pred=res.get_forecast(steps=dias)
270     graficar_pred(pred,diarios,
271                   "Casos diarios de COVID-19 en México nacionales esperados para el
        2022\ncon todos los datos (sin IP)",
272                   "Fecha",
273                   "Casos diarios esperados",
274                   "pred_22_total_noci",ci=False)
275
276     #genera una predicción dentro de muestra
277     pred=res.get_prediction(start="1/1/2022")
278     graficar_pred(pred,diarios,
279                   "Casos diarios de COVID-19 en México nacionales esperados a partir
        de Enero del 2022\ncon todos los datos (sin IP)",
280                   "Fecha",
281                   "Casos diarios esperados",
282                   "pred_enero_total_noci",ci=False)
```

Referencias bibliográficas

Bian, L., Gao, Q., Gao, F., Wang, Q., He, Q., Wu, X., Mao, Q., Xu, M., & Liang, Z. (2021). Impact of the Delta variant on vaccine efficacy and response strategies. *Expert review of vaccines*, *20*(10), 1201-1209.

Brockwell, P., & Davis, R. (2016). *Introduction to Time Series and Forecasting* (3ra ed.). Springer.

Brockwell, P. J., & Davis, R. A. (2009). *Time series: theory and methods* (2da ed.). Springer Science & Business Media.

CONACyT. (2022). Covid-19 México. https://datos.covid-19.conacyt.mx

Dehesh, T., Mardani-Fard, H., & Dehesh, P. (2020). Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models. *medRxiv*. https://doi.org/10.1101/2020.03.13.20035345

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, *74*(366), 427-431.

Fort, H. (2021). A very simple model to account for the rapid rise of the alpha variant of SARS-CoV-2 in several countries and the world. *Virus research*, *304*.

Gowrisankar, A., Priyanka, T., & Banerjee, S. (2022). Omicron: a mysterious variant of concern. *The European Physical Journal Plus*, *137*(1), 1-8.

Hogg, R., McKean, J., & Craig, A. (2019). *Introduction to Mathematical Statistics* (8va ed.). Pearson.

Kumar, N., & Susan, S. (2020). COVID-19 pandemic prediction using time series forecasting models. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1-7.

Kwiatkowski, D., Phillips, P., & Schmidt, P. (1991). *Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root?* (Inf. téc.). Cowles Foundation for Research in Economics, Yale University.

Malki, Z., Atlam, E.-S., Ewis, A., Dagnew, G., Alzighaibi, A. R., ELmarhomy, G., Elhosseini, M. A., Hassanien, A. E., & Gad, I. (2021). ARIMA models for predicting the end of COVID-19 pandemic and the risk of second rebound. *Neural Computing and Applications*, *33*(7), 2929-2948.

Ng, J. J. L., & Serrano, C. (2020). México| Covid-19 semana 16, proyecciones SIR b (t), ARIMA y Comparativo Internacional.

Tandon, H., Ranjan, P., Chakraborty, T., & Suhag, V. (2020). Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future. *arXiv preprint arXiv:2004.07859*.

Zumaya, A. R. A. (2021). La expansión de Covid-19 en México en 2020: un enfoque desde la econometría de series de tiempo. *Sobre México Temas de Economía*, (3), 34-66.