Universidad de Costa Rica

CA-0404 Modelos Lineales

Anteproyecto de investigación

Pronóstico demográfico de Costa Rica

Profesor
Luis Barboza Chinchilla

Autores
David Zumbado Fernández
Leonardo Blanco Villalobos
Ignacio Barrantes Valerio

9 de diciembre de 2022

Indice de contenidos

Listado de Figuras

Listado de Tablas

Introducción

Rees (2020) afirma que una de las mayores contribuciones de la demografía ha sido el desarrollo de modelos para la proyección de futuras poblaciones, el cual es un tema de gran importancia para cualquier país, pues impacta directamente en sectores tan diversos como lo pueden ser la planificación urbana, el desarrollo humano sostenible, la demanda de bienes y servicios o los sistemas de pensiones. Además, esta permite estimar indicadores tan relevantes como la esperanza de vida.

Dentro de las proyecciones poblacionales se puede encontrar la proyección de la mortalidad, que está asociada directamente a las defunciones ocurridas en una población. En este trabajo se busca contestar la siguente pregunta de investigación: ¿Cómo se puede realizar un pronóstico de la serie de defunciones totales anuales de Costa Rica? Como puede notarse, la pregunta está totalmente circunsctira en el marco de la proyección poblacional.

De forma más específica, en este trabajo se busca realizar un pronóstico de la cantidad de defunciones totales de Costa Rica para una ventana de tiempo de dos años (2019 y 2020) utilizando datos de 1950 al 2018. Para llevar a cabo este objetivo se consideran modelos ARIMA y modelos lineales generalizados, que son un caso particular de los modelos de espacio-estado.

Marco teórico

En términos generales, la demografía es “una ciencia que estudia las poblaciones humanas, su dimensión, estructura, evolución y características generales” (Bruno Ramírez, 2019) y además “estudia estadísticamente la estructura y la dinámica de las poblaciones, así como los procesos concretos que determinan su formación, conservación y desaparición […] tales procesos son los de fecundidad, mortalidad y migración: emigración e inmigración” (Bruno Ramírez, 2019).

De acuerdo al mismo autor, hay dos grandes tipos de demografía: la estática, que “estudia las poblaciones humanas en un momento de tiempo determinado desde un punto de vista de dimensión, territorio, estructura y características estructurales” (2019) y la dinámica, “que estudia las poblaciones humanas desde el punto de vista de la evolución en el transcurso del tiempo y los mecanismos por los que se modifica la dimensión, estructura y distribución geográfica de las poblaciones” (2019).

Bajo estas concepciones, el presente trabajo se enmarca en la demografía dinámica, pues se busca realizar

pronóstico sobre una serie histórica de defunciones. El pronóstico consiste en tomar modelos ajustados a datos históricos y utilizarlos para predecir observaciones futuras. Asimismo, una serie cronológica o srie de tiempo se le llama a "la que forman los valores sucesivos que una cierta variable ha tomado en el transcurso del tiempo" (Macció et al., 1985). Además, según Macció et al. (1985) la mortalidad o defunción se emplea para expresar la acción de la muerte sobre la población y esta se mide en valores absolutos y sobre el año calendario.

Ahora bien, las teorías demográficas "pretenden explicar los patrones del crecimiento de la población en diversos países del mundo dando cuenta de la estructura y la dinámica de la población y estableciendo leyes o principios que regirían esos fenómenos" (Bruno Ramírez, 2019). Más específicamente, una teoría demográfica "explicaría los cambios y acontecimientos de las poblaciones humanas, de su dimensión, estructura, evolución y características generales, tanto desde un punto de vista cuantitativo (estadístico) como cualitativo (biológico, sociológico, cultural y económico)" (Bruno Ramírez, 2019).

La primer teoría global de la población, es la malthusiana, enunciada por Thomas Robert Malthus en 1978 en su Ensayo sobre el principio de población, que consistió "en una teoría donde se establece una ley general que explica el crecimiento total de la población en relación con otra variable fuera del contexto social, como lo es la disponibilidad de alimentos" (Sarrible, 2009). A su vez, explica que la teoría malthusiana propone que "el ritmo de crecimiento de la población responde a una progresión geométrica, mientras que el ritmo de aumento de los recursos para su supervivencia lo hace en progresión aritmética" (Bruno Ramírez, 2019).

De acuerdo a Mariscal de Gante & Rodríguez (2018), una de las teorías demográficas más importantes es la Teoría de la Transición Demográfica (TTD). Esta consiste en una generalización empírica en función de observaciones pasadas y establece una conexión entre la evolución demográfica de la población y el crecimiento económico (Mariscal de Gante & Rodríguez, 2018).

El mismo autor explica que bajo el régimen demográfico, la TTD da pie a los modelos demográficos más importantes y esta "se produce cuando la natalidad y la mortalidad, o por lo menos uno de los dos fenómenos, ha dejado sus elevados niveles tradicionales para dirigirse hacia porcentajes más bajos, asociados a la fecundidad dirigida y al uso de métodos de lucha contra la natalidad, pasando de una demografía antigua y tradicional a otra moderna" (Alcalde, 2010).

La Teoría de la transición epidemiológica (TTE) surge como alternativa teórica de la ya menciona Teoría de transición demográfica (TTD) debido a la naturaleza descriptiva o evolucionista de esta. La TTE tiene un carácter más multidisciplinar y con esto multifactorial.

Esta teoría establece que los principales factores causantes de las transiciones demográficas son los factores ecobiológicos de la mortalidad, factores biopsicológicos, como el uso de contraconceptivos ante el aumento de la supervivencia infantil generando un aumento de la EV, factores médicos y de salud pública, factores psicológicos o emocionales, y factores socioeconómicos, ya que los desarrollos económicos establecen los primeros sistemas sanitarios, ayudando así al descenso de la mortalidad y la reducción de la incidencia de las enfermedades infecciosas.

Es importante mencionar que tal como lo establece Mariscal de Gante & Rodríguez (2018) la TTE, tiene como objeto de estudio la mortalidad, ya que se centra en aspectos como: patrones de enfermedad, causas de las muertes y la interacción de estas con patrones demográficos, económicos y sociológicos.

Como antecedentes, se observa que en múltiples investigaciones con temáticas relacionadas a defunciones como el de Adekanmbi et al. (2014) y el estudio por Ordorica (2004), se implementan modelos para el pronóstico de series defunciones mediante distintos enfoques, entre ellos pronóstico mediante el empleo de modelos ARIMA, sin embargo estudios más recientes hacen uso de los Modelos Dinámicos Lineales.

Tal como lo establece Petris et al. (2007) entre los modelos más utilizados para el análisis de series temporales se encuentra la clase de modelos de media móvil autorregresiva (ARMA). Para enteros, no negativos p y q, un modelo ARMA(p, q) es definido mediante la notación:

$$Y_t = \mu + \sum_{j=1}^{p} \phi_j (Y_{t-j} - \mu) + \sum_{j=1}^{p} \psi_j \epsilon_{t-j} + \epsilon_t$$

Donde (ϵ_t) es una ruido blanco Gaussiano con varianza σ_ϵ^2 y los parámetros $\phi_1, \phi_2, ..., \phi_p$ satisfacen la condición de estacionariedad.

Cuando los datos no presentan estacionariedad, se suele tomar las diferencias hasta que se obtenga esta, una vez obtenida se procede a ajustar el modelo ARMA(p, q) a los datos diferenciados.

Un modelo para un proceso cuya d-ésima diferencia sigue un modelo ARMA(p, q) es llamado un ARIMA(p, d, q).

Tal como lo establece Petris et al. (2007), los modelos dinámicos lineales son una clase de Modelos de Espacio-Estado también llamados Modelos de Espacio-Estado Lineales Gaussianos, estos poseen dos supuestos, la linealidad y el supuesto de distribuciones Gaussianas. Petris et al. (2007) señala que este último supuesto puede ser justificado mediante argumentos del teorema del límite central.

En general, el problema de pronóstico de k-pasos hacia adelante consiste en estimar la evolución del sistema θ_{t+k} para $k \geq 1$ y realizar un pronóstico de k-pasos para Y_{t+k}.

Según Petris et al. (2007) en los DLM, el filtro de Kalman proporciona las fórmulas para actualizar nuestra inferencia actual sobre el vector de estado conforme se disponga de nuevos datos.

Para un DLM, si se cumple que:

$$\theta_t | \mathcal{D}_t \sim \mathcal{N}(m_t, C_t), t \geq 1$$

Se tiene que:

La densidad de predicción de estado de k-pasos con $k \geq 1$ hacia adelante de θ_{t+k} dada la información pasada D_t es Gaussiana con media y varianza condicional dadas respectivamente por $a_t(k) = E[\theta_{t+k}|D_t] = G_{t+k} a_{t,k-1}$ y $R_t(k) = Var[\theta_{t+k}|D_t] = G_{t+k} R_{t,k-1} G_{t+k}^{'} + W_{t+k}$.

La densidad de predicción de k-pasos con $k \geq 1$ hacia adelante de Y_{t+k} dada la información pasada D_t, es Gaussiana con media y varianza condicional dadas respectivamente por $f_t(k) = E[Y_{t+k}|D_t] = F_{t+k} a_t(k)$ y $Q_t(k) = Var[Y_{t+k}|D_t] = F_{t+k} R_t(k) F_{t+k}^{'} + V_{t+k}$.

El DLM polinomial de orden 1 es descrito por el par de ecuaciones:

$$\begin{aligned} Y_t &= \mu_t + v_t, \qquad v_t \sim \mathcal{N}(0, V) \\ \mu_t &= \mu_{t-1} + w_t, \quad w_t \sim \mathcal{N}(0, W) \end{aligned}$$

donde (Y_t) es, para efectos del presente análisis, el proceso del total de defunciones anuales de Costa Rica.

