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

28 de noviembre de 2022

Indice de contenidos

Listado de Figuras

Listado de Tablas

Bitácora 1

En el proyecto se buscará realizar análisis demográfico, específicamente se centrará en el pronóstico de la cantidad de defunciones en Costa Rica.

Parte 1

Características generales de la tabla

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.

Variables de estudio

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 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.

Clasificación de las variables

Todas las variables de la tabla utilizada son numéricas donde cinco identifican cantidades absolutas registradas con respecto a la cantidad de nacimientos, defunciones, etc, y nueve variables son tasas de variabilidad con respecto al año anterior, o como un porcentaje de la población.

Parte 2

Pregunta central de investigación

La pregunta formulada es: ¿Cómo se puede realizar un pronóstico de la serie de defunciones totales anuales de Costa Rica?

Objeto de la investigación

La cantidad de defunciones totales anuales ocurridas en Costa Rica y registradas por el Instituto Nacional de Estadística y Censos del año 1950 hasta el año 2020.

Conceptos delimitadores de la pregunta de investigación

- Defunciones totales anuales: Según Macció et al. (1985), la mortalidad o defunción se emplea para expresar la acción de la muerte sobre la población. Esta se mide en valores absolutos y sobre el año calendario.
- Pronóstico: Según Brownlee (2020) la realización de predicciones sobre el futuro se denomina extrapolación en el tratamiento estadístico clásico de los datos de las series temporales. Los campos más modernos que se centran en el tema, lo denominan pronóstico de series temporales.El pronóstico consiste en tomar modelos ajustados a datos históricos y utilizarlos para predecir observaciones futuras.
- Serie: De acuerdo a Macció et al. (1985) una series cronológica o crónica se le llama a "la que forman los valores sucesivos que una cierta variable ha tomado en el transcurso del tiempo".

Principios

Rees (2020) afirma que una de las mayores contribuciones de la demografía ha sido el desarrollo de modelos para la proyección de futuras poblaciones, lo cuál está estrechamente relacionado con la proyección de defunciones. De acuerdo al autor, dos de los modelos más importantes y usados en Demografía cuando a proyección poblacional se refiere, están el de Cohorte-Componente (*The Cohort–Component Model for Projecting the Population*) y el de Proyecciones Probabilísticas (*Probabilistic Projections*).

El modelo de Cohorte-Componente utiliza datos de nacimientos, muertes y migración (a estos factores se les conoce como componentes de cambio) a nivel de cohortes. De acuerdo al Diccionario Demográfico Multilingüe, un cohorte es "conjunto de individuos que han vivido un cierto acontecimiento durante un mismo período" (1985); y en el marco del modelo puede referirse a grupos de edad. Usualmente, los cohortes están desagregados por sexo (Wilson & Rees, 2021).

Por otro lado, el modelo de Proyecciones Probabilísticas, el cual también es establecido en Rees (2020) trata de estimar la población futura a través de los distribuciones de error de ciertos componentes como la tasa de fertilidad. Con estas se pueden proyectar año a año y bajo la combinación de cientos de estos componentes se puede construir escenarios posibles. Cada una de estas distribuciones se puede derivar utilizando datos históricos y a través del uso de modelos de series de tiempo auto-regresivas, o de censos o opiniones experta. Bajo este enfoque se construyen varios escenarios alrededor de una mediana y que se dispersan producto de las distribuciones de incertidumbre pero que al final logran construir un análogo del intervalo de confianza.

Bitacora 2

Punto 1

En la Tabla 1 y Tabla 2 se presentan las primeras cinco líneas de la tabla de datos para las columnas 1 a 9 y 10 a 18 respectivamente. En conformidad con Wickham & Grolemund (2016), una tabla en formato *tidy* cumple que:

- Cada variable possee su propia columna
- Cada observación posee su propia fila
- Cada valor posee su propia celda

Se puede apreciar en las dos tablas mencionadas que la tabla de datos cumple con el formato *tidy*, pues las variables están dispuestas a lo largo de las columnas y las observaciones a lo largo de las filas.

Tabla 1: Primeras cinco filas y nueve columnas de la tabla de datos

Año	Población total	Población de hombres	Población de mujeres	Nacimientos	Defunciones	Defunciones infantiles	Defunciones neonatales	Defunciones fetales
1 950	868 934	438 185	430 749	37 248	9 769	3 358	974	1 103
1 951	897 630	452 702	444 928	39 239	9 631	3 467	953	1 062
1 952	929 173	468 613	460 560	42 461	9 902	3 811	1 083	1 012
1 953	962 485	485 464	477 021	42 817	10 312	4 017	1 161	1 021
1 954	997 535	503 219	494 316	48 157	9 713	3 856	1 210	1 135
1 955	1 035 424	522 378	513 046	48 903	9 998	4 065	1 233	1 095

Nota

El conteo de población se realiza al 30 de junio de cada año.

Tabla 2: Primeras cinco filas y segundas nueve columnas de la tabla de datos

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
31.62381	42.86632	11.242511	90.15249	26.14906	29.61233	-	-	-
32.98464	43.71400	10.729365	88.35597	24.28706	27.06491	-	-	-
35.04084	45.69763	10.656788	89.75295	25.50576	23.83364	-	-	-
33.77195	44.48589	10.713933	93.81788	27.11540	23.84567	-	-	-
38.53900	48.27600	9.737002	80.07143	25.12615	23.56874	-	-	-
37.57398	47.22993	9.655948	83.12373	25.21318	22.39126	6.96	-	-

Nota

El conteo de población se realiza al 30 de junio de cada año.

Tabla 3: 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 4: 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

Parte 2

Resumen de cinco números

En la tabla Tabla 3 se detalla el resumen de cinco números para algunas variables relevantes de la tabla de datos.

Medidas de dispersión

En la tabla Tabla 4 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 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.

