IMPUTACIÓN DE DATOS FALTANTES DE UNA SERIE DE TIEMPO BASADOS EN LOS VALORES CONOCIDOS DE LA PREDICCIÓN, Y SU APLICACIÓN A DATOS DE INFECCIÓN RESPIRATORIA AGUDA EN BOGOTÁ

Diego Johann Reyes Rojas

Trabajo de Grado para Optar al Título de Especialista en Estadística

Director

Andrés Sebastián Ríos Gutiérrez

Candidato a doctor

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Matemáticas

Especialización en Estadística

Bucaramanga

2024

Dedicado a Daniela

Agradecimientos

Dedico este trabajo principalmente a mis compañeros de la especialización, a Dumar, Laura, Laura Juliana, Jorge y Alejandra, quienes sin su apoyo y compañía en las largas jornadas de clase que tuvimos hubiera sido imposible darle continuidad al programa. También deseo darle especial agradecimiento a mi director, el profesor Andrés Sebastián Ríos por su amplio conocimiento sobre series temporales y procesos estocásticos, pues la teoría y los códigos estudiados en sus clases fueron ampliamente usados en este trabajo. Deseo dar un especial agradecimiento a los demás profesores de la especialización, en particular al profesor Oscar Yesid Castrillón que ahora se encuentra radicado en Puerto Rico y al profesor Henry Sebastián Rangel por la excelente labor docente que tuvieron y por hacer que mi gusto y curiosidad por la estadística se acrecentara aún más. Finalmente, pero no menos importante, darle las gracias a la familia quienes siempre han estado presentes desde el amor y la incondicionalidad. A Dios Padre, gracias.

Tabla de contenido

Lista de Tablas

Lista de Figuras

Lista de Imágenes

Lista de Apéndices

Resumen

Título: Imputación de datos faltantes de una serie de tiempo basados en los valores conocidos de la predicción, y su aplicación a datos de infección respiratoria aguda en Bogotá[*]

Autor: Diego Johann Reyes Rojas[**]

Palabras Clave: Imputación de datos faltantes, Series de tiempo estacionales, Modelos ARIMA multiplicativos, Estimados MSE

Descripción: Objetivos: Imputar datos faltantes de los años 2020 y 2021 del histórico de los casos de morbilidad por infección respiratoria aguda en Bogotá utilizando series de tiempo estacionales. Metodología: Usar la prueba de Dickey-Fuller aumentada para estudiar la estacionariedad de la serie temporal, luego diferenciar la serie a fin de volverla estacionaria, escoger 5 modelos ARIMA multiplicativos usando el criterio de Akaike y usar el estimador MSE para escoger cuál de estos se ajusta mejor a los datos disponibles del año 2022. Conclusiones: Los datos faltantes de la variable `infectados` de los años 2020 y 2021 pudieron ser imputados a través del modelo $ARIMA(p,d,q) \times (2,3,1)_{50}$ el cual tuvo el mejor ajuste a los datos disponibles del 2022 según el estimador MSE. No siempre los métodos para crear modelos $ARIMA(p,d,q) \times (P,D,Q)_s$ van a converger: Esto puede ocurrir debido a que la función usada para estimar los parámetros del modelo ARIMA está basada en la función `optim` que hace uso de métodos numéricos, como lo es el *método de Brent* para el cual no siempre se tiene garantizada la existencia de los valores que optimicen la función. Tener un buen ajuste de los datos históricos no necesariamente implica que

[*] Trabajo de Grado

[**] Facultad de Ciencias. Escuela de Matemáticas. Especialización en Estadística. Director: Andrés Sebastián Ríos Gutiérrez. Candidato a Doctor.

se realice una mejor predicción: aunque los modelos $ARIMA(5,1,5) \times (4,2,4)_{50}$ y $ARIMA(5,1,5) \times (2,3,4)_{50}$ presentaron un mejor valor AIC que el modelo $ARIMA(5,1,5) \times (2,3,1)_{50}$, este tiene mejor ajuste para el año 2022 según el MSE. El modelo predice que aproximadamente cada año (casi 52 semanas) se tendrá un pico de contagios por infección respiratoria aguda. Por otro lado, ya que la predicción no tiene comportamiento monótono, se concluye que el modelo no dio alerta de un aumento precipitado del número de casos por infección respiratoria aguda para el año 2022 y por lo tanto no dio alerta de una posible pandemia para ese año como realmente ocurrió.

Abstract

Title: Imputation of missing data in a time series based on known values of prediction, and its application to acute respiratory infection data in Bogotá[*]

Author(s): Diego Johann Reyes Rojas[**]

Key Words: Imputation of missing data, Seasonal time series, Multiplicative ARIMA models, MSE estimator

Description: Objectives: To impute missing data from the years 2020 and 2021 of the history of morbidity cases due to acute respiratory infection in Bogotá using seasonal time series. Methodology: Use the augmented Dickey-Fuller test to study the stationarity of the time series, then differentiate the series in order to make it stationary, choose 5 seasonal ARIMA models using the Akaike criterion and use the MSE estimator to choose which of these is suitable. best fits the available data for the year 2022. Conclusions: The missing data of the infected variable for the year 2020 and 2021 could be imputed through the $ARIMA(p,d,q) \times (2,3,1)_{50}$ model which had the best fit to the available 2022 data according to the MSE estimator. The methods for creating $ARIMA(p,d,q) \times (P,D,Q)_s$ models will not always converge: This may occur because the function used to estimate the parameters of the ARIMA model is based on the optimal function that makes use of numerical methods, such as the Brent method for which the existence of the values that optimize the function is not always guaranteed. Having a good fit of the historical data does not necessarily imply that a better prediction was made: although the $ARIMA(5,1,5) \times$

[*] Degree work

[**]Science Faculty. Math School. Specialization in Statistics. Director: Andrés Sebastián Ríos Gutiérrez. Doctoral Candidate.

$(4,2,4)_{50}$ and $ARIMA(5,1,5) \times (2,3,4)_{50}$ presented a better AIC value than the $ARIMA(p,d,q) \times (2,3,1)_{50}$ model, this has a better fit for the year 2022 according to the MSE. The model predicts that approximately every year (almost 52 weeks) there will be a peak in infections due to acute respiratory infection. On the other hand, since the prediction does not have monotonic behavior, it is concluded that the model did not warn of a precipitous increase in the number of cases due to acute respiratory infection for the year 2022 and therefore did not warn of a possible pandemic for that year as it really happened.

Introducción

La imputación de datos es el proceso por el cual se llenan los valores faltantes o *missing values* de una base de datos a través de técnicas estadísticas como la simulación. Estos datos simulados son generados a través de modelos estadísticos que son escogidos dependiendo de la naturaleza los datos (cantidad de datos, número y tipo de variables en la base de datos, etc.). Por ejemplo, cuando la base de datos reúne la información de una sola variable que se estudia a través del tiempo se buscan modelos para una serie de tiempo. Es común que las bases de datos presenten vacíos de información y una parte importante del análisis estadístico consiste en la limpieza y el arreglo de dichos datos. Es importante considerar la imputación de datos como objeto de estudio en sí mismo porque gran parte de la coherencia de los resultados de un estudio estadístico depende de esta tarea. Según Medina et al. (2007) "Está ampliamente documentado que la aplicación de procedimientos inapropiados de sustitución de información introduce sesgos y reduce el poder explicativo de los métodos estadísticos, le resta eficiencia a la fase de inferencia y puede incluso invalidar las conclusiones del estudio"(p.10). Para los propósitos de este trabajo se estudiará la base de datos correspondiente a los casos de morbilidad por infección respiratoria aguda IRA en Bogotá, desde el año 2010 hasta las primeras 30 semanas del 2022. Dicha base de datos presenta valores faltantes entre el 2020 y el 2021 y el objetivo principal de la investigación consistirá en imputar dichos datos faltantes a través de métodos que sean confiables dada la naturaleza de los datos que se disponen. Es decir, realizar la predicción para la serie de tiempo de la variable `infectados` la cual recoge la información del número de infectados por infección respiratoria aguda en cada semana epidemiológica en los años mencionados anteriormente.

Hay una gran diversidad de métodos para simular datos de series de tiempo en función de la cantidad de datos que se desean imputar. Cuando la cantidad de datos es menor puede hacerse uso de métodos clásicos como la interpolación por medio de splines cuadráticos, cúbicos etc. (ver (Ahlberg, Nilson, & Walsh, 1967)). Cuando estamos frente a una gran cantidad de datos faltantes se recurren a modelos más sofisticados, entre los cuales se encuentran los modelos autorregresivos de medias móviles ARMA, que captan las características regulares o autorregresivas de una serie de tiempo y al mismo tiempo sus características estacionarias a través de su componente de medias móviles. Cuando se requiere modelar series de tiempo que no son estacionarias se usan los modelos autorregresivos integrados de medias móviles o $ARIMA(p,d,q)$ que son una implementación de los modelos ARMA.

Otros modelos que pueden ser utilizados para estudiar las series de tiempo son los modelos de suavizado exponencial ETS (Hyndman & Athanasopoulos, 2018), los modelos de componentes no observables UCM (Harvey, 1990), los modelos GARCH (Bollerslev, 1986), o los modelos de regresión con efectos temporales (Cleveland & Devlin, 1988). En este trabajo se utilizan los modelos ARIMA por su sencilla implementación en lenguaje R (Team, 2020), los cuales pueden ser usados para pronosticar series de tiempo con tendencia monótona (creciente o decreciente).

Para el desarrollo de este trabajo se usaron los modelos ARIMA multiplicativos denotados $ARIMA(p,d,q) \times (P,D,Q)_s$, que a diferencia de los modelos ARIMA iniciales en estos se tiene en cuenta el comportamiento estacional de la serie, esto es, cuando existe correlación entre observaciones separadas periódicamente, como también del comportamiento regular o correlación entre observaciones sucesivas. La construcción de estos modelos se hace en dos etapas: la construcción de un modelo ARIMA para las subseries conformadas por las observaciones que

están separadas periódicamente y finalmente, la construcción de un modelo general a partir de los modelos anteriores.

Si bien es cierto que no hay un modelo que se ajuste perfectamente a una serie de tiempo, los modelos ARIMA son cada vez más usados para el estudio y simulación de estas, y para la predicción e imputación de datos faltantes de una variable que se estudia a lo largo del tiempo (Hyndman & Athanasopoulos, 2018).

A través de los modelos ARIMA multiplicativos se imputarán los datos faltantes de la base de datos correspondiente a la variable de estudio `infectados` en los años 2020 y 2021, la cual representa el número de casos de morbilidad por infección respiratoria aguda en la ciudad de Bogotá, desde el año 2010 hasta las primeras 30 semanas del año 2022. También se medirá el nivel de ajuste del modelo escogido a través del estimador MSE (promedio de la suma de los cuadrados) en los datos históricos del 2022 para reconocer qué tan adecuado es el uso de este tipo de modelos para el pronóstico y la imputación de datos faltantes de una serie de tiempo.

1. Justificación

Debido a la presencia de COVID-19 desde el año 2020, en Bogotá se registraron gran cantidad de casos de morbilidad por infección respiratoria aguda. Esto llevo a que los datos disponibles para los años 2020 y 2021 no se tuvieran en cuenta en parte porque su comportamiento difería bastante de los que ya se tenían desde el año 2010 hasta el 2019. Hacer una predicción de estos datos es importante por varias razones, entre las cuales se podrían considerar las siguientes:

- Planificación de recursos. Con estas predicciones los proveedores de atención médica pueden tener un conocimiento más acertado de cómo se deben asignar recursos de manera más efectiva a futuro, como cantidad de camas, número de vacunas, etc. considerando que no estamos exentos de un nuevo brote de COVID-19 o de otras enfermedades de la misma naturaleza (ver (Arciniegas Paspuel, Castro Morales, & Arias Collaguazo, 2021))
- Gestión de existencias y suministros a futuro (ver (Lebrato, Paradela, Lebrato, & Portales, 2018))
- Generación de campañas de prevención temprana (ver (Lima, 2020)) entre otras.