Por otro lado, el modelo DLM polinomial de orden 2 se describe mediante las siguientes ecuaciones:

$$\begin{aligned} Y_t &= \theta_{t,1} + v_t, & v_t &\sim \mathcal{N}(0, V) \\ \theta_{t,1} &= \theta_{t-1,1} + \theta_{t-2,1} + w_{t,1}, & w_{t,1} &\sim \mathcal{N}(0, W) \\ \theta_{t,2} &= \theta_{t-1,2} + w_{t,2} \end{aligned}$$

Descripción de los datos

La tabla de datos proviene del Instituto Nacional de Estadística y Censos (INEC) de Costa Rica y es de acceso público, descargable desde la página web del instituto, como puede consultarse en INEC (2021). Esta base presenta los principales indicadores demográficos anuales de Costa Rica durante el periodo 1950-2020. Incluye en total 18 variables, entre las cuales están el año, la población total al 30 de junio de cada año, desagregado también por sexo, así como la cantidad de defunciones. Las variables presentes en la tabla son las siguientes: **Población de estudio:**La población de estudio son aquellas personas que vivían en Costa Rica entre los años 1950-2020 y mueren en este periodo. **Muestra observada:** La muestra observada, son todas aquellas personas que vivían en Costa Rica y al morir son registrados por el Instituto Nacional de Estadística y Censo. **Unidad estadística o individuos**: La unidad estadística es el recuento anual de defunciones en Costa Rica. **Variables de estudio:** Son un total de 18 variables, las cuales según INEC (2020) se tiene: **Año:** Esta variable indica el año. **Total:** Esta variable registra la población total. **Hombres:** Esta variable registra la población total de hombres. **Mujeres:** Esta variable se encarga de registar el total de mujeres. **Nacimientos:** Esta variable registra el total de nacimientos. **Defunciones:** Esta variable registra el total de defunciones. **Defunciones infantiles:** Esta variable registra las defunciones de infantes (niños y niñas). **Defunciones neonatales:** Esta variable registra las defunciones de recién nacidos, hace referencia a la mortalidad de los nacidos antes de alcanzar los 28 días de edad. **Defunciones fetales:** Esta variable registra las defunciones de fetos, se refiere a la mortalidad de un bebé antes o durante el parto. **Tasa de crecimiento:** Esta variable registra la tasa de crecimiento de la población costarricense.Se refiere al crecimiento de la población entre dos fechas sin contemplar la migración. **Tasa de natalidad:** Esta variable registra la tasa de nacimientos registrados en la población costarricenses. **Tasa de mortalidad:** Esta variable registra la tasa de muertes en su totalidad. **Tasa de mortalidad infantil** Esta variable registra la tasa de muertes infantiles **Tasa de mortalidad neonatal:** Esta variable registra la tasa de muertes neonatales. **Tasa de mortalidad fetal:** En esta variable se registra la tasa de muertes fetales. **Tasa global de fecundidad:** Esta variable registra la tasa de fecundidad global (TGF), la cual indica cantidad de hijos e hijas que en promedio tendría cada mujer al final del periodo fértil, si durante su vida tiene sus hijos e hijas de acuerdo a las tasas de fecundidad por edad observadas en el país y año de interés y, además estas mujeres no están afectadas

Tabla 1: Resumen de cinco números para algunas variables de la tabla de datos

	Población total	Población de hombres	Población de mujeres	Nacimientos	Defunciones	Defunciones infantiles	Defunciones neonatales	Defunciones fetales
Mínimo	868 934	438 185	430 749.0	37 248	8 596.0	462.0	344.0	329
Primer cuartil	1 638 283	827 590	810 692.5	59 589	9 897.5	765.0	537.0	526
Mediana	2 646 142	1 339 573	1 306 569.0	70 004	11 376.0	1 345.0	807.0	691
Tercer cuartil	4 054 418	2 053 052	2 001 366.5	75 794	15 704.5	3 543.5	1 173.5	959
Máximo	5 111 238	2 575 550	2 535 688.0	84 337	26 209.0	4 889.0	1 688.0	1 261

Tabla 2: Estadísticos de dispersión para algunas variables de la tabla de datos

	Población total	Población de hombres	Población de mujeres	Nacimientos	Defunciones	Defunciones infantiles	Defunciones neonatales	Defunciones fetales
Desviación estándar	1 340 677	677 255.2	663 434.1	11 274.6	4 428.495	1 396.199	366.604	241.5763
Rango intercuartílico	2 416 136	1 225 461.5	1 190 674.0	16 205.0	5 807.000	2 778.500	636.500	433.0000

por la mortalidad desde el nacimiento hasta el final de periodo fértil. **Tasa bruta de reproducción:** Esta variable hace referencia a el cantidad de hijas que en promedio tendría cada mujer al final del periodo fértil, si durante su vida tiene sus hijos e hijas de acuerdo a las tasas de fecundidad por edad observadas en el país y año de interés y, además estas mujeres no están afectadas por la mortalidad desde el nacimiento hasta el final de periodo fértil. **Tasa neta de reproducción:** es el número de hijas que en promedio tendría cada mujer al final del periodo fértil, si durante su vida tiene sus hijos e hijas de acuerdo a las tasas de fecundidad por edad observadas en el país y año de interés y, además estas mujeres sí están afectadas por la mortalidad por edad observada en el país y año de interés desde el nacimiento hasta el final de periodo fértil.

Análisis descriptivo de los datos

En la tabla Tabla 1 se detalla el resumen de cinco números para algunas variables relevantes de la tabla de datos. Se observa un rango bastante amplio (diferencia entre el máximo y mínimo) en las defunciones totales, lo que es de esperarse dado el crecimiento poblacional que se vio a partir de los 1980's y la cantidad de años que se registran en los datos.

En la tabla Tabla 2 se detallan los estadísticos de desviación estándar y rango intercuartílico para algunas variables relevante de la tabla de datos. Se observa que la variable población de hombres presenta mayor dispersión respecto población de mujeres, esta mayor variabilidad en la población de hombres se puede deber a diversos factores, entre ellos : una diferencia en el número de defunciones, nacimientos, o número de migrantes en la población masculina en relación con la de población femenina.

Una observación importante, es que los nacimientos presentan una mayor dispersión respecto a el número de defunciones. Es decir, se observa que el número defunciones anuales presenta menor desviación respecto a la media (desviación estándar) y una menor diferencia entre tercer cuartil y primer cuartil de defunciones (IQR).

Análogamente, se observa que para los tipos de defunción: infantil, neonatales y fetales. Se evidencia una mayor

Tabla 3: Frecuencia de defunciones, población y tasa de mortalidad

Defunciones	Frec. de defunciones	Población total	Frec. de poblacion	Tasa de mortalidad	Frec. de tasa
(8.58e+03,1.04e+04]	23	(8.65e+05,1.29e+06]	11	(3.72,4.41]	39
(1.04e+04,1.21e+04]	18	(1.29e+06,1.72e+06]	9	(4.41,5.1]	8
(1.21e+04,1.39e+04]	4	(1.72e+06,2.14e+06]	8	(5.1,5.78]	1
(1.39e+04,1.56e+04]	8	(2.14e+06,2.57e+06]	6	(5.78,6.46]	3
(1.56e+04,1.74e+04]	5	(2.57e+06,2.99e+06]	6	(6.46,7.14]	3
(1.74e+04,1.92e+04]	4	(2.99e+06,3.41e+06]	5	(7.14,7.83]	3
(1.92e+04,2.09e+04]	3	(3.41e+06,3.84e+06]	5	(7.83,8.51]	6
(2.09e+04,2.27e+04]	2	(3.84e+06,4.26e+06]	6	(8.51,9.19]	1
(2.27e+04,2.44e+04]	3	(4.26e+06,4.69e+06]	7	(9.19,9.88]	3
(2.44e+04,2.62e+04]	1	(4.69e+06,5.12e+06]	8	(10.6,11.3]	4

dispersión para las defunciones infantiles, seguidas de las neonatales y por último, con menor dispersión las defunciones fetales.

Finalmente, se observa una diferencia significativa entre la medidas de dispersión desviación estándar y rango intercuartílico (IQR), esto se debe a la sensibilidad de la desviación estándar a valores extremos, es decir es posible que existan valores extremos (muy alto o muy bajos respecto a la media) y por esta razón ambas medidas difieran considerablemente, sin embargo para este caso en particular el orden en el grado de dispersión (observe que la dispersión disminuye al avanzar en la tabla de izquierda a derecha) se mantiene para ambos medidas.

De la Tabla 3 se observa que históricamente las defunciones totales de la mayoría de los años cae dentro del primero y segundo intervalo, con una distribución más uniforme dentro de los intervalos más altos. Por su parte la tasa de mortalidad bruta, la cual se aprecia en las columnas cinco y seis, se muestra que en la mayoría de años la tasa se ha mantenido entre 3.72 y 4.41. Esto quiere decir que la cantidad de defunciones relativo a la población ha sido más o menos constante en la mayoría de años. Por su parte, la población total se ha distribuido más uniformemente en todos los intervalos por lo que se concluye que el crecimiento población se ha movido más rápido que las defunciones.

En la Figura 1 se muestra la cantidad total de defunciones para el periodo 1950-2020. Destaca una tendencia creciente muy marcada a partir de cerca de 1980 y hasta el final del periodo considerado. Esto muestra que se debe hacer una transformación a los datos para lograr la estacionareidad que supone el modelo ARMA. También se resalta que para el 2020 las defunciones totales se encuentran claramente por arriba de la tendencia lo cual se debe probablemente a la pandemia del Covid-19.