Tabla 5: 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

Tablas de frecuencia

De la Tabla 5 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 tabla Tabla 6 se observa que los nacimientos en nuestro país a lo largo de los años tiene una distribución más uniforme para los últimos intervalos de la tabla.

Por otro lado, para el caso de la tasa de nacimientos se observa un comportamiento menos uniforme a diferencia de los observado para la tasa de mortalidad. En este caso, se observa que la frecuencia oscila entre valores menores a 6 pero con ciertos repuntes altos de frecuencia para ciertos intervalos por lo que la tasa de nacimientos no se mantiene en cierto intervalo específico.

Tabla 6: Frecuencia de nacimientos y tasa de natalidad

Nacimientos	Frec. de nacimientos	Tasa de nacimientos	Frec. de tasa
(3.72e+04,4.2e+04]	2	(11.3,15.4]	8
(4.2e+04,4.67e+04]	2	(15.4,19.6]	12
(4.67e+04,5.14e+04]	2	(19.6,23.7]	6
(5.14e+04,5.61e+04]	4	(23.7,27.8]	5
(5.61e+04,6.08e+04]	10	(27.8,31.9]	19
(6.08e+04,6.55e+04]	7	(31.9,36.1]	3
(6.55e+04,7.02e+04]	9	(36.1,40.2]	1
(7.02e+04,7.49e+04]	15	(40.2,44.3]	3
(7.49e+04,7.96e+04]	9	(44.3,48.4]	11
(7.96e+04,8.44e+04]	11	(48.4,52.6]	3