Además, las predicciones hechas por los modelos sirven como referentes para evaluar si estos realmente son los más apropiados para estudiar la(s) variable(s) que se tiene(n). Para este trabajo, por ejemplo, se tienen datos de las primeras 30 semanas del año 2022, y esto permite que se comparen las predicciones que hace el modelo obtenido para ese mismo año con dichos datos reales. Por otro lado, estimar o imputar valores de una serie de tiempo cuando la cantidad de información faltante es muy grande, como también la predicción de valores futuros, requiere el uso de métodos estadísticos más avanzados que los usados cuando se imputan pocos datos como

es el caso de la interpolación (ver (Ahlberg, Nilson, & Walsh, 1967)). En este sentido, el trabajo pretende ser también una puesta en práctica de conocimientos estadísticos más especializados como las series de tiempo, los procesos estocásticos y los modelos ARIMA entre otros, que deben ser conocidos y manejados en la práctica por un especialista en estadística.

2. Antecedentes

2.1 Antecedentes Nacionales:

- Rodríguez Morales, F. (2012). *Evaluación económica de la inclusión de salas de enfermedad respiratoria aguda en el modelo de atención primaria de la localidad de Ciudad Bolívar*.

La investigación se hizo con la finalidad de reducir los costos y el número de hospitalizaciones de pacientes con enfermedades respiratorias agudas en Bogotá, Ciudad Bolívar. Se usó un modelo ARIMA que ayudó a determinar los costos y las unidades de efectividad (pacientes los cuales fue evitable la hospitalización) en el caso en que se planeara dicha problemática sin intervenciones. Gracias al uso este modelo junto con otras herramientas estadísticas y computacionales como los árboles de decisión se logró construir el programa Salas ERA que redujo los costos en un valor cercano a $674 millones de pesos y el número 228 hospitalizaciones de un total 286 que se tuvieron en cuenta en este estudio.

- García García, C. (2014). *Caracterización epidemiológica de la infección respiratoria aguda grave y circulación viral en Boyacá, julio de 2012 a julio de 2013*.

En esta investigación se tuvo en cuenta número de casos de personas que ingresaban con síntomas de infección respiratoria aguda grave desde el mes de julio de 2012 hasta julio de 2013 por medio del Sistema de Vigilancia en Salud Pública en Boyacá. Para este estudio se tuvo en cuenta una población de 530 personas las cuales se le realizaron procedimientos para el reconocimiento de la morbilidad. Se concluyó dentro de otros resultados que, aproximadamente el 62.5% de las personas que fueron hospitalizadas en Boyacá tenía como causal una infección respiratoria aguda para ese intervalo de tiempo. El estudio realizado hace uso de métodos

estadísticos descriptivos para el estudio de la variable número de casos ingresados por infección respiratoria aguda grave a través de tiempo.

- Garces, C.J. (2023). *Modelización del covid-19 en Santander mediante series temporales*. Universidad Industrial de Santander, Bucaramanga.

En este trabajo se hace uso de un modelo ARIMA para predecir la cantidad de casos de personas con covid-19 en Santander con datos recolectados por el Instituto Nacional de Salud los cuales son de libre acceso. También se hizo uso un modelo GARCH con el mejor ajuste a los datos disponibles.

- Ovalles, Y.B & Velásquez, J.N. (s.f.). *Respiratory viral infections in pediatrics: generalities about physiopathogeny, diagnosis and clinical outcomes*.

En este trabajo se ahonda en la importancia de estudiar las infecciones respiratorias agudas y la incidencia que tiene en la morbilidad y mortalidad en la población infantil. Cuantitativamente se muestra que el 50% de las personas con infección respiratoria aguda está relacionada principalmente con los virus. Se resalta la importancia no solamente de tratar la enfermedad sino también de prevenirla y de tener información clara y novedosa que llega a las entidades de salud. La prevención de la enfermedad y su evolución a través del tiempo es un tema de mayor importancia como se menciona en este trabajo y está directamente relacionado con la imputación y pronóstico de datos por medio de modelos de series de tiempo.

2.2 Antecedentes Internacionales:

- Lugo Meléndez, A. *Capacitación a la Población sobre la Prevención y Manejo de IRAS y EDAS*.

En este estudio se resalta la importancia de realizar capacitaciones como de tener un conocimiento más claro y contextual de las infecciones respiratorias agudas, con datos precisos y

cifras pues es un tema de consulta recurrente, especialmente en poblaciones de personas mejores a 5 años.

- Pham, H. T., Do-Thi, T. T., Baek, J., Nguyen, C. K., Pham, Q. T., Nguyen, H. L., ... & Le, G. M. *Handling missing data methods for estimating the incidence rate of COVID-19 pandemic data: A case study in Vietnam.*

En este trabajo se mostró la importancia de manejar datos faltantes y como estos afectan a las estimaciones de las tasas de incidencia del covid-19. Se concluyó además que el método más adecuado para imputar los datos faltantes en el contexto de la enfermedad por covid-19 es imprescindible para garantizar la confiabilidad de los resultados estadísticos. Para este trabajo se usaron y compararon diferentes métodos como el sesgo bruto absoluto medio, el error cuadrático medio bruto y el cambio medio porcentual absoluto. Para la imputación de datos faltantes de la base se trabajó con modelos ARIMA.

- Li, Y., Liu, X., Li, X., Xue, C., Zhang, B., & Wang, Y. (2023). *Interruption time series analysis using autoregressive integrated moving average model: evaluating the impact of COVID-19 on the epidemic trend of gonorrhea in China*. *BMC Public Health*, *23*(1), 1-11.

En esta investigación se evaluó el impacto que tiene el covid-19 en el comportamiento epidémico y prevención de la gonorrea en la ciudad de China haciendo uso de un modelo ITS-ARIMA. Se tomó como variable el número de casos de gonorrea notificados en China desde enero de 2005 hasta septiembre de 2022. Además, se comparó el método mencionado con otro, de las series temporales estructurales bayesianas en términos de su mejor ajuste, teniendo este último mayor porcentaje de error que el primero. Se llegó a la conclusión de que el modelo $ARIMA\big(0,1,(1,3)\big) \times (0,1,1)_{12}$ y se evidencia en este trabajo que los modelos ARIMA y sus variantes son ampliamente usados en contextos de enfermedad por infección respiratoria aguda.

- Singh, D., Halder, S., Bhattacharyya, S., Nath, I., Sahana, S., Pal, S., & Alkhafaji, M. A. (2023, September). *COVID-19 case analysis in India using EDA and its prediction. In AIP Conference Proceedings (Vol. 2845, No. 1).*,

En este trabajo se tuvo en cuenta los casos de covid-19 en la India para estudiar su propagación, por medio de técnicas avanzadas como el Análisis Exploratorio de Datos y el uso de modelos de predicción bien conocidos entre los cuales se encuentran los modelos Prophet, ARIMA, XGBoost Regressor entre otros y con los cuales se hicieron las respectivas predicciones.

3. Objetivos

3.1 Objetivo General

- Imputar datos faltantes de los años 2020 y 2021 del histórico de los casos de morbilidad por infección respiratoria aguda en Bogotá utilizando series de tiempo estacionales

3.2 Objetivos Específicos

- Utilizar modelos estacionales multiplicativos para la predicción de valores futuros de la serie de tiempo del histórico de casos de morbilidad por infección respiratoria aguda en Bogotá
- Comparar mediante el cuadrado medio del error de las predicciones dadas por los modelos estacionales para los días con información faltante (años 2020 y 2021) bajo los multiplicativos previamente implementados, con los datos disponibles de predichos (años 2022)

4. Marco Conceptual

A continuación, se resumen los conceptos y las definiciones estadísticas que dan contexto al trabajo. La información se basó principalmente de Casimiro (2009).

Proceso estocástico. Un proceso estocástico es una familia de variables aleatorias ordenadas por un conjunto T. Esto es, una colección $\{X_t\}$ de variables aleatorias, donde $t \in T$. En este contexto T representa tiempos, por lo que, en general, se tiene que $T \subseteq [0, \infty)$. Una *serie temporal* es una realización del proceso, que es observado únicamente para un número finito de valores de t. Por ejemplo, $x_{t_1}, x_{t_2}, \dots, x_{t_n}$ es una serie temporal de longitud n donde $x_{t_k} \in X_{t_k}$ para todo k. Un proceso se dice de tiempo continuo si T es un intervalo, por ejemplo $[0, \infty)$ como se definió anteriormente, o de tiempo discreto si por ejemplo $T = \mathbb{N}$.

4.1 Procesos Estocásticos Estacionarios

Hay diferentes tipos de procesos estocásticos, entre los cuales se encuentran los procesos estacionarios. Estos pueden ser definidos de dos formas:

Proceso estocástico estrictamente estacionario. Sea $\{X_t\}$ un proceso estocástico. Se dice que el proceso es estrictamente estacionario si se cumple que

$$F\left[X_{t_1}, X_{t_2}, \dots, X_{t_n}\right] = F\left[X_{t_1+k}, X_{t_2+k}, \dots, X_{t_n+k}\right]$$

Para todo $t_1, t_2, \dots, t_n, k$.

En la definición anterior, la función F es llamada *función de distribución* y es usada para medir la distribución de un subconjunto finito de variables aleatorias en el proceso. La anterior condición dice que el proceso no se altera luego de desplazarse k periodos en el tiempo.

Proceso estocástico ampliamente estacionario. Sea $\{X_t\}$ un proceso estocástico. Se dice que el proceso es ampliamente estacionario si se cumplen las siguientes condiciones:

a. $E(X_t) = \mu_X$ para todo $t \geq 0$;

b. $Var(X_t) = \sigma_X^2$ para todo $t \geq 0$;

c. $Cov(X_s, X_t) = f_X(t - s)$ con $t \geq s \geq 0$.

Los $E(X_t)$ son llamados *niveles del proceso* y la función $f_X(t - s)$ se denomina *función de autocovarianza*. La tercera condición menciona que la covarianza entre dos variables cualesquiera del proceso, dependen únicamente (por medio de f_X) del número de periodos que separa dichas variables. Por otro lado, la segunda condición se obtiene de la última ya que $Cov(X_t, X_t) = f_X(0) = \sigma_X^2$. Por lo tanto, la varianza en un proceso estacionario en sentido amplio es constante a lo largo del tiempo e igual a $f_X(0)$. Además, se dice que un proceso es estacionario en media si se cumple la condición a. y es estacionario en covarianza si se cumple b. y c.

El siguiente resultado establece condiciones para la equivalencia entre los dos tipos de estacionariedad (ver (Casimiro, 2009, pág. 14)):

Teorema sobre estacionariedad. Sea $\{X_t\}$ un proceso estocástico. Si la distribución del proceso es normal, entonces el proceso es estacionario en sentido amplio si, y solo si, es estacionario en sentido estricto.

Función de autocorrelación (FAC). Sea $\{X_t\}$ un proceso estacionario. La función de autocorrelación de $\{X_t\}$ es una función ρ tal que a cada valor $k = 0,1,2,3, \ldots$ se le asigna un índice ρ_k, el cual mide el grado de asociación lineal entre dos variables aleatorias del proceso separadas por k periodos. Dicho índice para un valor k está definido como sigue:

$$\rho_k = \frac{Cov(X_t, X_{t+k})}{\sqrt{Var(X_t) \cdot Var(X_{t+k})}}$$

Donde $|\rho_k| \leq 1$ para todo k.

La forma de visualizar la función de autocorrelación es a través de un correlograma el cual es una representación gráfica que muestra las correlaciones entre valores de una serie de tiempo y sus valores pasados a medida que el tiempo avanza. Si se trata de un proceso estacionario dicho correlograma va a mostrar un comportamiento decreciente conforme se toman valores mayores de k y tendiendo exponencialmente a 0.