En la Figura 2 se comparan las defunciones infantiles, neonatales y fetales. Cabe añadir que la distancia vertical entre las defunciones infantiles y las neonatales resulta en las llamadas defunciones posneonatales, es decir, las que ocurren a partir de los 29 días de edad y hasta un año. Se advierte que a mediados de los años sesenta la cantidad de defunciones infantiles aparenta tener una tendencia decreciente. Al respecto, Rosero Bixby afirma que la caída más dramática en los años setenta “se logra gracias a los programas de atención primaria de la salud, ayudados por una extraordinaria reducción de la natalidad que permite un mejor desarrollo intrauterino, mejor cuidado del niño

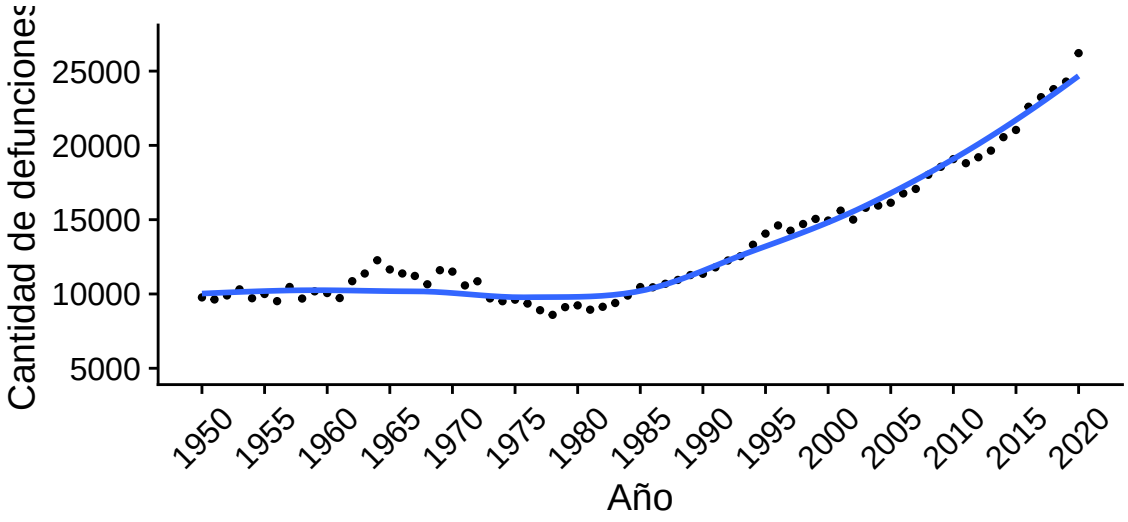


Figura 1: Cantidad de defunciones por año para el periodo 1950-2020

y reduce el riesgo de contagio" (2004). El mismo autor menciona que "el riesgo de morir de los menores de un año ha disminuido en forma poco menos que espectacular entre 1970-78, pues ha sido reducido a la tercera parte (de 62 a 22 muertes por cada mil nacimientos) en un lapso de apenas 8 años" (Rosero-Bixby, 2016). Por su parte, la cantidad de defunciones neonatales superó a las fetales desde mediados de los años cincuenta y hasta mediados de los años ochenta, donde se pierde un poco la noción de cuál suele ser mayor.

Métodos y resultados

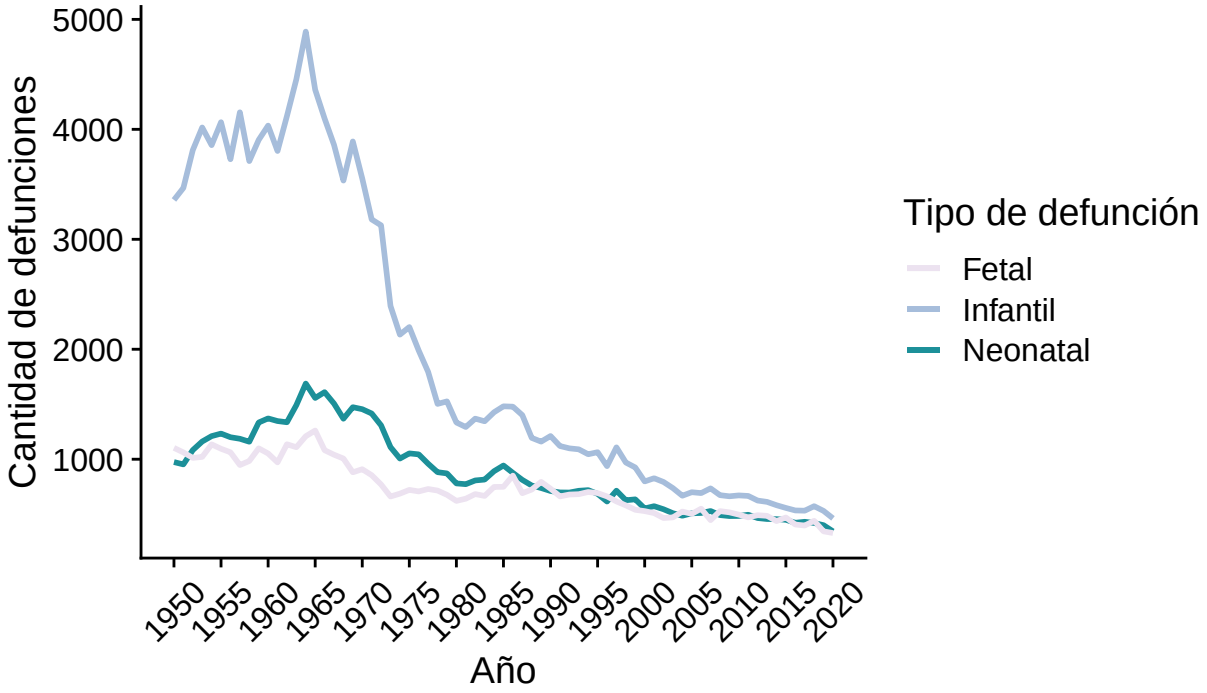
SARIMA

Para la implementación del modelo SARIMA, se observa en la Figura 1 que la serie de tiempo de defunciones totales muestra un comportamiento creciente de carácter lineal, por lo que se busca ajustar un modelo de la forma

$$D_t = \beta_0 + \beta_1 t + x_T$$

Donde D_t representa las defunciones totales, t es el tiempo y x_t es un proceso ARMA.

De la Figura 7 del anexo se observa que los residuos lucen estacionarios. Además, parece que el PACF se trunca después del primer rezago, mientras que el ACF decrece. Esto sugiere fuertemente un modelo autoregresivo AR(1). Tambiés es posible que se trunque el ACF luego del primer o segundo rezago, mientras que el ACF decrece, por lo que también se tienen los modelos candidatos MA(1) Y MA(2)



Fuente: Elaboración propia con datos del INEC

Figura 2: Defunciones infantiles, neonatales y fetales por año

Tabla 4: Resumen de diagnósticos de los modelos propuestos

Modelo	Estadístico W	Valor p Shapiro-Wilks	Estadístico Q	Valor p Ljung-Box	AIC	BIC	AICc
ARMA(1,0)	0.96	0.22	0.01	0.94	14.98	15.15	15.00
ARMA(0,1)	0.94	0.04	4.09	0.04	15.56	15.73	15.58
ARMA(0,2)	0.96	0.20	1.22	0.27	15.18	15.39	15.21

Tabla 5: Intervalos de predicción del modelo ARIMA(1,1,0)

Año	predicción	Inf80	Sup80	Inf95	Sup95
2019	23678.19	23186.70	24169.68	22926.52	24429.86
2020	23950.71	23286.99	24614.43	22935.64	24965.78

La Tabla 4 resume las pruebas Shapiro-Wilks de normalidad y Ljung-Box de independencia de los residuos estandarizados. Se da un no rechazo de las hipótesis nulas de independencia y normalidad de los residuos en el primer y tercer modelo. Además, se da un rechazo de la hipótesis de normalidad e independencia de los residuos en el segundo modelo. Para una desglose visual de los diagnósticos de los modelos SARIMA ajustadas se puede consultar el anexo. El modelo autoregresivo presenta las mejores medidas de AIC, AICc y BIC, aparte de ser el más parsimonioso de los tres. Por esta razón se decide escoger el modelo con componente ARMA(1,0) para hacer el pronóstico.

La Tabla 5 muestra el valor pronosticado para las defunciones totales en los años 2019 y 2020 y los intervalos de predicción al 80% y 95%. Al comparar con los valores reales de 24292 para el 2019 y 26209 para el 2020, se observa que el valor real para el 2019 es consistente con el intervalo de predicción al 95%, pero no al 80%. Por su parte, el valor real para el 2020 se sale de los intervalos a ambos niveles de confianza. Esto es esperable pues el primer año de pandemia tuvo una cantidad inusualmente alta de muertes. La diferencia entre los valores predichos y los reales para el 2019 y 2020 son dados por 613.80 y 2258.29 respectivamente.

DLM 1

Para utilizar un DLM polinomial de orden 1, se empieza por estimar los parámetros V y W correspondientes a la varianza de el ruido blanco gaussiano aditivo en las ecuaciones observada y del sistema, respectivamente. Esto se lleva a cabo por máxima verosimilitud, y se obtiene $\hat{V} = 0.37$ y $\hat{W} = 304117$. Nótese la gran diferencia en escala entre ambas varianzas. En Petris et al. (2007) se advierte que este modelo en particular es muy sensible respecto al valor de la razón $\frac{W}{V}$, llamada *radio señal-ruido*. Cabe mencionar también que este proceso se comporta asintóticamente como un ARIMA(0,1,1) (Petris et al., 2007).

Una vez obtenidos estos parámetros, se procede a aplicar el filtro de Kalman y con este se realiza el pronóstico de la cantidad total de defunciones para 2019 y 2020, donde se encuentra un inconveniente: para ambos años, el pronóstico es de 23786 defunciones. Más aún, esta cantidad es justamente el total de defunciones en 2018. Sin duda esto representa una gran desventaja de este modelo pues claramente no resulta útil en términos de predicción, incluso teniendo en cuenta que solo es prudente tener pronósticos de corto plazo. Este puede ser explicada por el gran desbalance que existe entre las varianzas de los ruidos de cada ecuación del modelo, lo que se traduce en un radio señal-ruido muy alto.