Parte 3: Propuesta de UVE

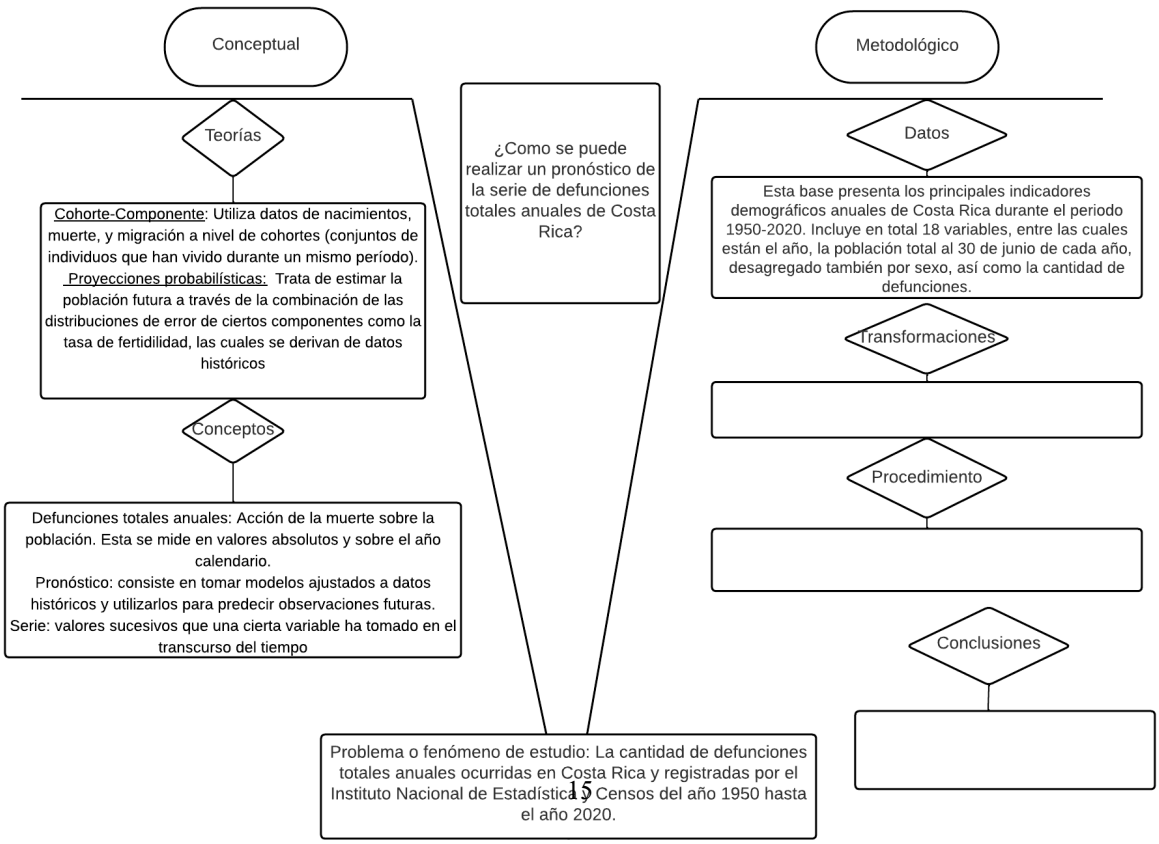


Figura 1: Borrador de la UVE Heurística

Bitacora 3

Distribución de la variable cuantitativa

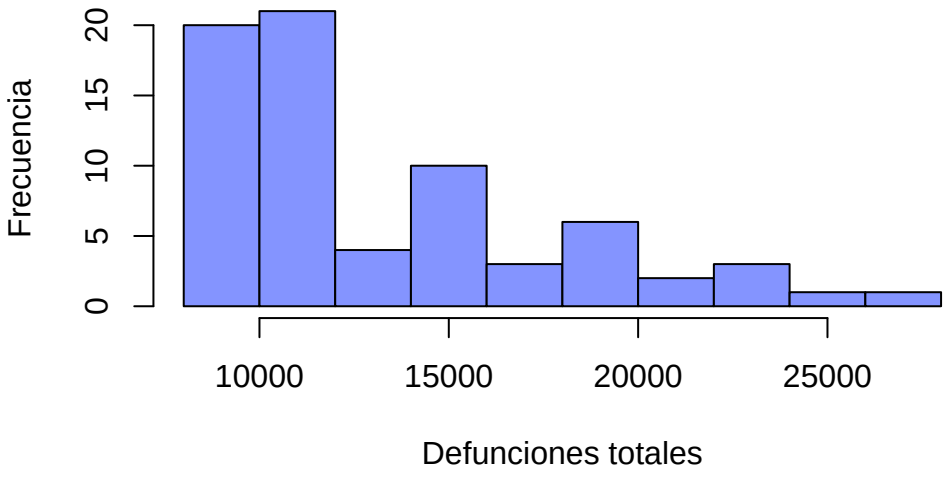


Figura 2: Histograma de las defunciones totales entre 1950 y 2020

La Figura 2 muestra la distribución de la variable cuantitativa de defunciones totales anuales, la cual es de principal interés en este trabajo. Se muestra poca simetría y una acumulación muy marcada en los cantidades más pequeñas, lo que quiere decir que la mayoría de años registrados presentaron defunciones totales de menos de 1500 al año.

Asoaciación de variables

En la Figura 3 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.

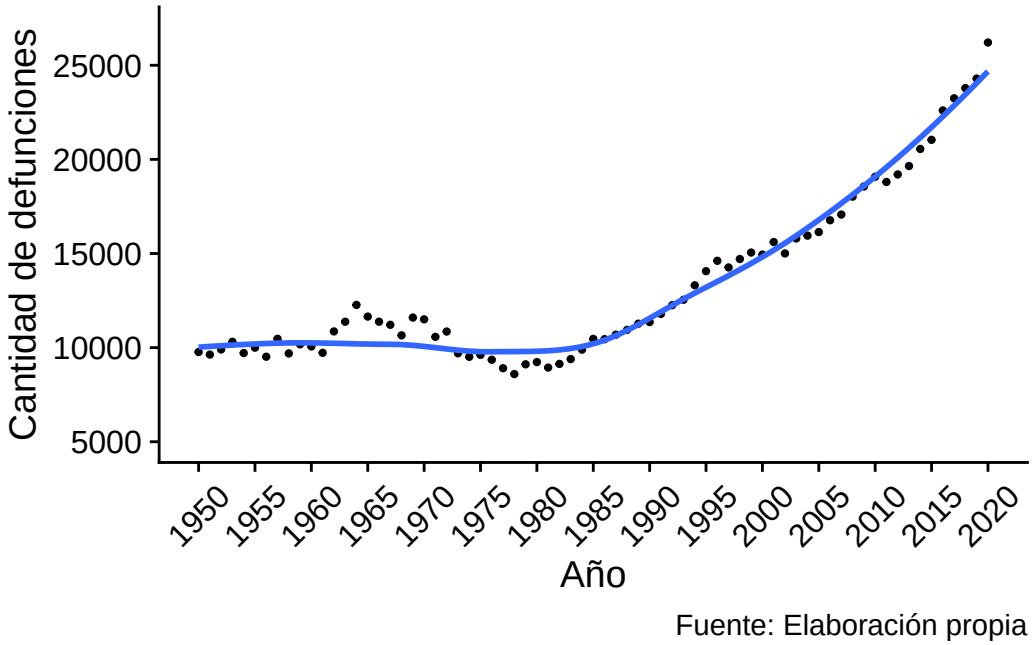


Figura 3: Cantidad de defunciones por año para el periodo 1950-2020

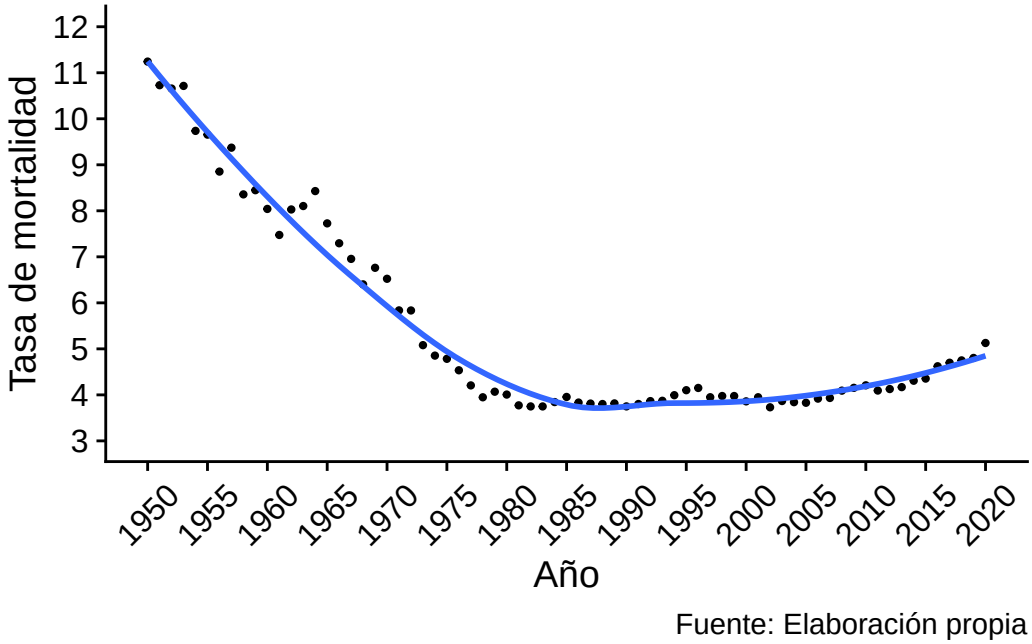


Figura 4: Tasa de mortalidad por año para el periodo 1950-2020

Este mismo comportamiento se puede observar en la tasa de mortalidad en la Figura 4 aunque de una forma mucho menos pronunciada, además, se puede apreciar que para el periodo 1950-1980 esta tasa decreció considerablemente.