Los siguientes procesos estocásticos son importantes porque hacen parte de la construcción de los modelos que se usarán en este trabajo:

Proceso de ruido blanco $RB(0, \sigma^2)$. Es un proceso $\{a_t\}$ tal que se cumplen las siguientes condiciones:

i. $E(a_t) = 0$;

ii. $Var(a_t) = \sigma_a^2 < +\infty$ para todo t;

iii. $Cov(a_t, a_s) = 0$ para todo $s \neq t$.

4.2 Representación Lineal de un Proceso Estocástico Estacionario

Para que un proceso estacionario pueda ser escrito como una combinación lineal infinita se deben cumplir las siguientes condiciones generales:

- Para cada tiempo t el proceso se pueda descomponer como suma de una *parte regular* o predecible y de otra totalmente aleatoria llamada *innovación* denotada por a_t. Es decir, para cada t se tenga que $X_t = parte\ regular_t + a_t$
- El proceso debe ser *no anticipante* e *invertible.* La primera condición hace referencia a que X_t no dependa de los valores futuros $X_{t+1}, X_{t+2}, \ldots$, y la segunda que la influencia de X_{t-k} sobre X_t disminuye cuando los valores de k se hacen cada vez más grandes.

Cuando el proceso estacionario cumple las condiciones anteriores, se puede representar de la siguiente forma llamada *representación puramente autorregresiva* denotada por $AR(\infty)$:

$$X_t = \pi_1 X_{t-1} + \pi_2 X_{t-2} + \pi_3 X_{t-3} + \cdots + a_t = \sum_{i=1}^{\infty} \pi_i X_{t-i} + a_t$$

donde $\sum_{k=1}^{\infty} \pi_k^2 < \infty$ y $a_t \sim RB(0, \sigma_a^2)$ para todo t.

El polinomio infinito $1 - \pi_1 L - \pi_2 L^2 - \cdots = \Pi_\infty(L)$ donde $L^k(X_t) = X_{t-k}$, es llamado *operador de retardos* que juega un rol importante en la construcción de los modelos ARIMA.

Por otro lado, el siguiente teorema afirma que también se puede construir un modelo general que dependa solamente de la perturbación contemporánea a_t y de su pasado infinito (ver (Balakrishnan, 2010)):

Teorema de Wold. Sea $\{X_t\}$ un proceso estocástico. Si $\{X_t\}$ es un proceso estacionario entonces este se puede ser representado linealmente como:

$$X_t = a_t + \psi_1 a_{t-1} + \psi_{t-2}\, a_{t-2} + \psi_{t-3} a_{t-3} + \cdots = \sum_{i=1}^{\infty} \psi_i a_{t-i}.$$

Donde $\psi_0 = 1, \quad \sum_{i=0}^{\infty} \psi_i^2 < \infty$ y $a_t \sim RB(0, \sigma_a^2)$ para todo t.

Esta representación es llamada *representación puramente de medias móviles* y se denota por $MA(\infty)$.

Las condiciones de ser invertible y no anticipante en estos procesos se respetan gracias a la convergencia de las series en ambos modelos generales y la forma lineal de los mismos, que depende únicamente de los tiempos pasados y del tiempo presente. Una consecuencia del Teorema de Wold es que si en los procesos de ruido blanco de las representaciones anteriores, las variables a_t son independientes entre sí, dichas representaciones de $\{X_t\}$ son únicas (ver (Casimiro, 2009)).

4.3 Modelos para Representar Procesos Estocásticos Estacionarios

Debido a que un modelo no puede tener una cantidad infinita de parámetros (no sería posible procesar o estimar una cantidad infinita de parámetros por medio de una computadora), se deben realizar restricciones a las representaciones $AR(\infty)$ y $MA(\infty)$. Estas restricciones dan paso a los modelos $AR(p)$, $MA(q)$ y $ARMA(p,q)$ que se construyen a partir del siguiente resultado gracias a la teoría de polinomios (Casimiro, 2009, pág. 21):

$$\Pi_\infty(L) \sim \frac{\phi_p(L)}{\theta_q(L)}$$

donde $\phi_p(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p$ y $\theta_q(L) = 1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_q L^q$ son llamados *polinomio autorregresivo* y *polinomio de medias móviles* del proceso respectivamente.

Modelo Autorregresivos $AR(p)$. Dado un proceso estacionario $\{X_t\}$, el modelo autorregresivo $AR(p)$ es una restricción a la representación $AR(\infty)$ del proceso, tomando solamente sus primeras p componentes

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t$$

donde $a_t \sim RB(0, \sigma_a^2)$ y $a_t \sim N(0, \sigma_a^2)$.

Estos modelos vistos como procesos estocásticos en sí mismos no necesariamente son estacionarios. Para garantizar su estacionariedad se debe cumplir que las raíces del polinomio autorregresivo $\phi_p(L)$ tengan siempre módulo mayor que 1 (Casimiro, 2009, pág. 27).

Modelo de Medias Móviles $MA(q)$. Dado un proceso estacionario $\{X_t\}$, el modelo de medias móviles $MA(q)$ es una restricción a la representación $MA(\infty)$ del proceso, tomando solamente sus primeras q componentes

$$X_t = a_t - \theta_1 a_{t-1} - \theta_{t-2}\, a_{t-2} - \theta_{t-3} a_{t-3} - \cdots - \theta_{t-q} a_{t-q}$$

donde $a_t \sim RB(0, \sigma_a^2)$ y $a_t \sim N(0, \sigma_a^2)$ para todo t.

Estos modelos, vistos como procesos estocásticos en sí mismos cumplen la propiedad de que siempre son estacionarios. Sin embargo, no siempre son invertibles y para garantizar que lo sean se debe cumplir que las raíces del polinomio de medias móviles $\theta_q(L)$ tengan siempre módulo mayor que 1 (Casimiro, 2009, pág. 37).

Los modelos $MA(q)$ son reescritos de la siguiente forma para la construcción de los modelos $ARMA(p, q)$:

$$X_t = a_t - \theta_1 a_{t-1} - \theta_{t-2}\, a_{t-2} + \theta_{t-3} a_{t-3} - \cdots - \theta_{t-q} a_{t-q}.$$

Modelo de Medias Móviles $ARMA(p, q)$. Dado un proceso estacionario $\{X_t\}$, el modelo autorregresivo $ARMA(p, q)$ combina ambos modelos $AR(p)$ y $MA(q)$ donde se considera tanto la componente regular como la de medias móviles:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t - \theta_1 a_{t-1} - \theta_{t-2}\, a_{t-2} + \theta_{t-3} a_{t-3} - \cdots - \theta_{t-q} a_{t-q}$$

donde $a_t \sim RB(0, \sigma_a^2)$ y $a_t \sim N(0, \sigma_a^2)$ para todo t.

No siempre los modelos $ARMA(p, q)$ son estacionarios e invertibles. Para garantizarlo se necesita que el módulo de las raíces del polinomio autorregresivo y las del polinomio de medias móviles sea mayor que 1 respectivamente (Casimiro, 2009, pág. 39). Por otro lado, la normalidad de $\{a_t\}$ en cada uno de los modelos anteriores se exige ya que esto permite concluir que dichas representaciones sean únicas (ver (Casimiro, 2009)).

4.4 Modelo Autorregresivo Integrado de Medias Móviles ARIMA(p,d,q)

Cuando el proceso que se busca modelar no es estacionario en media, es decir, que μ_X cambia con el tiempo, se puede usar una técnica a fin de generar un nuevo proceso que si es estacionario llamada *diferenciación*. Los modelos que se ajustan al nuevo proceso diferenciado se

llaman $ARIMA(p,d,q)$ donde d es el número de veces que se debe llevar a cabo dicha técnica a fin de que se obtenga un nuevo proceso con media constante (Casimiro, 2009, pág. 46).

Proceso integrado. Sea $\{X_t\}$ un proceso estocástico no estacionario. Se dice que $\{X_t\}$ es un proceso integrado de orden d si el proceso $\{\Delta^d X_t\}$ es estacionario, donde

$$\Delta^d X_t = (1-L)^d X_t.$$

Modelo autorregresivo integrado de medias móviles $ARIMA(p,d,q)$. Dado un proceso estocástico $\{X_t\}$, el modelo autorregresivo de medias móviles $ARIMA(p,d,q)$ esta definido de la siguiente forma:

$$\varphi_{p-d}(L)\Delta^d X_t = \delta + \theta_q(L)a_t$$

donde $\varphi_{p-d}(L) = \left(1-\frac{L}{L_1}\right)\left(1-\frac{L}{L_2}\right)\cdots(1-\frac{L}{L_{p-d}})$ y $L_1, L_2, \dots, L_{p-d}$ son las raíces estacionarias del polinomio $\phi_p(L)$ que tienen módulo mayor que 1, y δ es constante.

4.5 Modelos ARIMA Estacionales

Una serie de tiempo $x_1, \dots, x_T$ tiene comportamiento estacional si dado un tiempo t, la observación x_t y las observaciones de los periodos pasados $x_{t-s}, x_{t-2s}, x_{t-3s}, \dots$ están temporalmente correlacionadas. El valor s depende de la escala de tiempo usada. Por ejemplo, si la escala de tiempo son los meses y las observaciones que están correlacionadas se dan el mismo mes de cada año, entonces $s = 12$. A continuación, se mostrarán los modelos ARIMA estacionales que son usados cuando se tienen este tipo de series temporales.

Modelo ARIMA estacional puro $ARIMA(p,d,q)_s$. Dado un proceso estocástico $\{X_t\}$ el modelo ARIMA estacional puro $ARIMA(p,d,q)_s$ está definido de la siguiente forma:

$$\Phi_p(L^s)(1-L^s)^d X_t = \delta + \Theta_q(L^s)a_t$$

donde δ es constante, $a_t \sim RB(0, \sigma_a^2)$ y los polinomios $\Phi_P(L^s)$ y $\Theta_Q(L)$ son llamados *polinomios autorregresivos y de medias móviles estacionales* respectivamente. Estos modelos son idénticamente iguales a los modelos $ARIMA(p,d,q)$ con la diferencia de que los polinomios autorregresivos y de medias móviles no están en función de L sino de L^s. Estos modelos son más sencillos y se usan únicamente para captar las características estacionales de una serie temporal.