En efecto, en conformidad con Petris et al. (2007) el filtro de Kalman para este simple modelo puede ser expresado de la siguiente manera:$y_{1:t-1} \sim \mathcal{N}(m_{t-1}, R_t = C_{t-1} + W)$, $Y_t \mid y_{1:t-1} \sim \mathcal{N}(f_t = m_{t-1}, Q_t = R_t + V)$ y $\mu_t \mid y_{1:t} \sim \mathcal{N}(m_t = m_{t-1} + K_t e_t, C_t = K_t V)$; donde la notación $y_{1:s}$ hace referencia a las observaciones

$y_1, y_2, \ldots, y_s$, $e_t = Y_t - f_t$ y $K_t = \frac{R_t}{Q_t} = \frac{C_{t-1}+W}{C_{t-1}+W+V} = 1 - \frac{1}{\frac{C_{t-1}}{V}+\frac{W}{V}+1}$. De la última ecuación se tiene que para valores altos de $\frac{W}{V}$, K_t es cercano a 1. De las mismas ecuaciones que concede el filtro de Kalman se extrae la recursión $m_t = K_t y_t + (1 - K_t)\, m_{t-1}$ de forma que K_t funciona como un peso, y si $\frac{W}{V}$ es grande, entonces m_t es "similar" a y_t y, por lo tanto, el pronóstico a un paso se parece mucho a la última observación. Es precisamente este fenómeno el que podría estar desembocando en el poco provecho del modelo en términos de sus pronósticos. Debido a que este modelo no es útil para la investigación debido a lo ya indicado, se decide omitir los diagnósticos y los criterios de información.

DLM 2

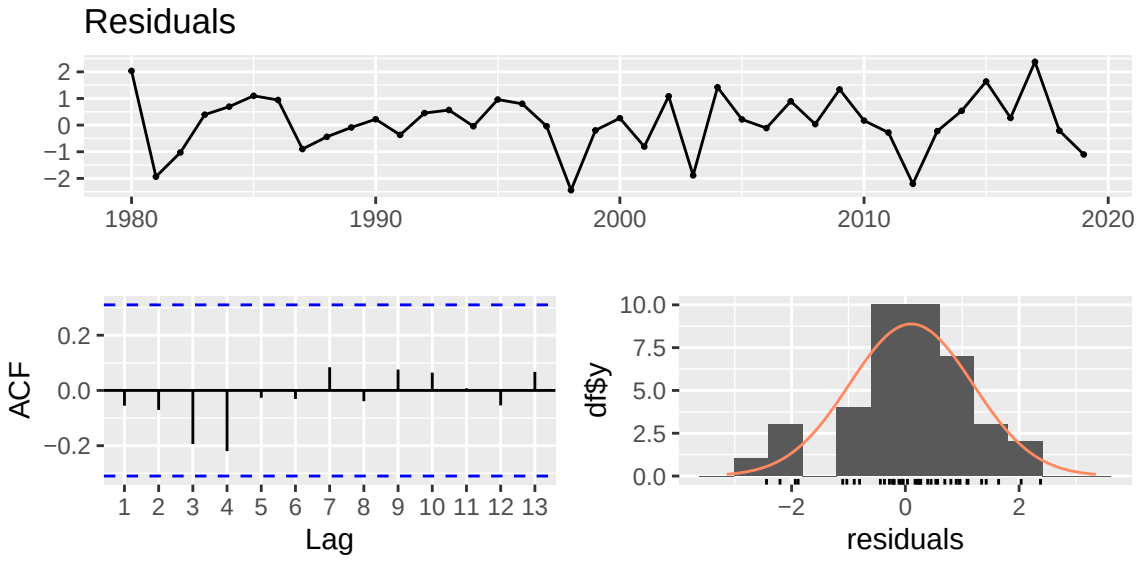


Figura 3: Diagnóstico de residuos para el modelo DLM polinomial de segundo orden

De los diagnósticos de los residuos Figura 3 se osberva del grafico de ACF que existe una baja correlación entre los residuos lo cual se confirma al aplicar la prueba Ljung-Box. Se observa del gráfico superior de residuos la presencia de outliers, esto se ve reflejado el en el histograma de los residuos en la parte inferior derecha donde se ve que los residuos presentan colas de mayor peso que la distribución normal, no obstante al aplicar la prueba Shapiro-Wilks de normalidad se observa que no se rechaza la hipotesis de normalidad.

De la Figura 4 se ve que el ajuste normal es bueno salvo hacia la cola izquierda.

El grafico Figura 5 muestra los p valores para el estadistico Ljung-Box donde se observa que los valores p se encuentran por encima del umbral con valor p de 0.05 (linea punteada), indicando que los residuos del modelo DLM polinomial de orden 2 son independientes. Concluimos, que el modelo ajusta bien o que el modelo no muestra una falta de ajuste.

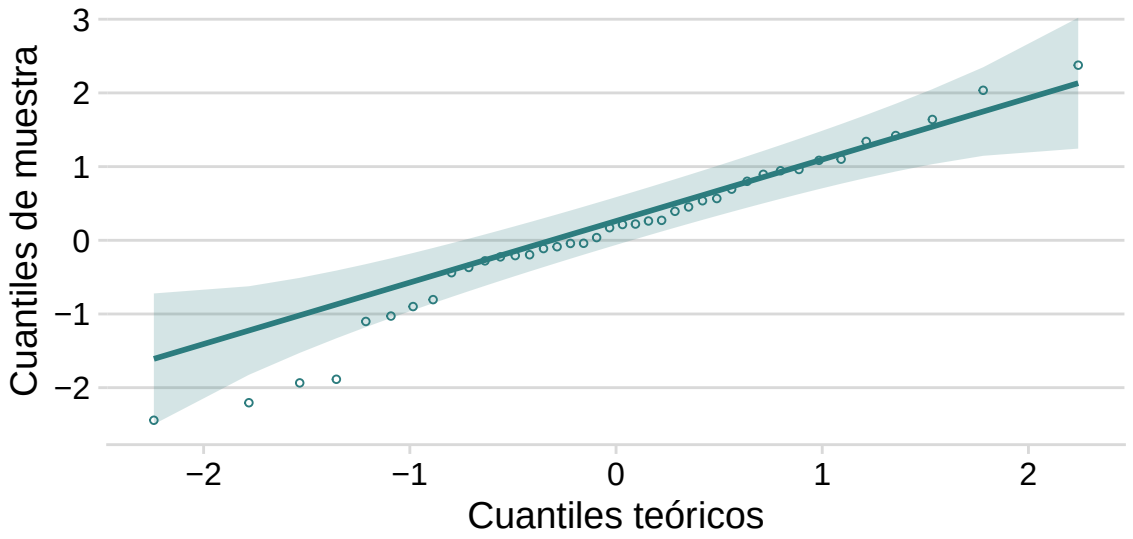


Figura 4: Gráfico cuantil-cuantil de los residuos para el modelo DLM polinomial de segundo orden

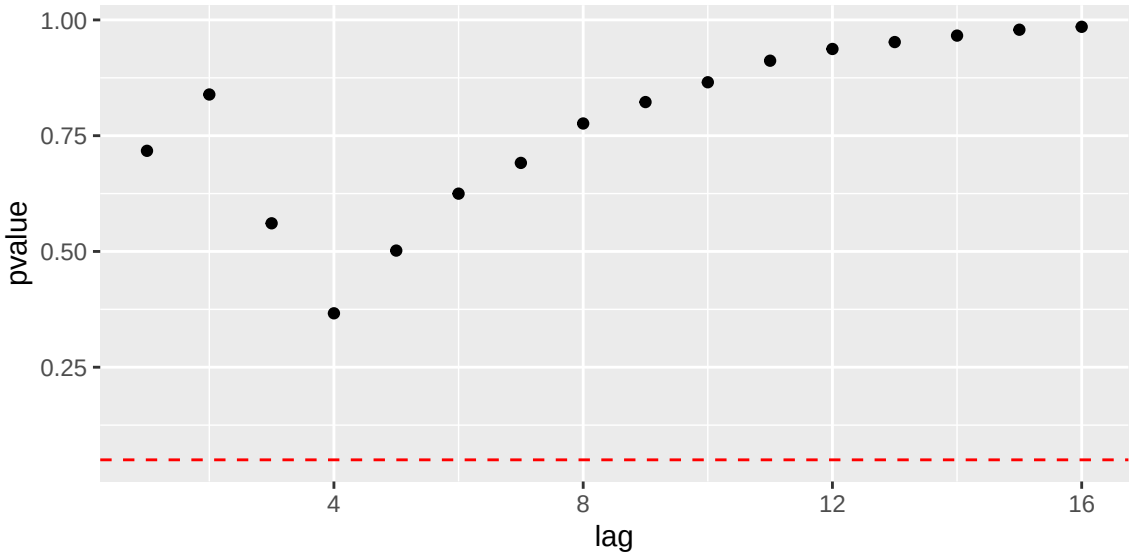


Figura 5: Valores p del estadistico Ljung-Box

Tabla 6: Resumen de diagnósticos de los modelos DLM polinomiales

Modelo	Estadístico W	Valor p Shapiro-Wilks	Estadístico Q	Valor p Ljung-Box	AIC	BIC	A
DLM polinomial orden 2	0.97	0.42	5.54	0.94	-21.1	-16.03	-20

Tabla 7: Intervalos de predicción del modelo DLM polinomial orden 2

Año	predicción	Inf80	Sup80	Inf95	Sup95
2019	24702	24170	25233	23889	25515
2020	25490	24690	26290	24267	26713

Análogamente a los modelos ARIMA , la Tabla 6 muestra las pruebas Shapiro-Wilks de normalidad y Ljung-Box de independencia de los residuos estandarizados. Para el caso de la prueba Shapiro-Wilks, no rechaza la hipótesis nula para el modelo DLM polinomial de orden 2, validando nuestra hipótesis de normalidad. Por otro lado, para la prueba Ljung-Box de independencia de los residuos estandarizados se da un no rechazo de la hipotesis nula, esto se observa al obtener un valor p de 0.16.

Los valores del AIC, AICc y BIC se observan que son menores a los obtenidos por los modelos ARIMA mostrados en la tabla Tabla 5 indicando que el modelo DLM polinomial de orden 2 mediante estos criterios es mejor modelo que los ARIMA.