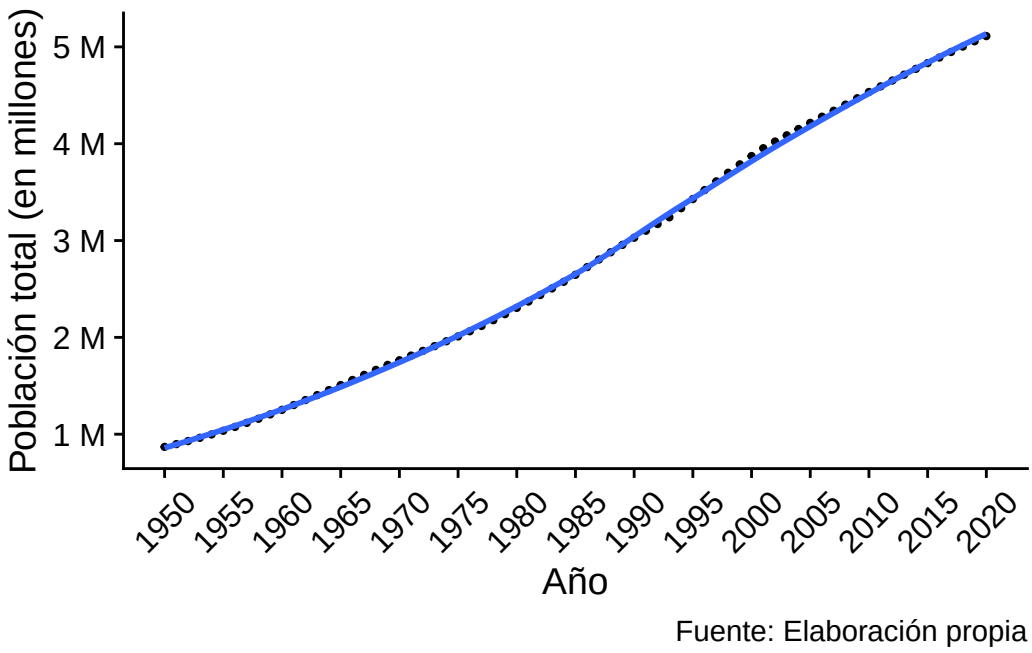


Figura 5: Población total por año para el periodo 1950-2020

Por su parte, en la Figura 5 se identifica una clara tendencia creciente de la población durante todo el periodo considerado.

Asimismo, de la Figura 6 se observa que la cantidad de nacimientos tuvo una tendencia creciente desde el inicio del periodo hasta cerca de 1990, en que empieza a descender hasta el último año considerado.

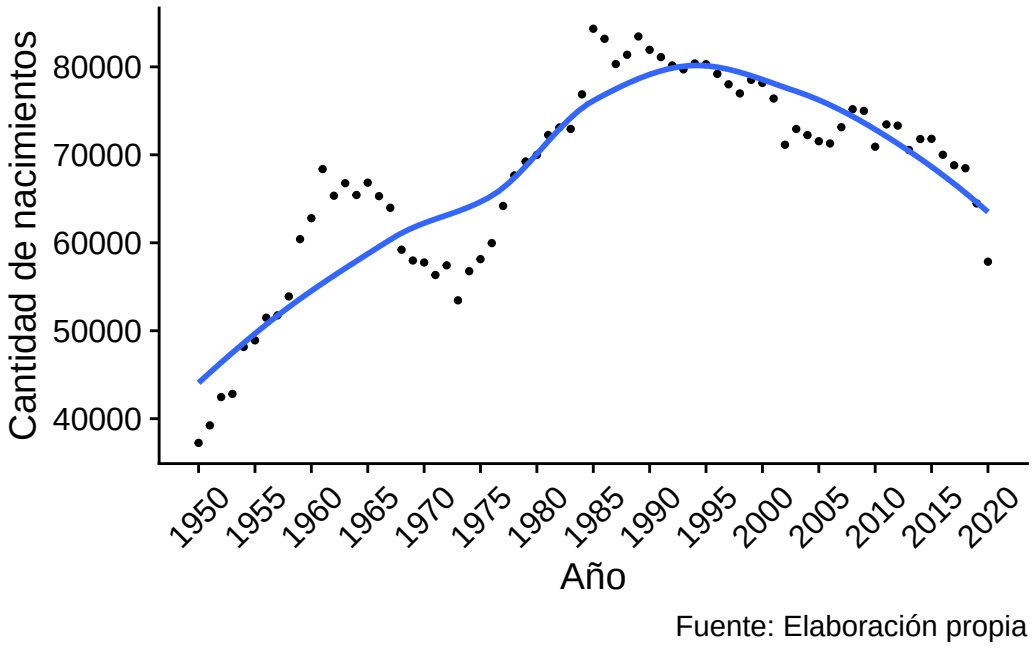
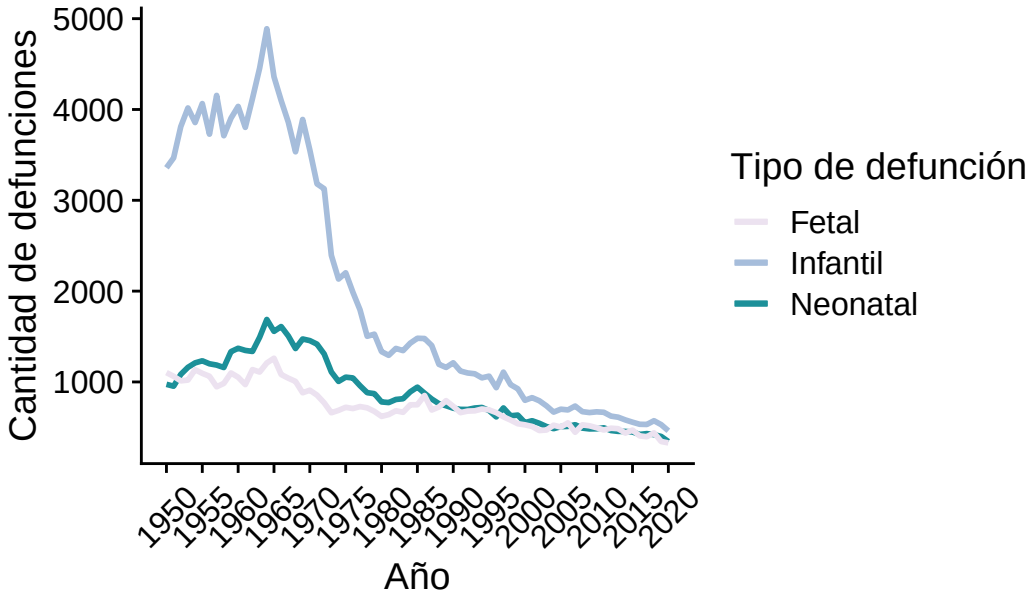