4.5.1 Modelos ARIMA Multiplicativos ARIMA(p,d,q) × (P,D,Q)$_s$

A diferencia de los modelos estacionales puros, que como se mencionó en la sección anterior solamente tienen en cuenta la relación lineal entre observaciones periódicas, los modelos que se van a construir en esta sección consideran también la relación dentro de los periodos, esto es, de observaciones consecutivas entre cada observación periódica. Sea $\{X_t\}$ un proceso estacional de periodo s bajo un modelo $ARIMA(p,d,q)_s$ y consideremos la serie $X_1, X_2, \dots, X_N$. Dicha serie puede separarse en s subseries las cuales se denotan por $X^{(1)}, X^{(2)}, \dots, X^{(s)}$ y definidas de la siguiente forma:

$X^{(1)}$	$X^{(2)}$	$\cdots$	$X^{(j)}$	$\cdots$	$X^{(s)}$
X_1	X_2	$\cdots$	X_j	$\cdots$	X_s
X_{1+s}	X_{2+s}	$\cdots$	X_{j+s}	$\cdots$	X_{2s}
X_{1+2s}	X_{2+2s}	$\cdots$	X_{j+2s}	$\cdots$	X_{3s}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$X_{j+(i-1)s}$	$X_{j+(i-1)s}$	$\cdots$	$X_{j+(i-1)s}$	$\cdots$	X_{is}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$X_{1+(n-1)s}$	$X_{1+(n-1)s}$	$\cdots$	$X_{j+(n-1)s}$	$\cdots$	X_{ns}

Cada subserie consta de n observaciones las cuales están dispuestas en cada columna del arreglo matricial anterior. Si se denota por $X_i^{(j)}$ la i −ésima observación de la j −ésima subserie, se tiene la siguiente ecuación que relaciona las observaciones de cada subserie con las de la serie inicial:

$$X_i^{(j)} = X_{j+s(i-1)}.$$

El arreglo anterior con esta última notación se vería de la siguiente forma:

$X^{(1)}$	$X^{(2)}$	$\cdots$	$X^{(j)}$	$\cdots$	$X^{(s)}$
$X_1^{(1)}$	$X_1^{(2)}$	$\cdots$	$X_1^{(j)}$	$\cdots$	$X_1^{(s)}$
$X_2^{(1)}$	$X_2^{(2)}$	$\cdots$	$X_2^{(j)}$	$\cdots$	$X_2^{(s)}$
$X_3^{(1)}$	$X_3^{(2)}$	$\cdots$	$X_3^{(j)}$	$\cdots$	$X_3^{(s)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$X_i^{(1)}$	$X_i^{(2)}$	$\cdots$	$X_i^{(j)}$	$\cdots$	$X_i^{(s)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$X_n^{(1)}$	$X_n^{(2)}$	$\cdots$	$X_n^{(2)}$	$\cdots$	$X_n^{(s)}$

Ya que el periodo de la serie inicial es s, cada una de las subseries $X^{(j)}$ anteriores no es estacionaria y puede ser modelada por medio de un modelo $ARIMA(p,d,q)$ no estacional. Para esta construcción se asumirá que todos los modelos $ARIMA(p,d,q)$ tienen los mismos parámetros p,d,q y, por lo tanto, cada $\{X_i^{(j)}\}$ se puede representar como:

$$\left(1-\phi_1 L-\cdots \phi_p L^p\right)(1-L)^d X_i^{(j)} = \left(1-\theta_1 L-\cdots \theta_q L^q\right)a_i^{(j)}$$

$$\phi_p(L)(1-L)^d X_i^{(j)} = \theta_q(L)a_i^{(j)}.$$

Para todo $1 \leq j \leq s$, $1 \leq i \leq n$, y donde $a_i^{(j)} \sim RB(0, \sigma^2)$.

La expresión $a_i^{(j)}$ representa la innovación para cada tiempo $t = j + s(i-1)$. El número de diferenciaciones debe ser necesariamente mayor a 0 puesto que si $d = 0$, cada subserie sería estacionaria y esto implicaría que la serie completa sería estacionaria, lo cual es contradictorio puesto que esta es estacional.

Por otro lado, se tiene que $L = L^s$. En efecto, ya que

$$L = \frac{X_{i-1}^{(j)}}{X_i^{(j)}}$$

Que equivale también a que $LX_i^{(j)} = X_{i-1}^{(j)}$, entonces

$$LX_i^{(j)} = X_{i-1}^{(j)} = X_{j+s(i-1-1)} = X_{j+s(i-1)-s} = X_{i-s}^{(j)} = L^s X_i^{(j)}.$$

Y esto implica que

$$L^s = \frac{X_{i-1}^{(j)}}{X_i^{(j)}} = L.$$

Esta propiedad permite escribir los polinomios autorregresivos y de medias móviles en función de L^s lo cual reduce significativamente la cantidad de parámetros del modelo buscado lo cual es deseable (principio de parsimonia) y la representación del modelo para cada subserie queda así:

$$\phi_p(L^s)(1-L^s)^d X_i^{(j)} = \theta_q(L^s) a_i^{(j)}.$$

Ya que cada valor de t desde 1 hasta N es igual a algún subíndice $j + s(i-1)$, el modelo anterior puede reescribirse de siguiente forma para modelar la serie de tiempo completa:

$$\phi_p(L^s)(1-L^s)^d X_t = \theta_q(L^s) a_t.$$

Sin embargo, hay que tener en cuenta que, aun cuando para cada j, la subserie $a_i^{(j)}$ es un proceso de ruido blanco, el proceso estocástico $\{a_t\}_{t\geq 0}$ no es un ruido blanco porque en la serie original $X_1, X_2, \dots, X_N$ hay dependencia lineal entre observaciones sucesivas y por lo tanto también se tendrá esta dependencia entre valores consecutivos para el proceso $\{a_t\}_{t\geq 0}$. Por consiguiente, el proceso $\{a_t\}_{t\geq 0}$ se puede modelar a través de un proceso $ARIMA(P, D, Q)$ que tiene la siguiente forma:

$$\Psi_P(L)(1-L)^D a_t = \Gamma_Q(L)u_t$$

donde $u_t \sim RB(0, \sigma^2)$.

Finalmente, integrando estos dos últimos modelos se obtiene el siguiente modelo denominado *modelo ARIMA estacional multiplicativo* $ARIMA(p, d, q) \times (P, D, Q)_s$:

$$\phi_p(L^s)\Psi_P(L)(1-L^s)^d(1-L)^D X_t = \theta_q(L^s)\Gamma_Q(L)u_t.$$

Este modelo recoge las siguientes características:

- Las estacionalidades estocásticas de la serie;
- Las tendencias estocásticas;
- Las posibles interacciones entre y dentro de las subseries que definen las estacionalidades.

5. Metodología

Para la elaboración y el cumplimiento de los objetivos de este trabajo la metodología usada fue la siguiente:

5.1 Base de Datos `casos_IRA`

La base de datos corresponde a los casos de morbilidad por infección respiratoria aguda en la ciudad de Bogotá, los cuales fueron tomados en las 52 semanas epidemiológicas de cada año desde el 2010 hasta las primeras 30 semanas del 2022. Esta base de datos fue proporcionada indirectamente por el director de esta investigación, y hace parte del Ministerio de Salud y Protección Social en Colombia. Consta de una sola variable de estudio, a saber, la variable `infectados` y el tamaño de la muestra está determinado por la cantidad de valores que toma esta, de 550 datos (excluyendo el conteo de los datos faltantes en el 2020 y 2021).

5.2 Información de la Variable de Estudio

La variable de estudio `infectados` corresponde a la cantidad de infectados por infección respiratoria aguda en Bogotá por cada semana epidemiológica desde el 2010 hasta el 2022. Es una variable de tipo discreto y con escala de medición de razón.

5.3 Modelización a Través de los Modelos ARIMA

Para la modelización usando series de tiempo multiplicativas se toma lo realizado en el artículo de (Alfaki & Masih, 2015), donde se realiza el siguiente procedimiento:

- Se determina el número de diferenciaciones regulares utilizando la función `diff` del paquete `forecast, l`a cual determina el orden d.
- Se determina los órdenes p y q realizando las funciones de autocorrelación parcial y autocorrelación, respectivamente. En (Casimiro, 2009) se observa que en general

el orden de p es aquel bajo el cual la función de autocorrelación parcial tiende exponencialmente a cero y el orden de q es el cual la función de autocorrelación tiende exponencialmente a cero.

- Se determina la estacionalidad de la serie de tiempo utilizando el periodograma: donde se observa un máximo se tiene que para esa frecuencia hay un comportamiento estacional de acuerdo con (Reisen, C, MS, & Taqqu, 2017).
- Para los órdenes P, D, Q menores o iguales a 4 se determinan los criterios de Akaike. Se escogen los modelos para los cuales se tiene convergencia del método[††], algunos de los cuáles divergen. En (Casimiro, 2009) recomiendan no tomar órdenes muy altos para conservar un modelo parsimonioso.

Esta metodología se ha probado con simulaciones en donde se observa un adecuado ajuste de los datos con respecto a los pronósticos del modelo ARIMA con mejor ajuste encontrado.

5.4 Criterio de Akaike para la Escogencia de los Modelos

Se considera que $a_t \sim N(0, \sigma^2)$ con valores dados por $u_t = X_t - f(t)$ donde $f(t)$ corresponde al valor pronosticado por el modelo estacional multiplicativo. Bajo esto, en (Brockwell & A, 2002) se toma el criterio de Akaike por:

$$AIC = -2 \ln f(a_1, \dots, a_t) + 2K$$

Con K igual al número de parámetros, y

[††] La función que se utiliza para estimar los parámetros del modelo ARIMA está basado en la función optim (ver (Truong, Nguyen, Truong, Ho, & Diem-Chinh, 2019)) basadas en métodos numéricos, entre los cuáles está el método de Brent para el cual no siempre se tiene garantizada la existencia de los valores que optimizan la función, como se dice en (Gegenfurtner, 1992).

$$f(a_{t_1},\dots,a_{t_n}) = \frac{1}{\sqrt{(2\pi\sigma^2)^n r_1\cdots r_n}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\left(X_{t_i}-\hat{X}_{t_i}\right)^2}{r_i}\right\},$$

Donde $r_i = Var(a_{t_i})$ para todo i, cuya ecuación se puede determinar analíticamente como se hace en el modelo ARMA (ver (Casimiro, 2009, pág. 115)).

5.5 Imputación de los Datos Faltantes y Pronóstico del 2022

La base de datos cuenta con 105 datos faltantes los cuales están situados entre del 2020 y el 2021. Para el pronóstico se determinó primero el modelo $ARIMA(p,d,q)\times(P,D,Q)_s$ de mejor ajuste y luego se hizo uso de la función `forecast()` del paquete `forecast` que realiza dicha predicción. Algunos otros datos faltantes que se presentaban en menor cantidad (máximo grupos de dos datos faltantes) fueron imputados a través de métodos clásicos (interpolación usando splines (Ahlberg, Nilson, & Walsh, 1967)).

5.6 Escogencia del Modelo con Mejor Ajuste a los Datos Del 2022

Para la escogencia del modelo $ARIMA(5,1,5)\times(P,D,Q)_s$ que se ajuste mejor a los datos históricos del 2022 (las primeras 30 semanas que se tienen registro en la base de datos) se usó como estimador el cuadrado medio del error (ver (Hastie, Tibshirani, & Friedman, 2009)), el cual mide el promedio de los errores al cuadrado. Su fórmula está definida como:

$$MSE = \frac{1}{N}\sum_{k=1}^{N}[x_k - \hat{x}_k]^2.$$

Donde x_k y $\hat{x}_k$ son los valores reales y estimados por el modelo respectivamente.

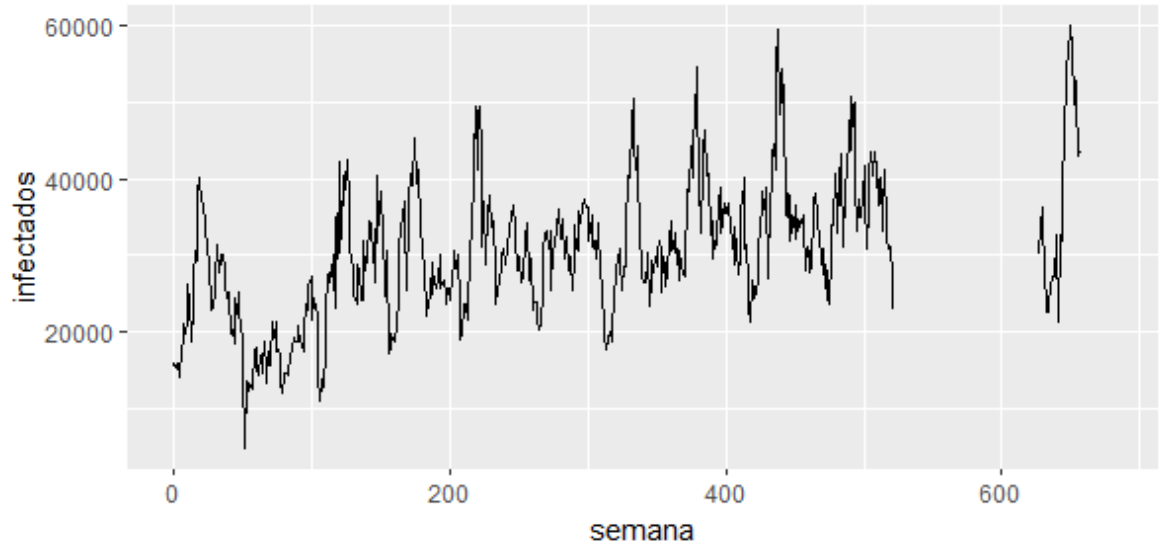
6. Resultados

A continuación, se muestra el procedimiento estadístico que se llevó a cabo para la escogencia de los modelos ARIMA multiplicativos que se tuvieron en cuenta en el estudio, usando la función de autocorrelación y de autocorrelación parcial que fijan los parámetros p, d, q del componente regular de los procesos $ARIMA(p, d, q)$ no estacionales y luego haciendo uso del criterio de Akaike para la escogencia de los parámetros P, D, Q de la componente estacional.