En Tabla 7 se muestran los pronósticos de defunciones realizados por el modelo para el año 2019 y 2020. Se muestran los intervalos de confianza al 80% y 95% de dichos pronósticos. Al comparar con los valores reales de 24292 para el 2019 y 26209 para el 2020 se observa que hay una diferencia de 410 y 719 defunciones aproximadamente.

Conclusiones

Ambos modelos estudiados logran hacer pronósticos a dos años de las defunciones que son consistentes con las reales. En el caso del ARIMA, el dato real para el 2020 se sale del intervalo de confianza al 80% y al 95%, mientras que el DLM polinomial de orden dos logran crear intervalo más certeros. Además, el segundo modelo parece ser menos conservador con las predicciones para la serie utilizada.

Si bien el modelo DLM polinomial de orden no logra ser tan preciso como el ARIMA para el año 2019, sí logra ser más preciso para el año 2020, mostrando un mejor ajuste para lo que se consideraría un año extraordinario por ser el primer año de pandemia.

Los pronósticos hechos por el modelo DLM polinimal de orden 1 no resultan útiles producto de la razón señal-ruido tan alta, Sin embargo, este resultado es consistente con el modelo.

Limitaciones

La limitación más clara del modelo ARIMA, es que al hacer pronósticos a largo plazo rápidamente se va a la media del proceso por lo que se imposibilita hacer pronósticos con una ventana de tiempo más amplia de la presentada en este proyecto.

Tal como lo establece Mary (2006) el modelo DLM polinomial de orden 2 ha demostrado ser suficiente para pronósticos a corto plazo, no obstante una de sus grandes limitaciones al igual que los modelos ARIMA es que los pronósticos a largo plazo son deficientes. Dicho esto concluimos por tanto, en general nuestro estudio está limitado a pronósticos de defunciones totales a corto plazo.

Recomendaciones

La mayoría de trabajos sobre demografía estadística para el caso costarricense que se consultaron fueron de corte descriptivo, con la excepción notable del de Aguilar Fernández (2013), quien justamente realiza una proyección sobre la mortalidad. En este trabajo se emplea el método Lee-Carter, que es un modelo de muy recurrente mención en la literatura sobre proyecciones demográficas, siendo también empleado por el mismo Instituto Nacional de Estadísticas y Censos (INEC, 2013) para la proyección de la mortalidad, por lo cuál debería considerarse la posibilidad de emplearlo y/o compararlo con otros métodos. Además, una de las características de estos trabajos es que no consideran la mortalidad de forma aislada, como se hace en el presente análisis, sino que las proyecciones involucran a los tres componentes del cambio poblacional (fecundidad, mortalidad y migración). También, debe mencionarse que estas proyecciones trabajan sobre datos mucho más específicos y pormenorizados, teniendo en cuenta conteos desagregados por edad (o grupo de edad) y sexo. En ese sentido, se cree para futuras extensiones de esta investigación y obedeciendo a la literatura afín, sin duda alguna se recomienda tratar de localizar datos que, primero, al menos estén desagregados por sexo y, segundo, que estos además tengan en cuenta distintos grupos de edad. También, podía valorarse la inclusión de covariables, como lo pueden ser tasas de criminalidad, indicadores de salud o también una variable de la población total rezagada.

Referencias

Adekanmbi, D., Ayoola, F., & Idowu, A. (2014). Demographic Time Series Modelling of Total Deaths in Nigeria. *Population Association of Southern Africa*, *15*(1), 21-48.

Aguilar Fernández, E. (2013). *Estimación y proyección de la mortalidad para Costa Rica con la aplicación del método Lee-Carter con dos variantes= Fitting and forecasting mortality for Costa Rica applying the Lee-Carter method with two variants*.

Alcalde, F. P. (2010). La teoría de la transición demográfica: recursos didácticos. *Enseñanza de las ciencias sociales: revista de investigación*, 129-137.

Bruno Ramírez, X. S. (2019). *Análisis del crecimiento y proyección poblacional del distrito Chulucanas, 2017-2025*.

INEC. (2013). *Estimaciones y Proyecciones de Población por sexo y edad 1950 - 2050*.

INEC. (2020). *Indicadores demográficos 2019*.

INEC. (2021). *Estadísticas demográficas. 1950-2020. Principales Indicadores Demográficos*. %7Bhttps://www.inec.go.cr/documento/estadisticas-demograficas-1950-2020-principales-indicadores-demograficos%7D

Macció, G. A., Centro Latinoamericano de Demografia, S., et al. (1985). *Diccionario demográfico multilingue*.

Mariscal de Gante, Á., & Rodríguez, V. (2018). *Tres teorías demográficas, las evidencias disponibles y el paso de la descripción del cómo al entendimiento del porqué: Una aplicación y una crítica de tres hipótesis poblacionales en base a los casos de España y de la India (1950-2020)*.

Mary, K. (2006). *Determining optimal architecture for dynamic linear models in time serires applications*.

Ordorica, M. (2004). Pronóstico de las defunciones por medio de los modelos autorregresivos integrados de promedios móviles. *249–264*, *10*(42), 21-48.

Petris, G., Petrone, Sonia, & Patrizia, C. (2007). *Dynamic Linear Models with R*. Springer.

Rees, P. (2020). Demography. En A. Kobayashi (Ed.), *International Encyclopedia of Human Geography (Second Edition)* (Second Edition, pp. 239-256). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-08-102295-5.10252-5

Rosero-Bixby, L. (2004). *Situación demográfica general de Costa Rica*.

Rosero-Bixby, L. (2016). La situación demográfica en Costa Rica. *Población y Salud en Mesoamérica*, *13*(2), 237-311.

Sarrible, G. (2009). *Teoría de la población*.

Anexo

Código

Carga de paquetes

```
options(knitr.kable.NA = '-', echo = FALSE)
library(readxl)
library(kableExtra)
library(dplyr)
library(janitor)
library(ggplot2)
library(wesanderson)
library(cowplot)
library(viridis)
library(GGally)
library(scales)
library(ggplot2)
library(ggpubr) # qqplots en ggplot2
#Paquetes para series de tiempo
library(astsa)
library(tseries)
library(forecast)


#Paquetes para DLM:

# library(devtools)

#devtools::install_github("https://github.com/cran/dlm")
```

```r
# install.packages("dlm", repos = "https://cran.rstudio.com/",  dependencies = TRUE)

library(dlm)
library(ggmcmc) #Para ACF Y PACF DLM .
```

Carga de datos y depuración

```r
base <- data.frame(read_excel("sepoblacev1950-2020_0.xls",
    range = "B9:S80", col_types = "numeric"))

colnames(base) <- c("Año",
                    "Población total",
                    "Población de hombres",
                    "Población de mujeres",
                    "Nacimientos",
                    "Defunciones",
                    "Defunciones infantiles",
                    "Defunciones neonatales",
                    "Defunciones fetales",
                    "Tasa de crecimiento",
                    "Tasa de natalidad",
                    "Tasa de mortalidad",
                    "Tasa de mortalidad infantil",
                    "Tasa de mortalidad neonatal",
                    "Tasa de mortalidad fetal",
                    "Tasa global de fecundidad",
                    "Tasa bruta de reproducción",
                    "Tasa neta de reproducción"
)

base <- base %>% clean_names()
```

Modelo SARIMA

División del conjunto de entrenamiento y de prueba.

```
entrenamiento <- base$defunciones[30:69]
serie <- ts( entrenamiento, start=c(1980,1), frequency=1)
```

Gráficos de la serie de tiempo

```
plot(serie, main="", xlab="Tiempo", ylab='Defunciones totales')
```

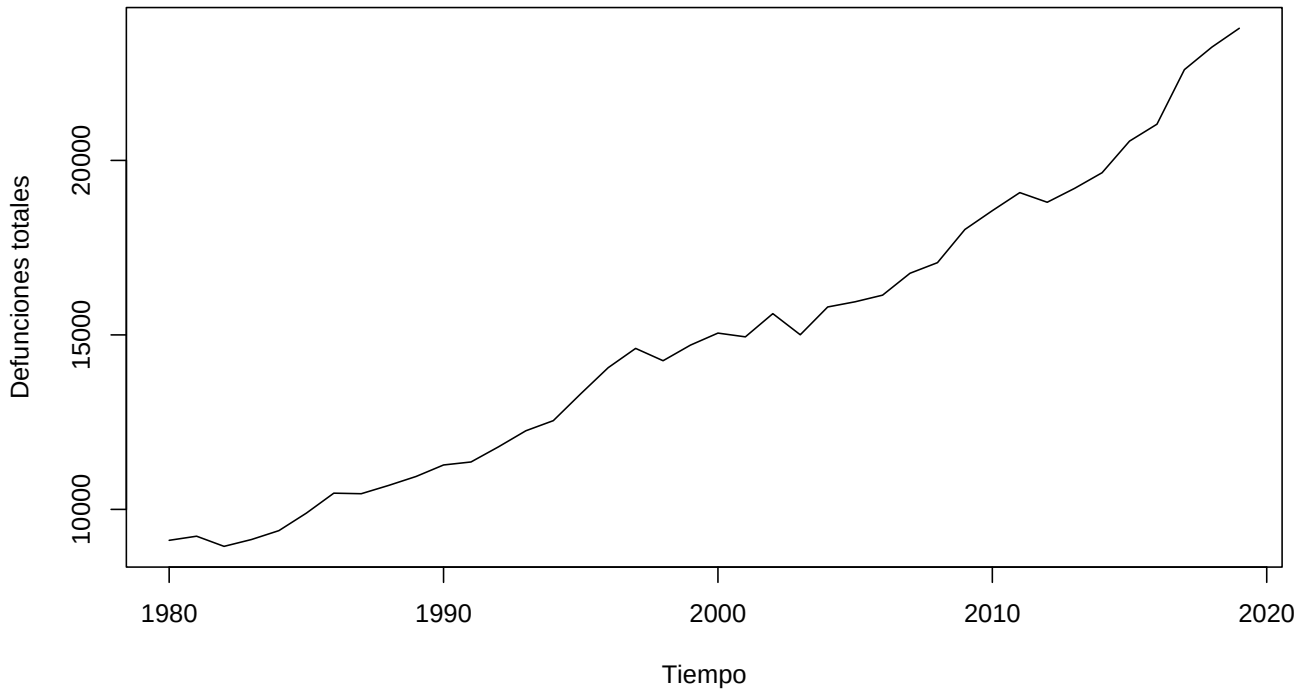


Figura 6: Serie de tiempo de defunciones totales

Implemenatación de los modelos SARIMA y diagnósticos

```
fit1 <- lm(serie ~ time(serie))
t <- acf2(resid(fit1), main='')
```

```
modelo1 <- sarima(serie, 1,0,0, xreg=time(serie))
```