Figura 6: Cantidad de nacimientos por año para el periodo 1950-2020

En la Figura 7 se comparan las defunciones infantiles, neonatales y fetales. Cabe añadir que la distancia verticar 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 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.



Fuente: Elaboración propia con datos del INEC

Figura 7: Defunciones infantiles, neonatales y fetales por año

Para estudiar la asociación entre la tasa de defunciones neonatales y la de defunciones fetales, se crea una variable categórica indicadora de si la primera es mayor a la segunda. Dado que ambas tasas se calculan respecto al mismo denominador (la cantidad de nacimientos), esta variable equivale a hacer los mismo con la cantidad de defunciones respectiva. En la Figura 8 se muestra la distribución de esta variable y se ve que en 15 de los 71 años considerados la tasa de mortalidad fetal fue mayor, donde de la Figura 7 se sabe que esto sucedió mayoritariamente entre 1985 y 2020. En consecuencia, en 56 años fue mayor la tasa de mortalidad neonatal, lo cuál sucede sobre todo en los primeros 35 años a partir de 1950. Entonces, en los 71 años de estudio, 78.87% de las ocasiones fue mayor la tasa de mortalidad neonatal, contra un 21.13% de la fetal.

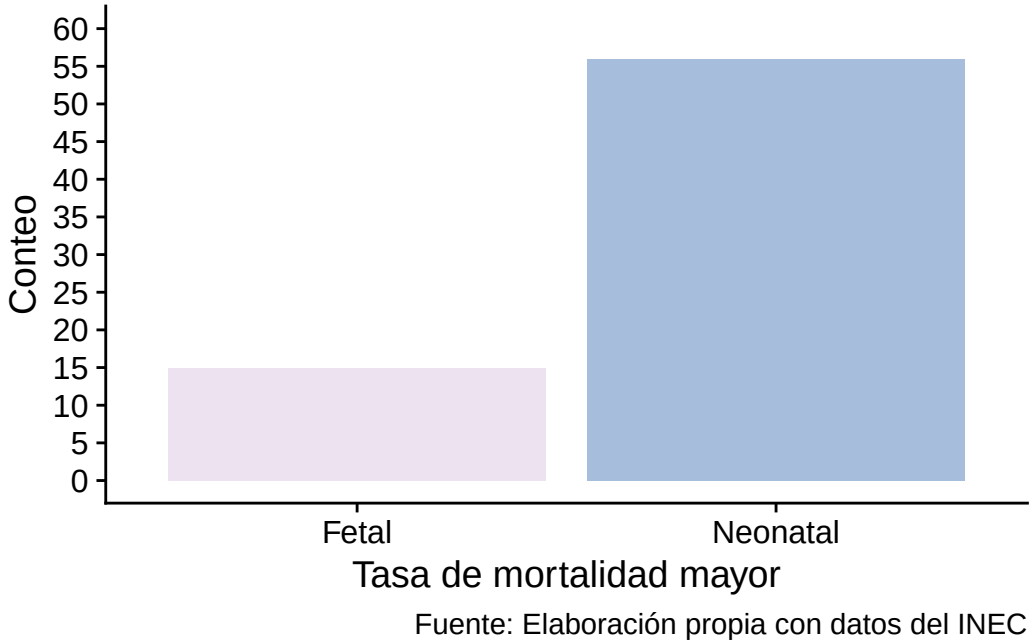


Figura 8: Distribución variable indicadora de la mayor tasa entre las de mortalidad fetal y la neonatal

Debido a que, por razones de escala, la cantidad de defunciones, nacimientos y población total resultan difíciles de comparar gráficamente, una alternativa se muestra en la Figura 9, donde se muestran las tasas de crecimiento, mortalidad y natalidad por año. Se identifica la gran similitud en el comportamiento de las tasas de crecimiento y natalidad, lo cuál no es ninguna sorpresa pues la segunda es una componente aditiva de la primera. Ahora bien, este gráfico permite apreciar que desde mediados del siglo XX y hasta cerca de 1980, tanto la tasa de crecimiento como la de mortalidad aparentan haber tenido una tendencia a la baja, lo cuál se rompe cerca de dicho año y la tasa de mortalidad empieza a empinarse ligeramente, mientras que la de crecimiento continúa en su tenencia con su tendencia decreciente. Nuevamente, se ve que el año 1980 parece haber un cambio en el comportamiento demográfico del país.