6.1 Descripción de la Serie de Tiempo de la Variable Respuesta `infectados`

Usando la función `ggplot()` del paquete `ggplot2` se generó la gráfica de la serie de tiempo de la variable de estudio `infectados`.

Figura 1. *Gráfica de la serie de tiempo de la variable `infectados`*



Los valores en el eje horizontal `semana` corresponden a las semanas epidemiológicas (52 por cada año desde el 2010 hasta el 2022) y el eje vertical `infectados` al conteo de infectados para cada valor de tiempo. La serie tiene valores faltantes entre las semanas 520 y 624 en las cuales se localizan los años 2020 y 2021, que son los años en donde se sabe que hay ausencia de datos.

Se observó tendencia o trend en la gráfica de la serie puesto que hay presencia de más de un nivel a lo largo del cual la serie sube y baja (drift), por lo tanto, se dedujo gráficamente que el proceso no es estacionario en media. Aunque el estudio de la estacionariedad en varianza de la serie no se hace muy evidente a través de la gráfica, al aplicar la prueba de Dickey-Fuller aumentada a la serie (función `adf.test()` del paquete `aTSA`) se obtuvo la siguiente salida:

Imagen 1. *Prueba de Dickey-Fuller para la serie* `infectados`

```
> adf.test(serie_arima_ira)
Augmented Dickey-Fuller Test
  alternative: stationary

 Type 1: no drift no trend
     lag   ADF p.value
[1,]   0 -1.47   0.154
[2,]   1 -1.12   0.279
[3,]   2 -1.12   0.279
[4,]   3 -1.08   0.294
[5,]   4 -1.27   0.226
[6,]   5 -1.14   0.273
Type 2: with drift no trend
     lag   ADF p.value
[1,]   0 -5.83    0.01
[2,]   1 -4.73    0.01
[3,]   2 -4.81    0.01
[4,]   3 -4.84    0.01
[5,]   4 -5.62    0.01
[6,]   5 -5.19    0.01
Type 3: with drift and trend
     lag   ADF p.value
[1,]   0 -6.84    0.01
[2,]   1 -5.48    0.01
[3,]   2 -5.64    0.01
[4,]   3 -5.70    0.01
[5,]   4 -6.82    0.01
[6,]   5 -6.37    0.01
                 ----
Note: in fact, p.value = 0.01 means p.value <= 0.01
```

La prueba evidenció que, con una confianza del 99%, la serie temporal es del tipo 3. Esto es, con una confianza del 99%, se comprobó que la serie no es estacionaria ni en media ni en varianza.

6.2 Determinación de los Parámetros Regulares p, d, q y Creación del Modelo Base ARIMA(p,d,q)

Una vez se estudió la estacionariedad de la serie, se determinó el número de diferenciaciones que se debían aplicar a fin de transformar la serie en estacionaria usando la función `ndiffs()` del paquete `forecast`. En este caso la función aplicada a la serie imprimió un valor de $d = 1$, con lo cual se concluyó que la serie diferenciada una vez resultó ser estacionaria. Está es otra forma de determinar la no estacionariedad de la serie, pues si la serie fuera estacionaria el número de diferenciaciones habría sido de $d = 0$.

Con las funciones de autocorrelación y autocorrelación parcial de la serie (funciones `pacf()` y `acf()` del paquete base de R) se encontraron los parámetros p y q respectivamente, más apropiados para que el modelo base $ARIMA(p, d, q)$ tuviera el mejor ajuste a la serie. Las funciones generaron los siguientes gráficos:

Figura 2. *Gráfica de ACF para la serie de* `infectados`

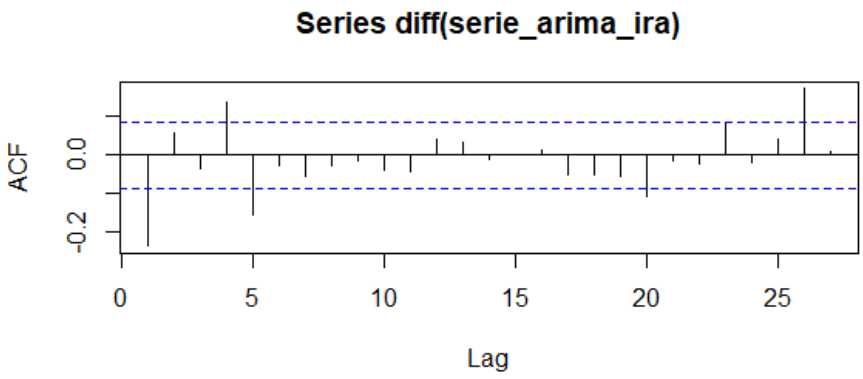
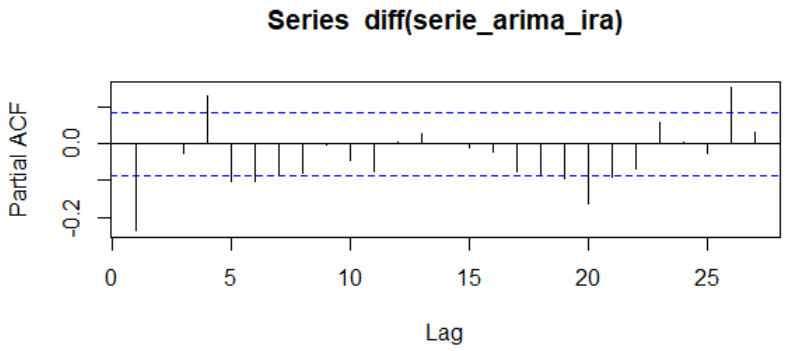


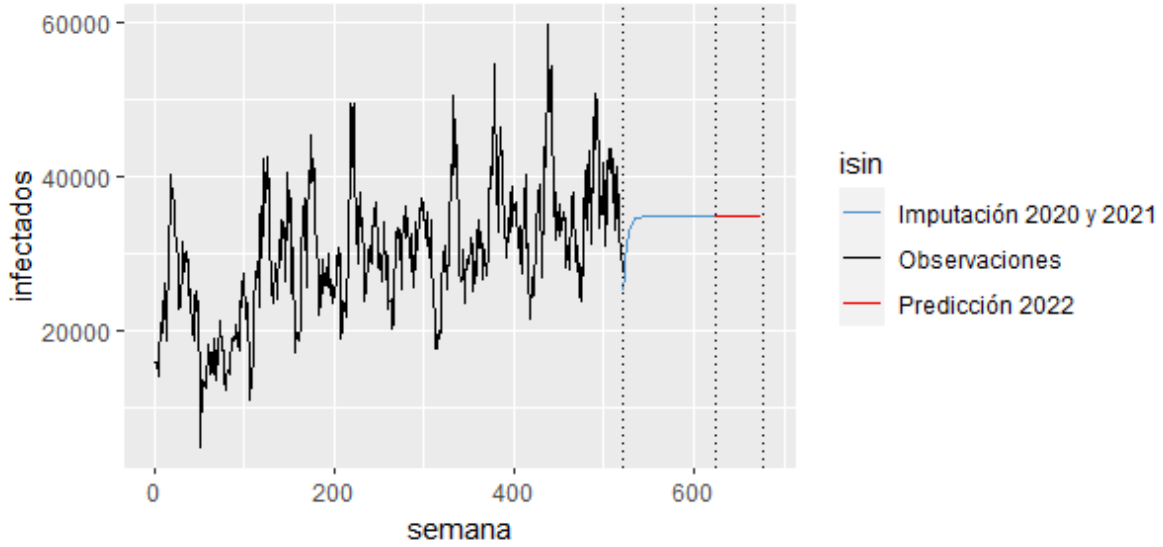
Figura 3. *Gráfica de PACF para la serie de* `infectados`



En la primera gráfica, correspondiente a la función de autocorrelación, se observó (aproximadamente) que, para valores en la abscisa horizontal mayores que 5, los segmentos de líneas verticales tienden a estar dentro de la franja delimitada por las líneas punteadas azules, para lo cual, se concluyó que es apropiado fijar el valor de q en 5. De manera similar se puede hacer el análisis para la gráfica de la función de autocorrelación parcial, concluyendo que es apropiado fijar el valor p en 5. Por lo tanto, el modelo base que se tuvo para este estudio es el $ARIMA(5,1,5)$.

Se observó lo siguiente: si se usara el modelo $ARIMA(5,1,5)$ para ajustar la serie de tiempo usando la función `Arima()` del paquete `forecast`, la imputación de los datos faltantes del 2020 al 2021 de la serie de tiempo y la predicción para el año 2022 con función `forecast()` del paquete `forecast` y un tamaño de predicción de $h = 52 * 3$, la gráfica de la predicción se vería de la siguiente forma:

Figura 4. *Gráfica de la predicción del modelo ARIMA(5,1,5)*



Donde la línea azul representa la imputación de los valores faltantes del 2020 al 2021, y la gráfica roja es corresponde a la predicción para el año 2022. Es claro que el modelo no sería el apropiado para modelar la serie de tiempo puesto que el comportamiento de la predicción dista mucho del ajuste para los datos históricos. Esto se debió fundamentalmente a que el modelo no captó el comportamiento estacional de la serie. Para la creación de un modelo que tenga mejor ajuste considerando esta última condición, se buscó complementar el modelo $ARIMA(5,1,5)$ a uno que tuviera parte estacional, como es el caso de los modelos ARIMA multiplicativos $ARIMA(5,1,5) \times (P,D,Q)_s$.

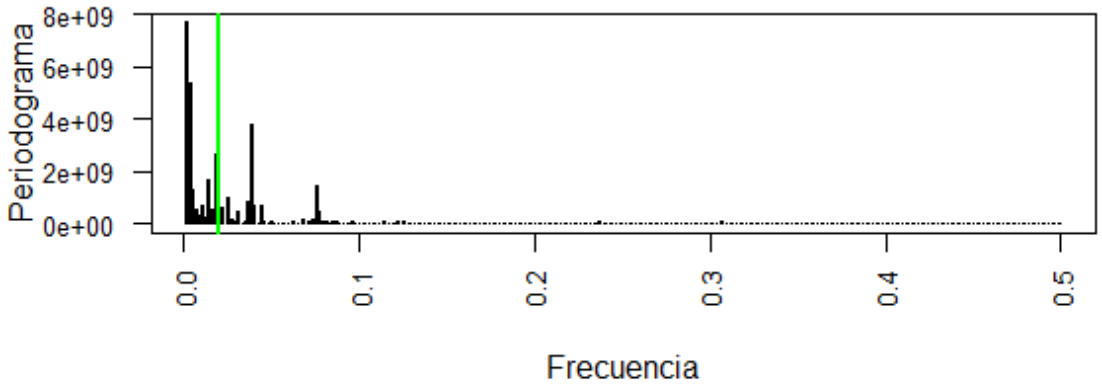
6.3 Escogencia de los Modelos Multiplicativos ARIMA(5,1,5) × (P,D,Q)$_s$

A continuación, se muestra el procedimiento que se llevó a cabo a fin de escoger el modelo que mejor se justara a la serie de tiempo, usando el criterio de Akaike para escoger los 5 modelos con los parámetros P, D, Q más apropiados y el cuadrado medio del error para determinar cual tenía mejor ajuste a los primeros 30 datos históricos del año 2022.