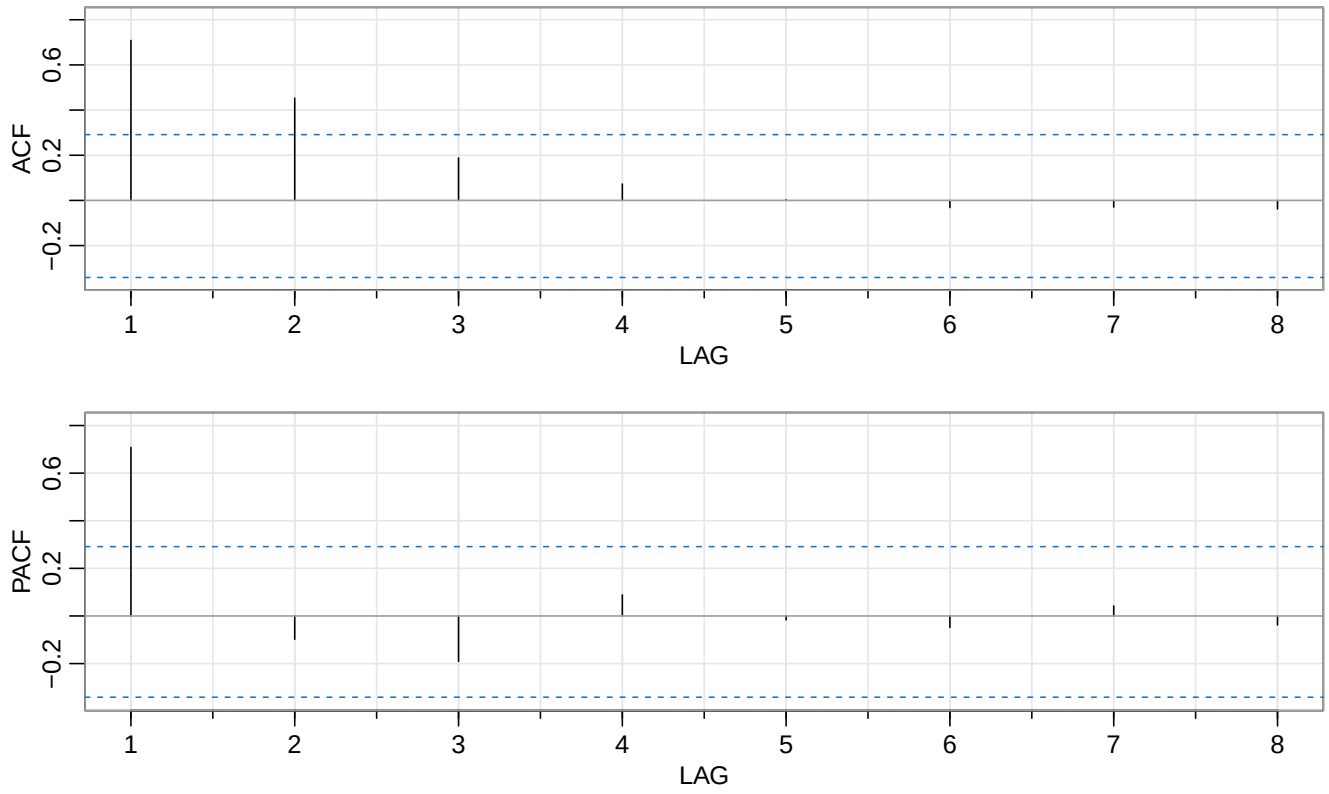


Figura 7: ACF y PACF de los residuos de la regresión

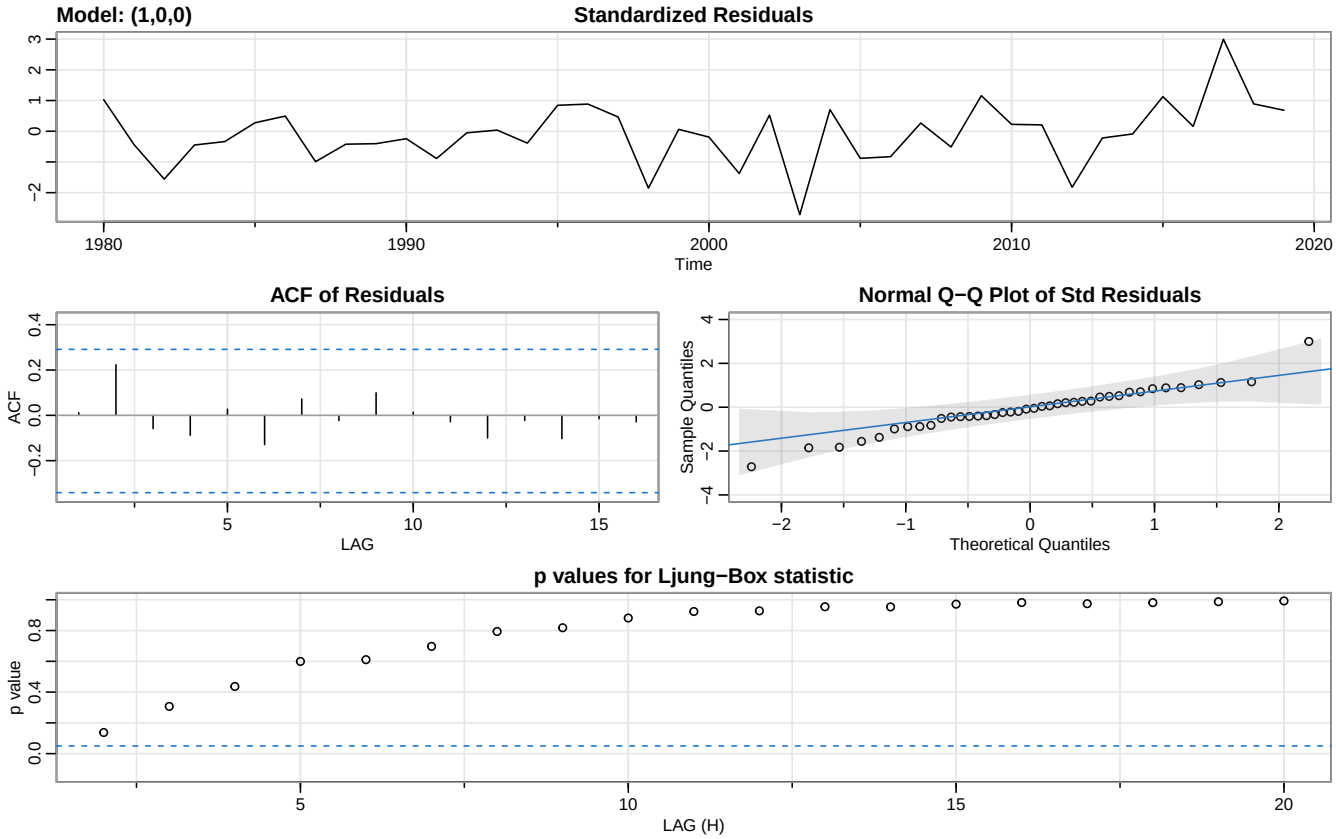


Figura 8: Diagnósticos del modelo con componentente ARMA(1,0)

```
modelo2 <- sarima(serie, 0,0,1, xreg=time(serie))
```