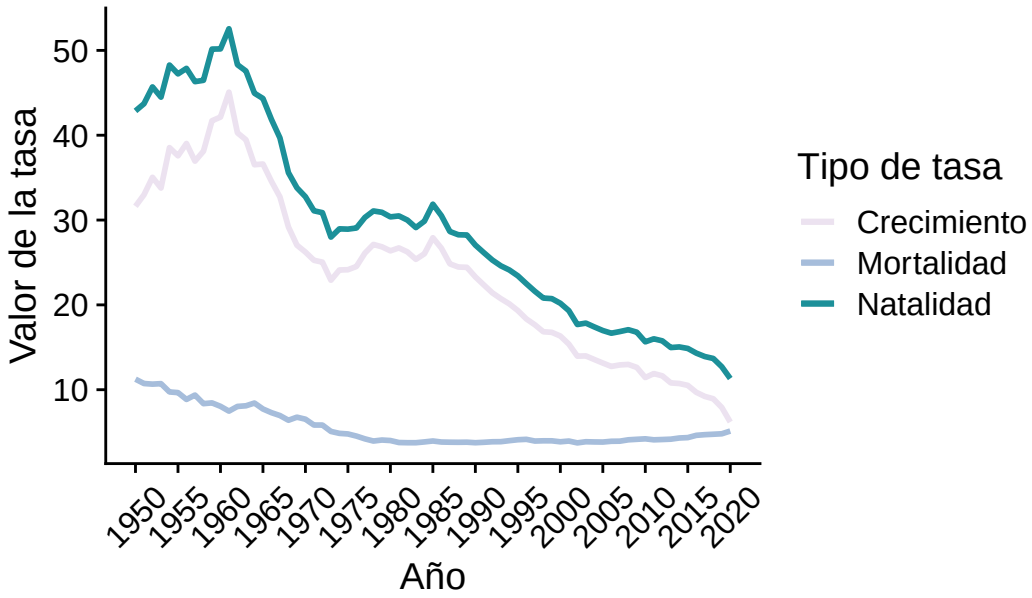


Figura 9: Tasas de crecimiento, mortalidad y natalidad por año

Para analizar la asociación de la tasa de mortalidad con la de crecimiento, se crea una variable categórica que indica si entre periodos consecutivos hubo coincidencia en la monotonía de ambas tasas, esto es, la variable es verdadera si ambas decrecieron o subieron, y es falsa si esto no se cumplió. En las figuras Figura 10 se muestra el histograma de esta variable, donde se ve que fue más común la no coincidencia en el comportamiento de las tasas entre periodos consecutivos.

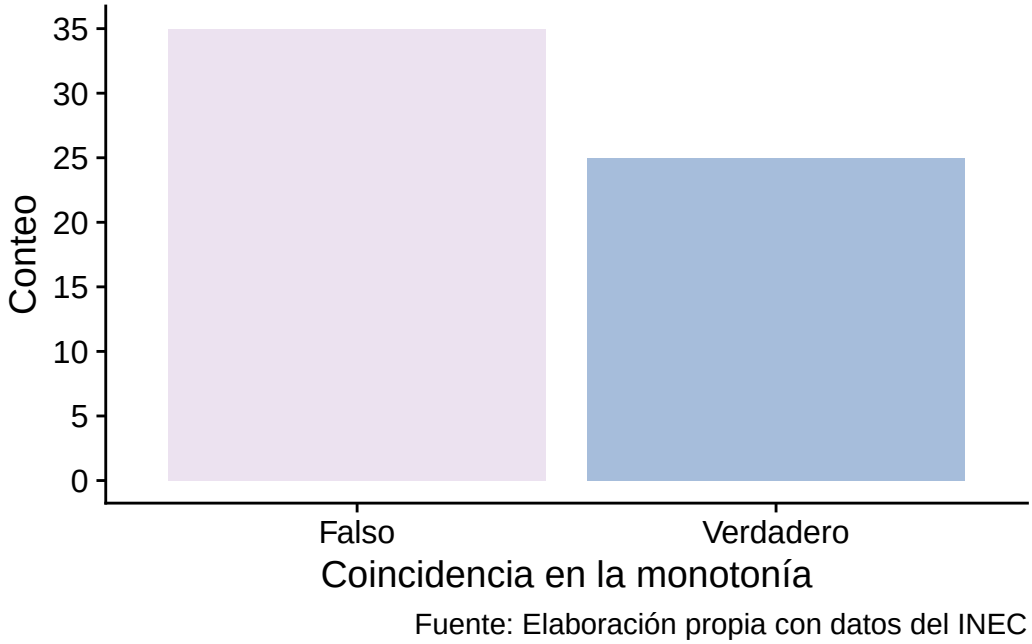
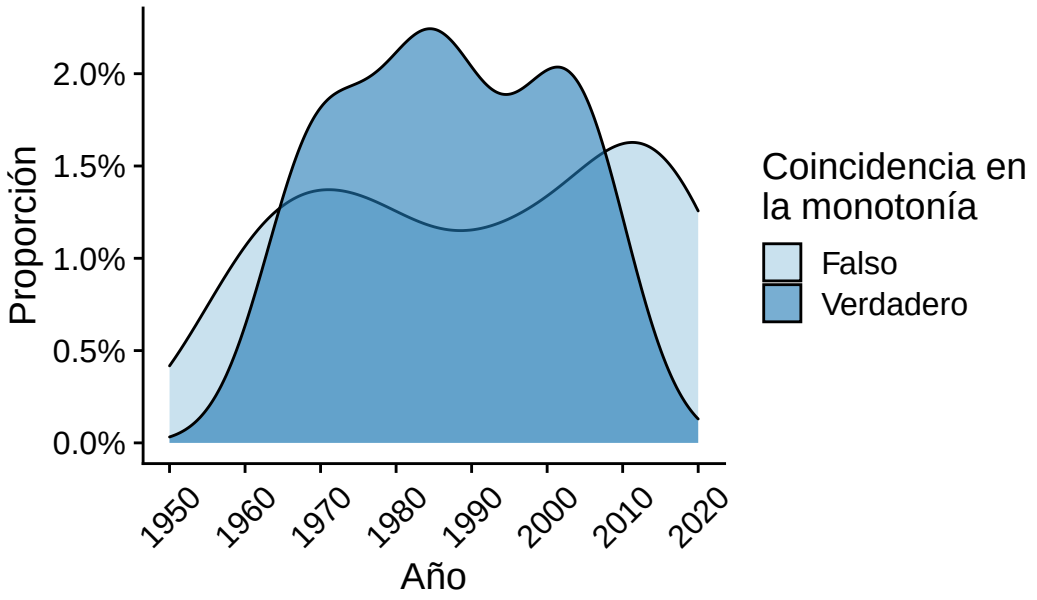