6.3.1 Determinación del Periodo y la Frecuencia de la Serie a Estudiar

Para determinar el periodo y la frecuencia de la serie de tiempo se hizo uso de su periodograma el cual se genera con la función `periodogram()` del paquete `stats`.

Figura 5. *Periodograma de la serie de tiempo de la variable* `infectados`



El gráfico muestra las posibles frecuencias a las que ocurren los ciclos o patrones y la escogencia de la más apropiada se hace teniendo en cuenta los valores donde la altura de la barra vertical no sea, ni muy alta, ni muy baja. Ya que el periodo es uno sobre la frecuencia escogida ($f = 0.02$ donde se marca la línea en color verde) se consideró que la serie de tiempo tenía un periodo de

$$s = \frac{1}{0.02} = 50.$$

6.3.2 Determinación de los Parámetros Estacionales P, D, Q del Modelo Multiplicativo

En el estudio se tuvieron en cuenta modelos ARIMA multiplicativos de la forma $ARIMA(5,1,5) \times (P, D, Q)_s$ los cuales se escogieron asignando diferentes valores enteros a P, D y Q entre 1 a 4 y calculando el valor del AIC que tenían (criterio de Akaike). No se tomaron en

cuenta valores más grandes para estos parámetros puesto que se buscó que el modelo fuera el mejor posible con la menor cantidad de parámetros o más parsimonioso.

Algunos de estos modelos tuvieron que ser excluidos puesto que el procedimiento de estimación utilizado no pudo converger hacia una solución adecuada y se imprimía una salida como la siguiente:

Imagen 2. *Salida en R al procesar un modelo divergente*

```
> ARIMA515_111= Arima(serie_arima_ira, order = c(5, 1, 5),
seasonal = list(order = c(1, 1, 1), period = 50))
Error in optim(init[mask], armafn, method = optim.method, hessian = TRUE,  :
non-finite finite-difference value [1]

      >
```

Para los demás modelos, en los cuales el método sí pudo converger, se aplicó el criterio de Akaike (función `AIC()` del paquete base de R) el cual determina el grado de ajuste que tienen con la serie de tiempo en los datos históricos desde el año 2010 hasta el año 2019. Finalmente se seleccionaron únicamente los que tuvieron los cinco mejores índices los cuales aparecen a continuación:

Tabla 1. *Modelos ARIMA(p,d,q) × (P,D,Q)$_s$ escogidos según su AIC*

MODELO	$ARIMA(5,1,5) \times (1,3,1)_{50}$	$ARIMA(5,1,5) \times (2,3,1)_{50}$	$ARIMA(5,1,5) \times (2,2,2)_{50}$	$ARIMA(5,1,5) \times (4,2,4)_{50}$	$ARIMA(5,1,5) \times (2,3,4)_{50}$
AIC	7819.268	7797.45	8477.445	7789.054	7789.054

Ya que la variable a estudiar en este trabajo es `infectados`, que depende del tiempo t, para una cantidad total de 30 datos que es la cantidad de datos reales conocidos del 2022, el valor MSE estaría dado por:

$$MSE = \frac{1}{30}\sum_{t=624}^{654}\left[infectados_t - \widehat{infectados}_t\right]^2.$$

Donde la suma inicia desde $t = 624$ que corresponde a la primera semana epidemiológica del año 2022 y $t = 654$ la 30 del mismo año.

6.3.3 Determinación del Mejor Modelo Según el MSE

En la siguiente tabla se resumen los índices MSE para cada uno de los modelos ARIMA escogidos anteriormente:

Tabla 2. *Modelos ARIMA(p,d,q) × (P,D,Q)$_s$ escogidos y sus respectivos índices MSE*

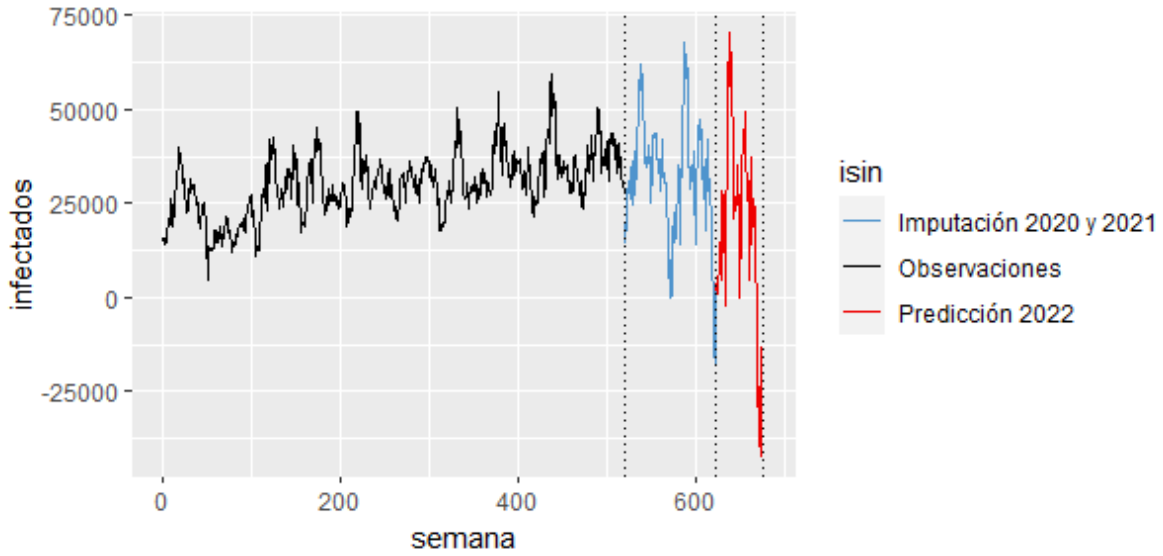
MODELO	$ARIMA(5,1,5)$ × (1,3,1)	$ARIMA(5,1,5)$ × (2,3,1)	$ARIMA(5,1,5)$ × (2,2,2)	$ARIMA(5,1,5)$ × (4,2,4)	$ARIMA(5,1,5)$ × (2,3,4)
MSE	222763579	155409647	192926639	183292777	179060509

Como se observa en la tabla anterior, el índice más pequeño corresponde al modelo $ARIMA(5,1,5) \times (2,3,1)_{50}$ con un valor MSE de 155409647.

6.4 Análisis Descriptivo del Modelo con Mejor Ajuste ARIMA(5,1,5) × (2,3,1)$_{50}$

Luego de escoger al modelo $ARIMA(5,1,5) \times (2,3,1)_{50}$ con mejor ajuste a los primeros datos reales del 2022, se dio paso a imputar los datos faltantes del 2020 y 2021 de la serie de tiempo con un tamaño de predicción de $h = 52 * 3$. El motivo por el cual se quiso hacer la predicción del 2022 fue para tener como referencia los datos reales de ese año y así reconocer que tan bien empataba el modelo con los datos, antes y después del vacío de información del 2020 y 2021.

Figura 6. *Gráfica del modelo ARIMA con mejor ajuste: ARIMA(5,1,5) × $(2,3,1)_{50}$ y AIC = 7797.4*



La línea azul representa los datos imputados según el modelo de mejor ajuste y la línea roja la predicción del año 2022. Las gráficas de los demás modelos se encuentran en la sección de gráficas.

7. Conclusiones

Según los objetivos planteados en esta investigación y luego de interpretar los resultados obtenidos, se concluyen lo siguiente:

Los datos faltantes de la variable `infectados` del año 2020 y 2021 pudieron ser imputados a través del modelo $ARIMA(p,d,q) \times (2,3,1)_{50}$ el cual tuvo el mejor ajuste a los datos disponibles del 2022 según el estimador MSE. Se puede afirmar que, en general, los modelos ARIMA permiten hacer imputación de datos faltantes para prolongados periodos de tiempo con información desconocida.

Dada la naturaleza de la variable `infectados` es coherente que el modelo ARIMA escogido para hacer la imputación y predicción de los datos faltantes tuviera una componente estacional. Está ampliamente documentado que las infecciones respiratorias agudas reaparecen cada tanto tiempo, donde factores externos como la temperatura, las condiciones climatológicas en algunos meses del año o el número máximo de habitantes en una población determinan que la cantidad de infectados por infección respiratoria aguda como por ejemplo en el caso de los virus, reaparezcan periódicamente en algunos meses específicos del año (ver (López-Perea, Méndez, López-Cuarado, Cámara, & de Mateo Ontañón, 2011, pág. 153)).

A pesar del poder predictivo de los modelos $ARIMA(p,d,q) \times (P,D,Q)_s$ para el estudio de las series de tiempo estacionales y del estimador MSE para la escogencia del modelo más optimo, estos modelos no se pueden considerar perfectos ni tampoco pueden predecir con total exactitud los valores reales de la serie a estudiar. Aun así, los modelos ARIMA en general siguen siendo muy usados para modelar series de tiempo y se sigue trabajando en su implementación.

No siempre los métodos para crear modelos $ARIMA(p,d,q) \times (P,D,Q)_s$ van a converger (ver *Imagen 2*). Esto puede ocurrir debido a que la función usada para estimar los parámetros del modelo ARIMA está basada en la función `optim` (ver (Truong, Nguyen, Truong, Ho, & Diem-Chinh, 2019)) que hace uso de métodos numéricos, como lo es el *método de Brent* para el cual no siempre se tiene garantizada la existencia de los valores que optimizan la función. (Ver (Gegenfurtner, 1992)).

Tener un buen ajuste de los datos históricos no necesariamente implica que se realicé una mejor predicción. Para este estudio los modelos $ARIMA(5,1,5) \times (4,2,4)_{50}$ y $ARIMA(5,1,5) \times (2,3,4)_{50}$ presentaron un mejor valor AIC que el modelo y $ARIMA(5,1,5) \times (2,3,1)_{50}$, es decir, que tuvieron un mejor ajuste en general de toda la serie, y sin embargo el estimador MSE mostró que el modelo más apropiado para la predicción del año 2022 era este último.

La gráfica de la imputación de los datos del 2020 al 2021 y gráfica de la predicción de los datos para el 2022 muestran un comportamiento periódico, similar a la gráfica para los datos históricos, con un periodo de $s = 50$ semanas. En particular, el modelo predice que aproximadamente cada año (casi 52 semanas) se tendrá un pico de contagios por infección respiratoria aguda. Por otro lado, ya que la predicción para el año 2022 no tiene comportamiento monótono (curva creciente o decreciente), se concluye que el modelo no muestra un aumento precipitado del número de casos por infección respiratoria aguda para el año 2022 y, por lo tanto, no dio alerta de una posible pandemia para ese año como realmente ocurrió.

Referencias Bibliográficas

Ahlberg, J., Nilson, E., & Walsh, J. (1967). *The Theory Of Splines and Their Applications.* Mathematics in Science and Engineering.

Alfaki, M. M., & Masih, S. B. (2015). Modeling and forecasting by using time series ARIMA models. *International Journal of Engineering Research & Technology*, 4(3).

Arciniegas Paspuel, O. G., Castro Morales, L. G., & Arias Collaguazo, W. M. (2021). Análisis y predicción de la recaudación tributaria en el Ecuador ante la COVID-19, aplicando el modelo ARIMA. *Dilemas contemporáneos: educación, política y valores*.

Balakrishnan, N. (2010). *Methods and Applications of Statistics in Business, Finance, and Management Science.* Wiley.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, *31*(3), 307-327.