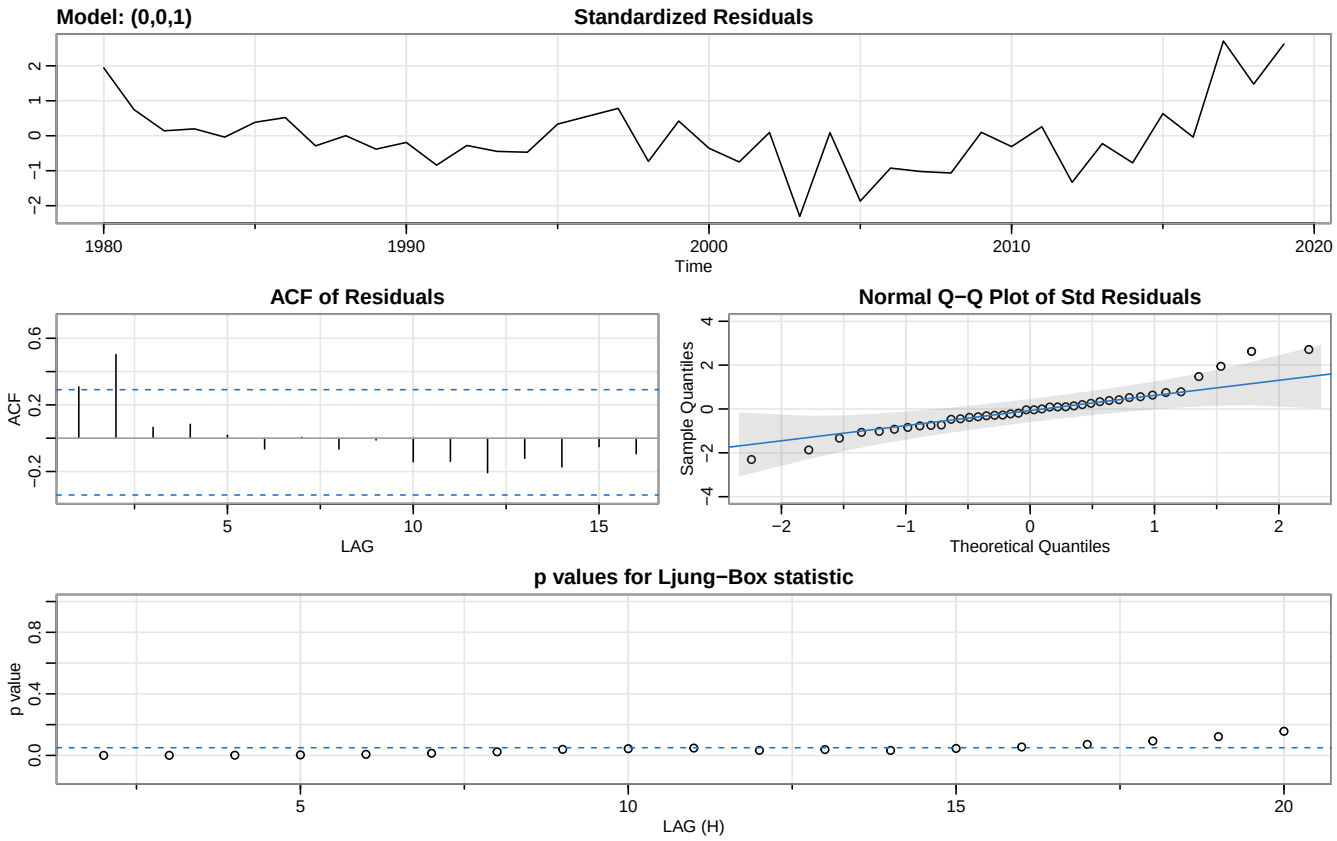


Figura 9: Diagnósticos del modelo con componente ARMA(0,1)

```
modelo3 <- sarima(serie, 0,0,2, xreg=time(serie))
```

Pronóstico

```
def.for <- sarima.for(serie, n.ahead = 2, 1,0,0, xreg=time(serie),
                      newxreg = c(2019,2020))
```

Intervalos de predicción

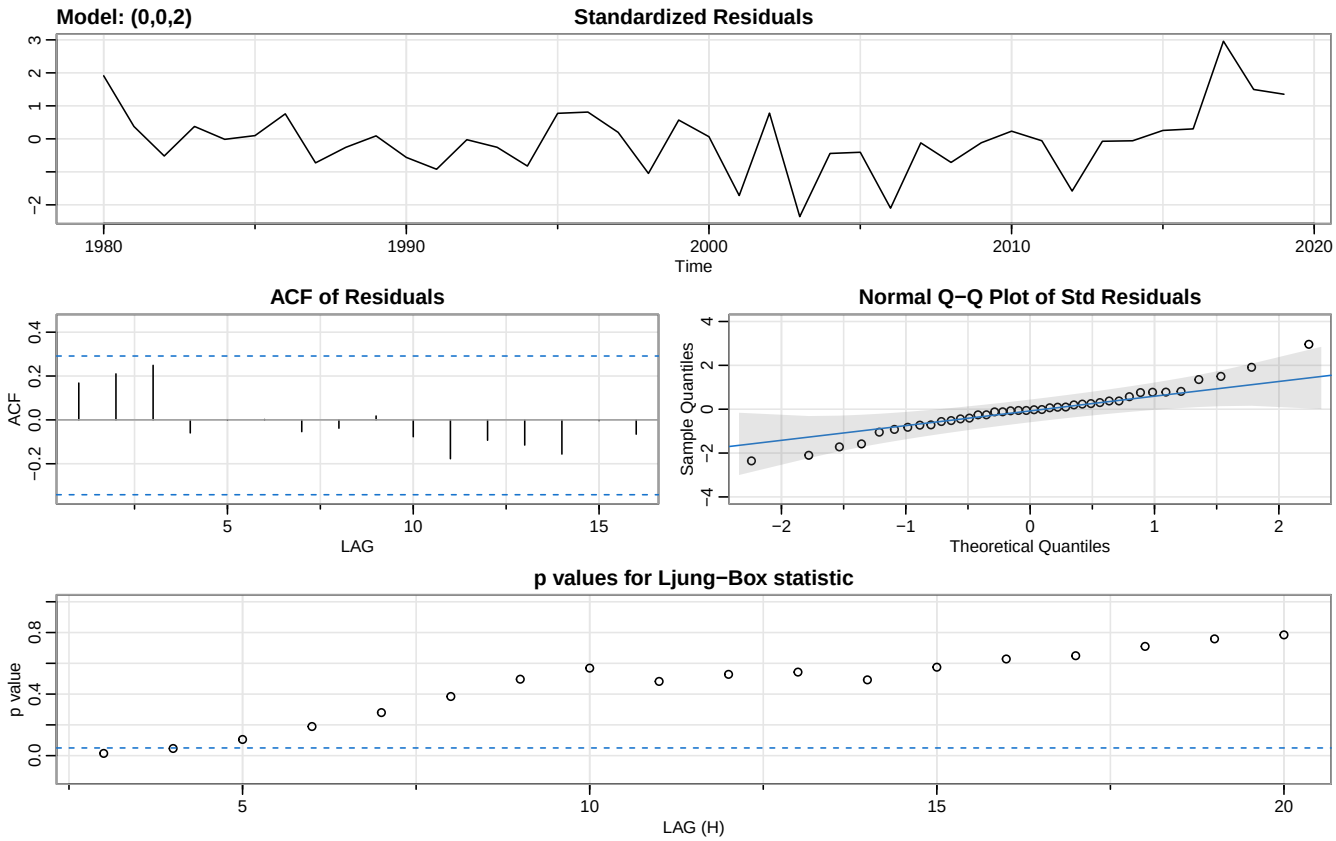


Figura 10: Diagnósticos del modelo ARMA(0,2)

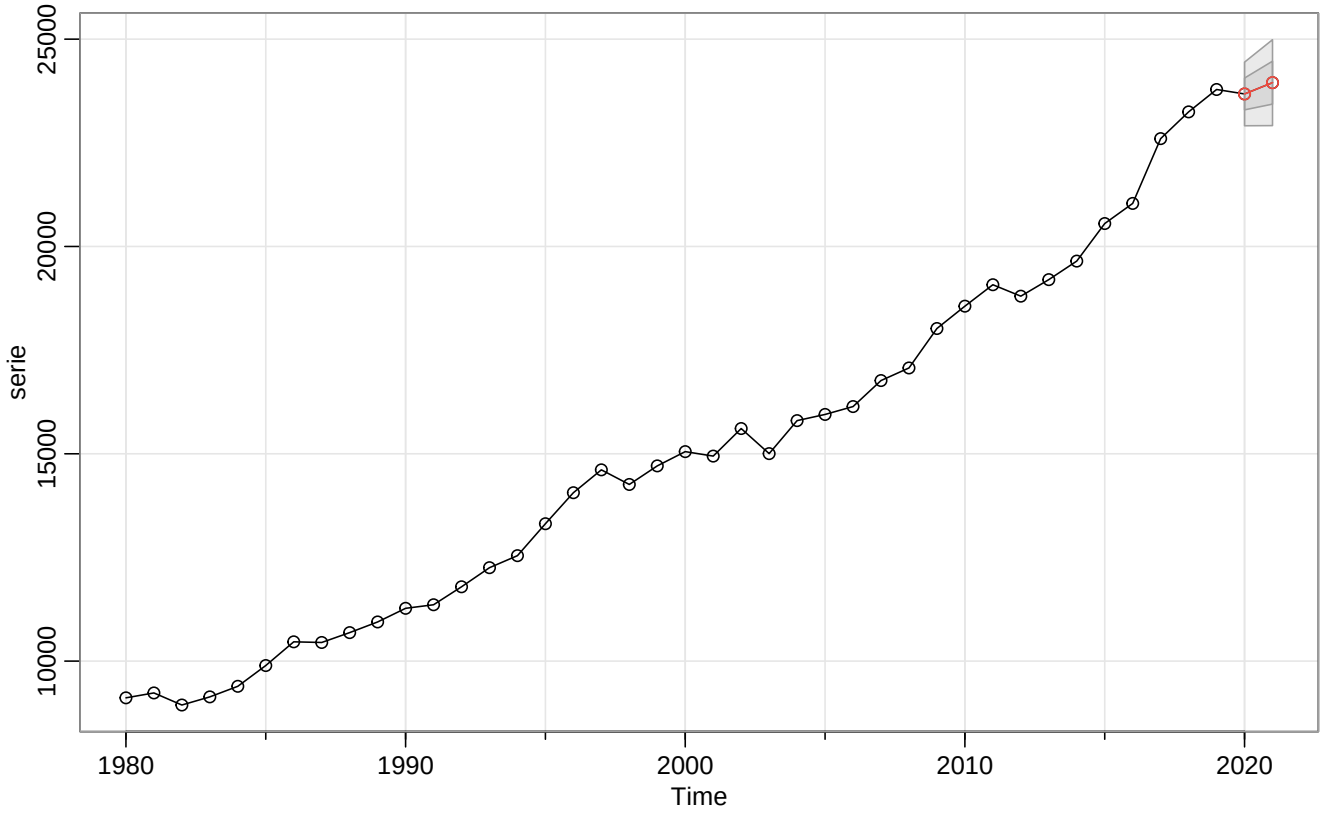


Figura 11: Pronóstico del modelo con componente ARMA(1,0) de las defunciones totales

```r
#Al 95%
U95 <- def.for$pred + def.for$se*qnorm(1-0.05/2)
L95 <- def.for$pred - def.for$se*qnorm(1-0.05/2)

#Al 80%
U80 <- def.for$pred + def.for$se*qnorm(1-0.2/2)
L80 <- def.for$pred - def.for$se*qnorm(1-0.2/2)


for.int <- data.frame(c(2019, 2020), def.for$pred, L80, U80, L95, U95)
```