Figura 10: Histograma de la coincidencia en la monotonía de la tasa de mortalidad y la de crecimiento

En la Figura 11 se muestra la distribución de los años según la coincidencia en la monotonía de la tasa de mortalidad y la de crecimiento y se ve que esta fue relativamente uniforme. Hacia 2016, se afirmaba que "la actual tasa de mortalidad general de Costa Rica (4 muertes anuales por cada mil habitantes) es una de las más bajas del mundo, inferior incluso a la de los superdesarrollados países de la Europa Noroccidental (11 por mil) […] esta situación tan favorable se debe, en parte, a una estructura por edades muy particular de la población costarricense, caracterizada por una alta concentración en las edades donde la mortalidad es más baja (adultos jóvenes)" (Rosero-Bixby, 2016).



Fuente: Elaboración propia con datos del INEC

Figura 11: Distribución de los años según la coincidencia en la monotonía de la tasa de mortalidad y la de crecimiento

Se realiza el mismo ejercicio con la población total y las defunciones totales, lo que resulta en el histograma en la Figura 12 y la distribución de los años en la Figura 13.

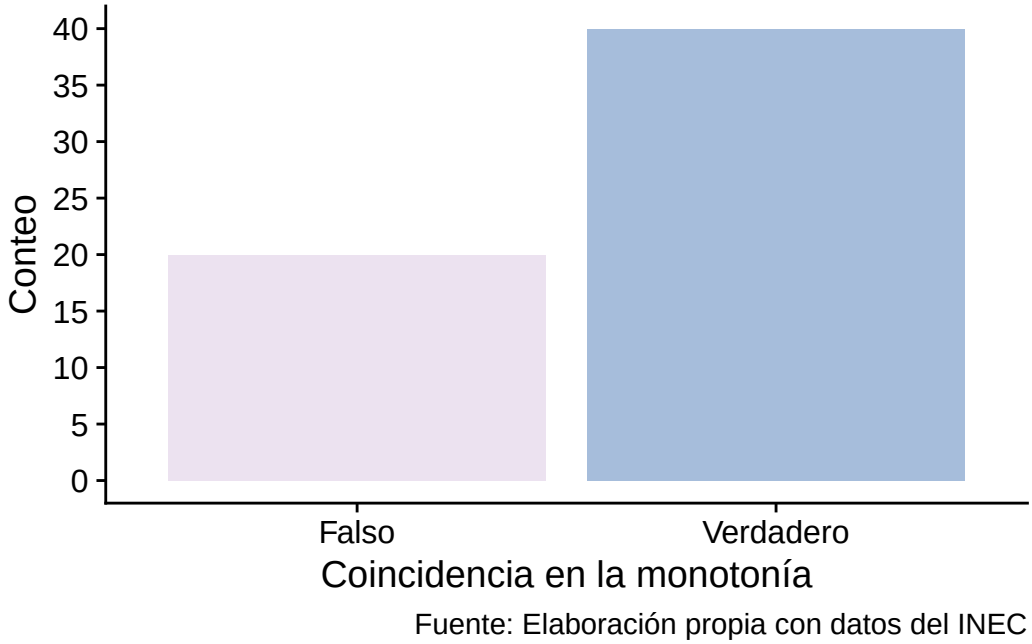
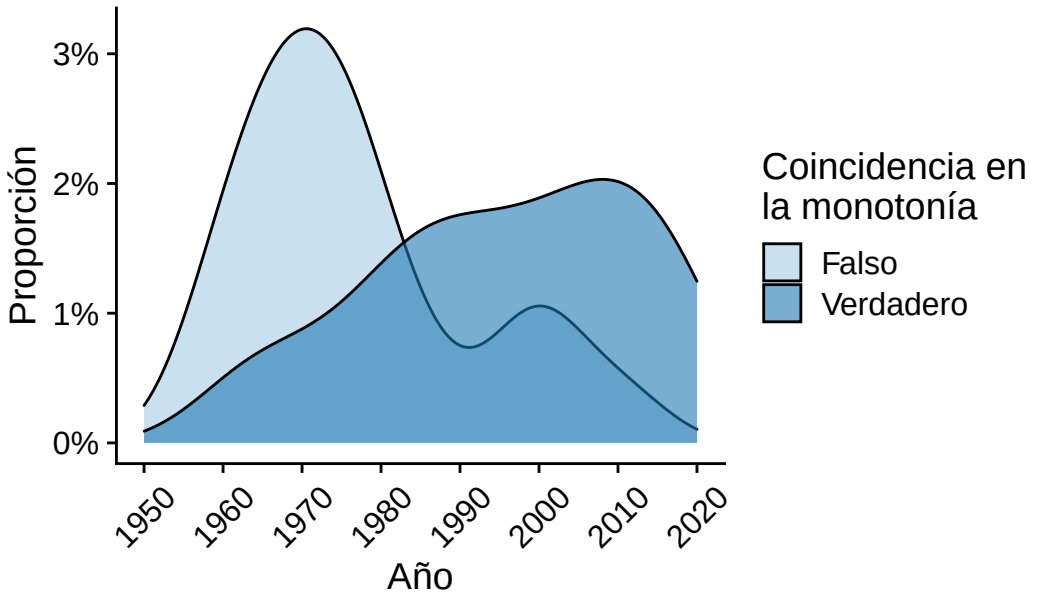


Figura 12: Histograma de la coincidencia en la monotonía de la población total y las defunciones totales

De la primera se ve que fue más común la coincidencia en este caso, y de la segunda que es relativamente marcado que la coincidencia fue falsa de mediados de los años ochenta para atrás, y fue más común que la coincidencia se diera desde cerca de 1980. Nuevamente, se haya evidencia descriptiva de que en la década de los ochenta hay algún tipo de cambio, lo que debe tenerse en cuenta en la sección metodológica, quizás restringiendo el periodo de estudio.