Brockwell, P. J., & A, R. (2002). Introduction to time series and forecasting.

Casimiro, M. P. (2009). *Análisis de Series Temporales. Modelos ARIMA.* (SARRIKO-ON, Ed.) País Vasco. Obtenido de https://addi.ehu.es/handle/10810/12492

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical associatio*, 596-610.

Fernando Medina, M. G. (2007). *Imputación de datos: teoría y práctica.* Santiago de Chile.

Gegenfurtner, K. R. (1992). PRAXIS: Brent's algorithm for function minimization. *24*, 560-564.

Harvey, A. C. (1990). Forecasting, structural time series models and the Kalman filter. *Cambridge university press*.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. *2*.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice.

Lebrato, L. T., Paradela, T. M., Lebrato, L. T., & Portales, Z. V. (2018). Promoción y educación para la salud en la prevención de las infecciones respiratorias agudas. *Humanidades Médicas*, 122-136.

Lima, P. d.-i. (2020). Córdova Sotomayor, Daniel Angel and Chávez Bacilio, Clara Guadalupe and Bermejo Vargas, Elisabet Winiferson and Jara Ccorahua, Ximena Nicole and Santa Maria Carlos, Flor Benigna. *Horizonte Médico (Lima)*, 54-60.

López-Perea, N., Méndez, L., López-Cuarado, T., Cámara, A., & de Mateo Ontañón, S. (2011). Estimación de la mortalidad atribuible a gripe estacional en España. *Boletín epidemiológico semanal, 19(11)*, 150-158.

Reisen, V., C, L.-L., MS, & Taqqu. (2017). An m-estimator for the long-memory parameter. *Journal of Statistical Planning and Inference, 187*, 44–55. doi:10.1016/j.jspi.2017.02.008

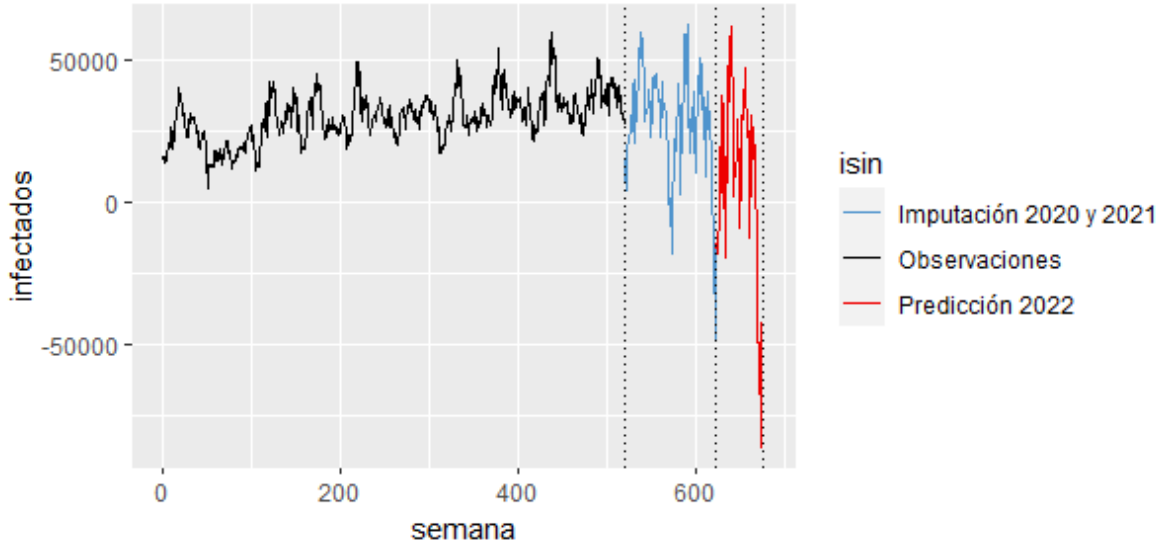
Team, R. C. (2020). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Obtenido de https://www.R-project.org/

Truong, Nguyen, B.-C. a., Truong, V.-B. a., Ho, H.-V. a., & Diem-Chinh, T. (2019). Comparison of optim, nleqslv and MaxLik to estimate parameters in some of regression models. *Journal of Advanced Engineering and Computation*, *3*(4), 532--550.

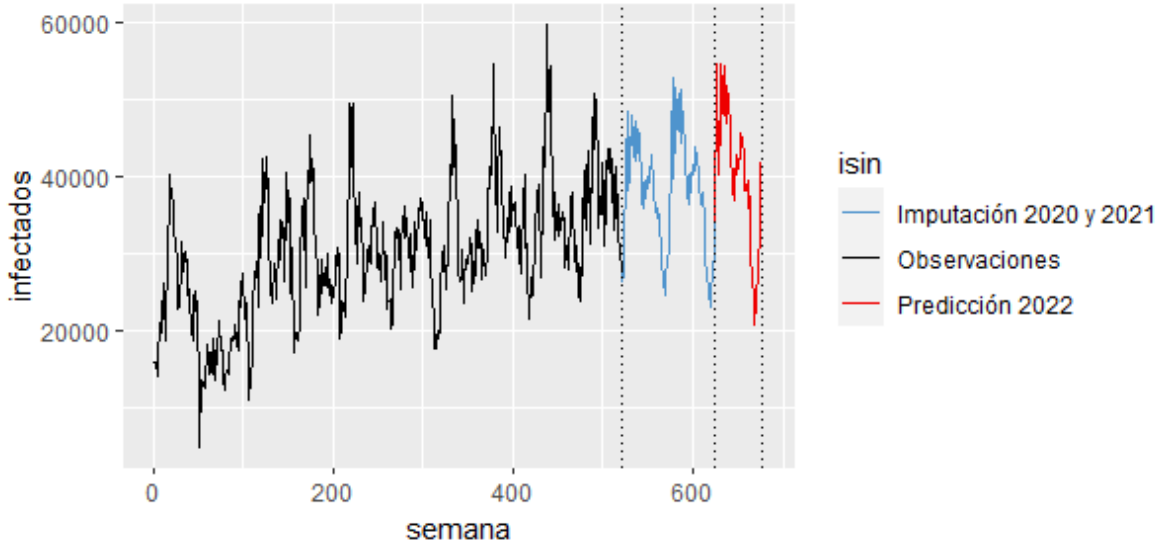
Apéndices

Apéndice A. Gráficas de los Modelos ARIMA(p,d,q) × $(P,D,Q)_s$ y sus AIC

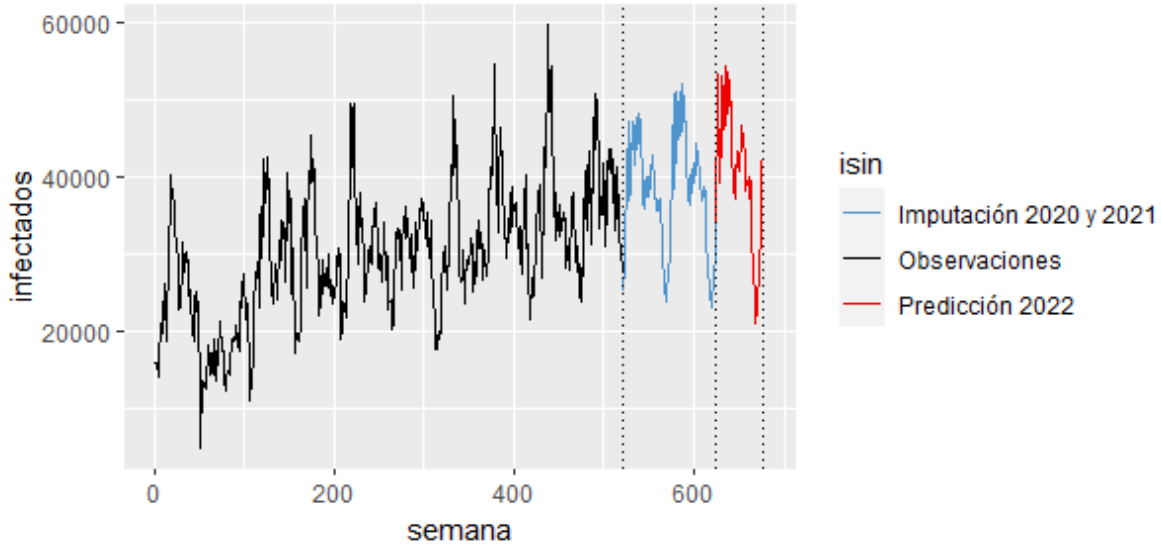
Gráfica del modelo ARIMA(5,1,5) × $(1,3,1)_{50}$ con AIC = 7819.268



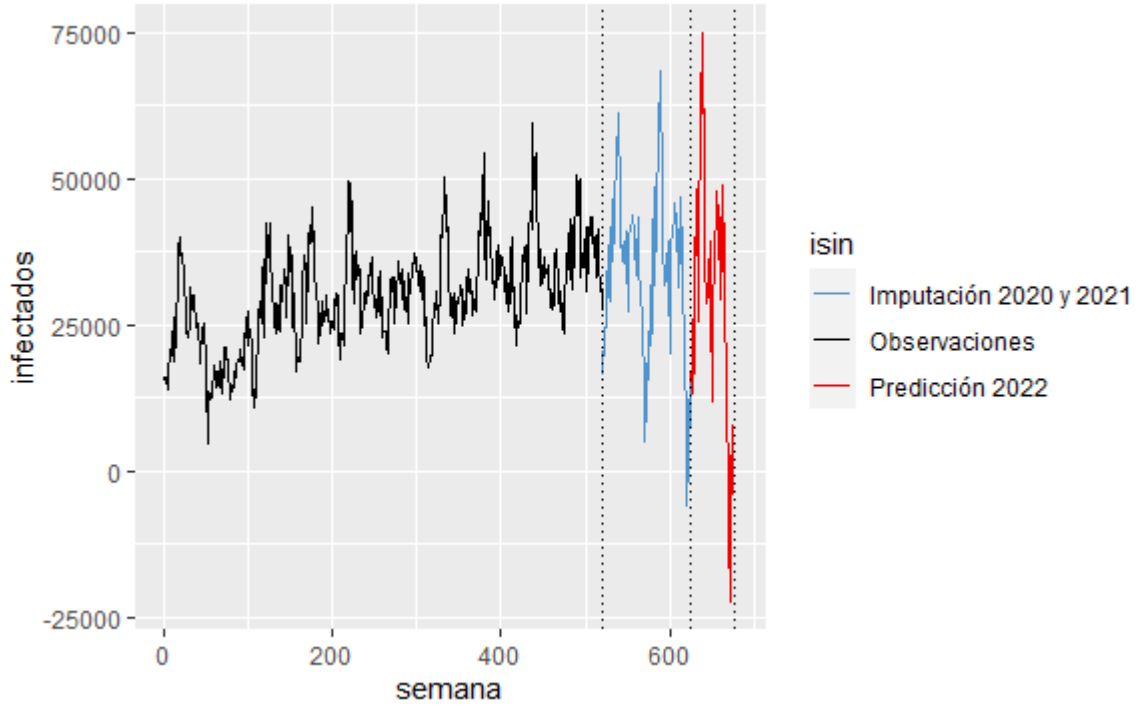
Gráfica del modelo ARIMA(5,1,5) × $(2,2,2)_{50}$ con AIC = 8477.445