Modelo DLM polinomial de primer grado

```r
serie1980 <- window(serie, start=1980)

# serie_insumo <- window( serie, end = 1979)
# m0_prueba = mean(serie_insumo)
# C0_prueba = var(serie_insumo)

# Se crea la estructura del modelo DLM orden 1:

DLMOrden1 <- function(parm) {
      dlmModPoly(order = 1, dV = exp(parm[1]), dW = exp(parm[2]))
}

# Se procede hacer ajuste de parametros vía maxima verosimilitud:

ajusteDlmOrden1 <- dlmMLE(serie1980, rep(0, 2), build = DLMOrden1, hessian = TRUE)


# Se crea el modelo con los parametros obtenidos vía maxima verosimilitud:

modeloDlmOrden1 <- DLMOrden1(ajusteDlmOrden1$par)

# round(modeloDlmOrden1$V, 2)
# round(modeloDlmOrden1$W, 2)
```

```
DLMOrden1 <- function(parm) {
      dlmModPoly(order = 1,
                 dV = exp(parm[1]),
                 dW = exp(parm[2]))
}
# Ajuste máxima verosimilitud
ajusteDlmOrden1 <- dlmMLE(serie1980, rep(0, 2), build = DLMOrden1, hessian = TRUE)
# Se aplica el filtro de Kalman:
filtroDLM1 <-  dlmFilter(serie1980, mod = modeloDlmOrden1)
# Se realiza el pronóstico
forecastDLM1 <-  dlmForecast(mod = filtroDLM1, nAhead = 2)
# forecastDLM1$f
```

Pronósticos

```
# Se aplica el filtro de Kalman:

filtroDLM1 <-  dlmFilter(serie1980, mod = modeloDlmOrden1)

# Procedemos a hacer forecast, en este caso se hace se hace forecating 2 pasos adelante.

forecastDLM1 <-  dlmForecast(mod = filtroDLM1, nAhead = 2)

# forecastDLM1
```

Diagnósticos

```
residsDLM1 <- residuals(filtroDLM1, sd = FALSE)
# checkresiduals( resids, test = F)

with_theme_cowplot <- function(expr) {
  orig <- theme_get()
  theme_set(theme_cowplot())
  force(expr)
  theme_set(orig)
}
g <- with_theme_cowplot(checkresiduals( residsDLM1, test = F, lag = 40))
```

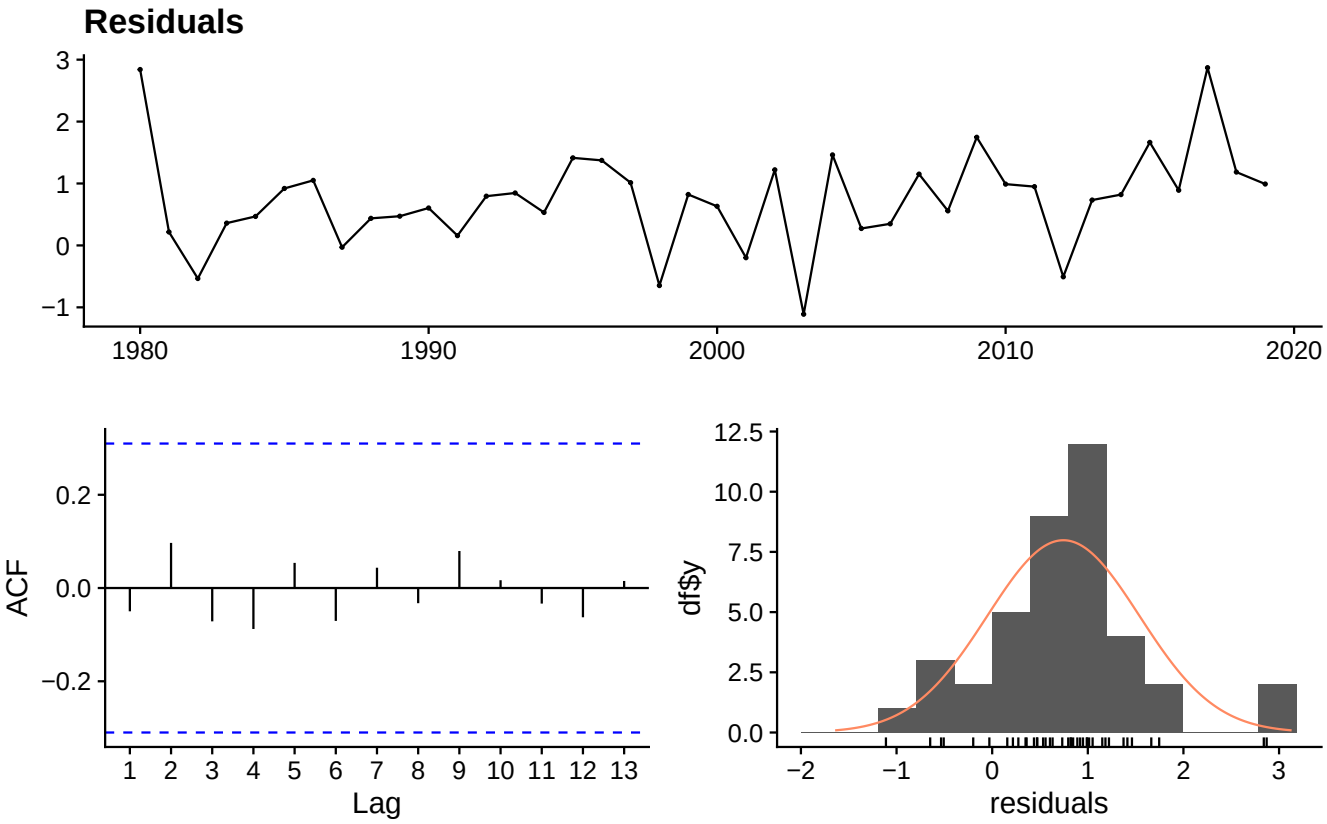


Figura 12: Algunos diagnósticos descriptivos de los residuos para el modelo DLM polinomial de primer orden

```r
# Se crea codigo para plotear los p values del estadistico Ljung-Box:


acfLDM1 <- acf(residsDLM1, plot = F)

acf_df <- data.frame(
  acf = acfLDM1$acf,
  lag = acfLDM1$lag
)


#Se calculan los p values del estadistico:
acf_df$pvalue <- sapply(acf_df$lag, function(i) Box.test(residsDLM1, lag=i, type="Ljung-Box")$p

acf_df <-  acf_df[-1,] # Se elimina lag en 0.
```

Modelo DLM polinomial de segundo orden

Implementación del modelo

```r
#Se crea la estructura del modelo DLM orden 2:
DLMOrden2 <- function(parm) {
    dlmModPoly(order = 2, dV = exp(parm[1]), dW = exp(parm[2:3]))

}
#Se procede hacer ajuste de parametros via maxima verosimilitud:

ajusteDlmOrden2 <- dlmMLE(serie, rep(0, 3), build = DLMOrden2 , hessian = TRUE)


#Se crea el modelo con los parametros obtenidos via maxima verosimilitud:

modeloDlmOrden2 <- DLMOrden2(ajusteDlmOrden2$par)
```

```r
#Se aplica filtro de Kalman:

filtroDLM2 <-  dlmFilter(serie, mod = modeloDlmOrden2  )
```

```r
#Procedemos hacer forecast, en este caso se hace se hace forecating 2 pasos adelante.


forecastDLM2 <-  dlmForecast(mod = filtroDLM2, nAhead = 2 )
```

Diagnósticos de residuos

```r
# Diagnostico de los residuos:

resids <-  residuals(filtroDLM2, sd = FALSE)
# Hacen varios plots para el chequeo de residuos.
checkresiduals( resids, test = F) + theme_minimal()
```

Diagnósticos adicionales y pruebas de bondad de ajuste

```r
#Se calcula el AIC, BIC y AICc
loglikDLM2 <- dlmLL(serie, dlmModPoly(2))
numeroParametros <- 3
n <- length(serie)
AICDLM2 <- 2 * (numeroParametros) - 2*log(loglikDLM2)
BICDLM2 <- (log(n)) * (numeroParametros)- 2*log(loglikDLM2)

AICcDLM2 <-  AICDLM2 + (2*(numeroParametros^(2) + numeroParametros))  /(n- numeroParametros-1)


#Se aplica el test Box-Ljung:

LjungTestDLM2 <-  Box.test(resids, lag = 12, type = "Ljung")
# Valor p del test Box-Ljung:
valorPLjungTestDLM2 <-  LjungTestDLM2$p.value
# estadistico del test Box-Ljung Q:

estadPLjungTestDLM2 <- LjungTestDLM2$statistic


#Se aplica el test Shapiro:

ShapiTestDLM2 <-  shapiro.test(resids)
# Valor p del test Shapiro:
```

```
valorPShapiTestDLM2 <-  ShapiTestDLM2$p.value

# estadistico del test Shapiro W:

estadShapiTestDLM2 <- ShapiTestDLM2$statistic
```

Intervalos de predicción

```
#Se construyen intervalos de confianza para DLM polinomial orden 2:

#Varianza:

varDLM2 <-  unlist( forecastDLM2$Q)

#Desviacion estandar:

sdDLM2 <-  sqrt(varDLM2)

#Intervalo de confianza al 95%:

U95DLM2 <- forecastDLM2$f + sdDLM2*qnorm(1-0.05/2)
L95DLM2 <- forecastDLM2$f - sdDLM2*qnorm(1-0.05/2)

#Intervalo de confianza al 80%:

U80DLM2 <- forecastDLM2$f + sdDLM2*qnorm(1-0.2/2)
L80DLM2 <- forecastDLM2$f - sdDLM2*qnorm(1-0.2/2)

datosICDLM2 <- data.frame(c(2019, 2020),forecastDLM2$f  ,L80DLM2, U80DLM2, L95DLM2, U95DLM2)
```