Gráfica del modelo ARIMA(5,1,5) × $(4,2,4)_{50}$ con AIC = 8488.202



Gráfica del modelo $ARIMA(5,1,5) \times (2,3,4)_{50}$ con AIC = 7789.054



Apéndice B. Código Usado en R

```
install.packages("tidyverse")
```

```
install.packages("vars")
library(readxl)
library(epitools)
library(imputeTS)
library(forecast)
library(TSA)
library(ggplot2)
library(dplyr)
library(tidyverse)
library(vars)
library(aTSA)
datos_IRA <- read_excel("F:\\TG\\datos IRA.xlsx")
datos_IRA = as.matrix(datos_IRA)
datos_IRA <- datos_IRA[, -1]
semana = c(1:572)
infectados = c(datos_IRA[1,], datos_IRA[2,], datos_IRA[3,], datos_IRA[4,],
datos_IRA[5,], datos_IRA[6,], datos_IRA[7,], datos_IRA[8,], datos_IRA[9,],
datos_IRA[10,], datos_IRA[11,])
datos_IRA1df = data.frame(semana,infectados)
ggplot(datos_IRA1df, aes(x=semana, y=infectados)) + geom_line()
casos_IRA = infectados
año_epidemiologico = c(casos_IRA[1:(5*52)], NA,casos_IRA[(5*52 + 1):(10*52)],
rep(NA, (2*52 + 1)), casos_IRA[(10*52 + 1):(10*52 + 30)],rep(NA, (24)))
st = data.frame(c(1:680),año_epidemiologico[1:680])
colnames(st) = c("semana","infectados")
ggplot(st, aes(x=semana , y=infectados)) + geom_line() serie_arima_ira =
ts(año_epidemiologico[1:521])
interpolation(serie_arima_ira, "spline")

### Prueba de Dickey-Fuller
adf.test(serie_arima_ira)
```

```
ndiffs(serie_arima_ira) # d=1
acf(diff(serie_arima_ira)) # q = 5
pacf(diff(serie_arima_ira)) # p = 5
ajuste_arima_ira = Arima(serie_arima_ira, order = c(5, 1, 5))
prediccion = data.frame(forecast(ajuste_arima_ira, h = 52*3))[,1]
length(prediccion)
length(st$semana)
length(st$semana[520:680])
df1 = data.frame(semana = st$semana[1:520], infectados =
st$infectados[1:520], isin = "Observaciones")
df2 = data.frame(semana = st$semana[520:624], infectados = prediccion[1:105]
, isin = "Imputación 2020 y 2021")
df3 = data.frame(semana = st$semana[(52*12):680], infectados =
prediccion[105:161] , isin = "Predicción 2022")
df = rbind(df1, df2, df3)
### GRAFICA DE LA IMPUTACIÓN Y DE LA PREDICCIÓN CON EL MODELO ARIMA(5,1,5) NO
ESTACIONAL
ggplot(df, aes(x = semana, y = infectados, color = isin)) + geom_line() +
geom_vline(xintercept = (52*10), linetype = 3, color = 1) +
geom_vline(xintercept = (52*12),linetype = 3, color = 1)+
geom_vline(xintercept = (52*13),linetype = 3, color = 1) +
scale_colour_manual(values=c(Observaciones='black', "Imputación 2020 y
2021"='#4F94CD', "Predicción 2022"='#EE0000'))
periodogram(serie_arima_ira, xlab = "Frecuencia", ylab = "Periodograma",las
=2, cex.axis = 0.8)
abline(v = 0.02, lwd = 2, col = "#00FF00")
periodo = 1/0.02
periodo
### ESCOGENCIA DE LOS MODELOS ARIMA MULTIPLICATIVOS Y SU AIC
##ARIMA515_131
```

```
#ARIMA515_131_2 = Arima(serie_arima_ira, order = c(5, 1, 5), seasonal =
list(order = c(1, 3, 1), period = 50))
#AIC(ARIMA515_131_2)
load("F:\\TG\\ARIMA515_131_2.rda")
prediccion_131 = data.frame(forecast(ARIMA515_131_2, h = 52*3))[,1]
#upper<-prediccion_131+2*sd(prediccion_131)
#lower<-prediccion_131-2*sd(prediccion_131)
df1 = data.frame(semana = st$semana[1:520], infectados =
st$infectados[1:520], isin = "Observaciones")
df2_131 = data.frame(semana = st$semana[520:624], infectados =
prediccion_131[1:105] , isin = "Imputación 2020 y 2021")
df3_131 = data.frame(semana = st$semana[(52*12):680], infectados =
prediccion_131[105:161] , isin = "Predicción 2022")
#h = data.frame(semana = st$semana[(52*12):680], infectados = upper[105:161]
,isin = "upper")
#l = data.frame(semana = st$semana[(52*12):680], infectados = lower[105:161],
isin = "lower")
df_131 = rbind(df1, df2_131, df3_131)
ggplot(df_131, aes(x = semana, y = infectados, color = isin)) + geom_line() +
geom_vline(xintercept = (52*10), linetype = 3, color = 1) +
geom_vline(xintercept = (52*12),linetype = 3, color = 1)+
geom_vline(xintercept = (52*13),linetype = 3, color = 1) +
scale_colour_manual(values=c(Observaciones='black', "Imputación 2020 y
2021"='#4F94CD', "Predicción 2022"='#EE0000'))
#save(ARIMA515_131_2, file = "D:\\TG\\ARIMA515_131_2.rda")
dif_131 = casos_IRA[(10*52 + 1):(10*52 + 30)]-prediccion_131[1:30]
cuad_131 = (dif_131)^2
sum_131 = sum(cuad_131)
MSE_131=(1/30)*(sum_131)
MSE_131
##ARIMA515_231
```

```
#ARIMA515_231 = Arima(serie_arima_ira, order = c(5, 1, 5), seasonal =
list(order = c(2, 3, 1), period = 50))
#AIC(ARIMA515_231)
load("F:\\TG\\ARIMA515_231.rda")
prediccion_231 = data.frame(forecast(ARIMA515_231, h = 52*3))[,1]
df1 = data.frame(semana = st$semana[1:520], infectados =
st$infectados[1:520], isin = "Observaciones")
df2_231 = data.frame(semana = st$semana[520:624], infectados =
prediccion_231[1:105] , isin = "Imputación 2020 y 2021")
df3_231 = data.frame(semana = st$semana[(52*12):680], infectados =
prediccion_231[105:161] , isin = "Predicción 2022")
df_231 = rbind(df1, df2_231, df3_231)
ggplot(df_231, aes(x = semana, y = infectados, color = isin)) + geom_line() +
geom_vline(xintercept = (52*10), linetype = 3, color = 1) +
geom_vline(xintercept = (52*12),linetype = 3, color = 1)+
geom_vline(xintercept = (52*13),linetype = 3, color = 1) +
scale_colour_manual(values=c(Observaciones='black', "Imputación 2020 y
2021"='#4F94CD', "Predicción 2022"='#EE0000'))
#save(ARIMA515_231, file = "D:\\TG\\ARIMA515_231.rda")
dif_231 = casos_IRA[(10*52 + 1):(10*52 + 30)]-prediccion_231[1:30]
cuad_231 = (dif_231)^2
sum_231 = sum(cuad_231)
MSE_231=(1/30)*(sum_231)
MSE_231
##ARIMA515_222
#ARIMA515_222 = Arima(serie_arima_ira, order = c(5, 1, 5), seasonal =
list(order = c(2, 2, 2), period = 50))
#AIC(ARIMA515_222)
load("F:\\TG\\ARIMA515_222.rda")
prediccion_222 = data.frame(forecast(ARIMA515_222, h = 52*3))[,1]
```

```
df1 = data.frame(semana = st$semana[1:520], infectados =
st$infectados[1:520], isin = "Observaciones")
df2_222 = data.frame(semana = st$semana[520:624], infectados =
prediccion_222[1:105] , isin = "Imputación 2020 y 2021")
df3_222 = data.frame(semana = st$semana[(52*12):680], infectados =
prediccion_222[105:161] , isin = "Predicción 2022")
df_222 = rbind(df1, df2_222, df3_222)
ggplot(df_222, aes(x = semana, y = infectados, color = isin)) + geom_line() +
geom_vline(xintercept = (52*10), linetype = 3, color = 1) +
geom_vline(xintercept = (52*12),linetype = 3, color = 1)+
geom_vline(xintercept = (52*13),linetype = 3, color = 1) +
scale_colour_manual(values=c(Observaciones='black', "Imputación 2020 y
2021"='#4F94CD', "Predicción 2022"='#EE0000'))
#save(ARIMA515_222, file = "D:\\TG\\ARIMA515_222.rda")
dif_222 = casos_IRA[(10*52 + 1):(10*52 + 30)]-prediccion_222[1:30]
cuad_222 = (dif_222)^2
sum_222 = sum(cuad_222)
MSE_222=(1/30)*(sum_222)
MSE_222
##ARIMA515_424
#ARIMA515_424 = Arima(serie_arima_ira, order = c(5, 1, 5), seasonal =
list(order = c(4, 2, 4), period = 50))
#AIC(ARIMA515_424)
load("F:\\TG\\ARIMA515_424.rda")
prediccion_424 = data.frame(forecast(ARIMA515_424, h = 52*3))[,1]
df1 = data.frame(semana = st$semana[1:520], infectados =
st$infectados[1:520], isin = "Observaciones")
df2_424 = data.frame(semana = st$semana[520:624], infectados =
prediccion_424[1:105] , isin = "Imputación 2020 y 2021")
df3_424 = data.frame(semana = st$semana[(52*12):680], infectados =
prediccion_424[105:161] , isin = "Predicción 2022")
```

```
df_424 = rbind(df1, df2_424, df3_424)
ggplot(df_424, aes(x = semana, y = infectados, color = isin)) + geom_line() +
geom_vline(xintercept = (52*10), linetype = 3, color = 1) +
geom_vline(xintercept = (52*12),linetype = 3, color = 1)+
geom_vline(xintercept = (52*13),linetype = 3, color = 1) +
scale_colour_manual(values=c(Observaciones='black', "Imputación 2020 y
2021"='#4F94CD', "Predicción 2022"='#EE0000'))
#save(ARIMA515_424, file = "D:\\TG\\ARIMA515_424.rda")
dif_424 = casos_IRA[(10*52 + 1):(10*52 + 30)]-prediccion_424[1:30]
cuad_424 = (dif_424)^2
sum_424 = sum(cuad_424)
MSE_424=(1/30)*(sum_424)
MSE_424
##ARIMA515_234
#ARIMA515_234 = Arima(serie_arima_ira, order = c(5, 1, 5), seasonal =
list(order = c(2, 3, 4), period = 50))
#AIC(ARIMA515_234)
load("F:\\TG\\ARIMA515_234.rda")
prediccion_234 = data.frame(forecast(ARIMA515_234, h = 52*3))[,1]
df1 = data.frame(semana = st$semana[1:520], infectados =
st$infectados[1:520], isin = "Observaciones")
df2_234 = data.frame(semana = st$semana[520:624], infectados =
prediccion_234[1:105] , isin = "Imputación 2020 y 2021")
df3_234 = data.frame(semana = st$semana[(52*12):680], infectados =
prediccion_234[105:161] , isin = "Predicción 2022")
df_234 = rbind(df1, df2_234, df3_234)
ggplot(df_234, aes(x = semana, y = infectados, color = isin)) + geom_line() +
geom_vline(xintercept = (52*10), linetype = 3, color = 1) +
geom_vline(xintercept = (52*12),linetype = 3, color = 1)+
geom_vline(xintercept = (52*13),linetype = 3, color = 1) +
```

```
scale_colour_manual(values=c(Observaciones='black', "Imputación 2020 y
2021"='#4F94CD', "Predicción 2022"='#EE0000'))
#save(ARIMA515_234, file = "D:\\TG\\ARIMA515_234.rda")
dif_234 = casos_IRA[(10*52 + 1):(10*52 + 30)]-prediccion_234[1:30]
cuad_234 = (dif_234)^2
sum_234 = sum(cuad_234)
MSE_234=(1/30)*(sum_234)
MSE_234
#Modelo ARIMA con el mejor MSE
MSE_min = min(MSE_131,MSE_222,MSE_231,MSE_234,MSE_424)
MSE_min
```