Fuente: Elaboración propia con datos del INEC

Figura 13: Distribución de los años según la coincidencia en la monotonía de la población total y la cantidad de defunciones

Se vuelve a proceder de la misma manera, esta vez con las defunciones infantiles y las defunciones totales, lo que resulta en el histograma en la Figura 14 y la distribución de los años en la Figura 15.

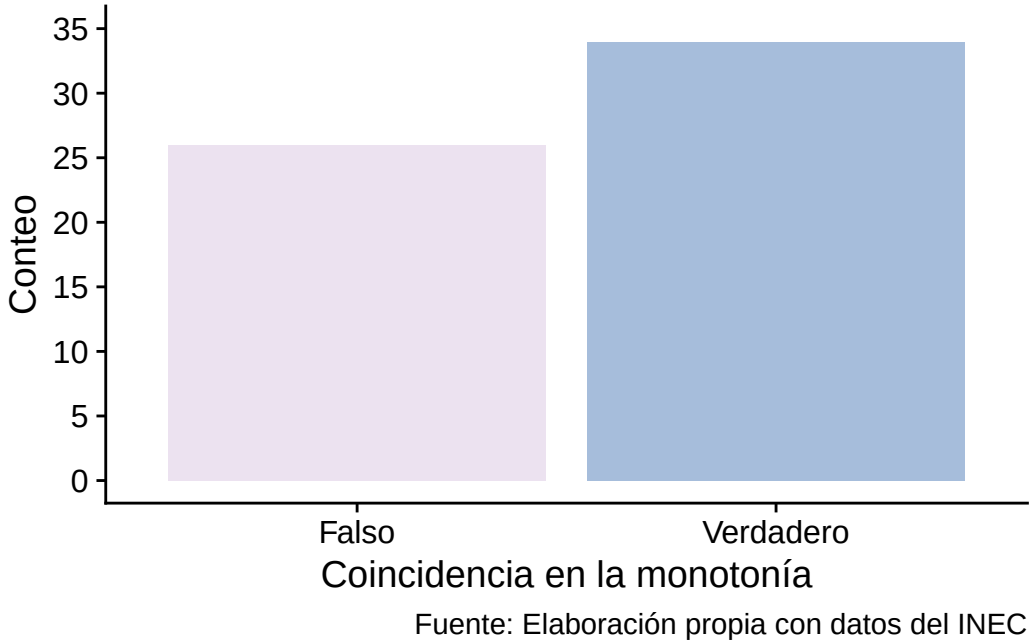
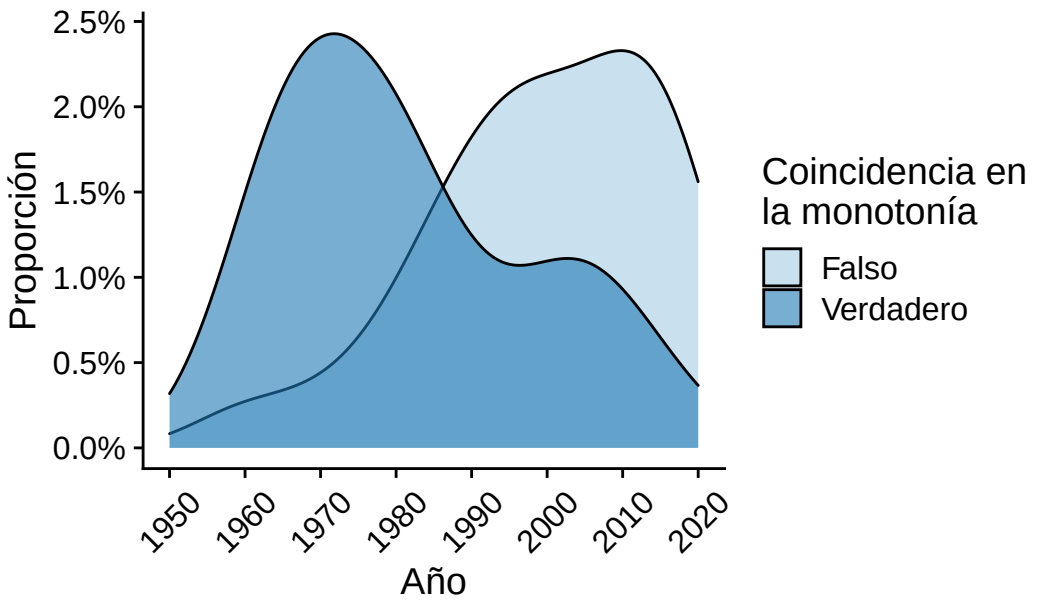


Figura 14: Histograma de la coincidencia en la monotonía de las defunciones infantiles y las defunciones totales

En este caso, los conteos están más parejos pero ocurrió más que sí se diera la coincidencia.



Fuente: Elaboración propia con datos del INEC

Figura 15: Distribución de los años según la coincidencia en la monotonía de las defunciones infantiles y la cantidad de defunciones

Además, de las distribuciones se aprecia que la discordancia se dio cerca de 1980 hasta el final, mientras que la coincidencia ocurrió sobre todo en años anteriores a 1990. De nuevo, 1980 parece ser un punto de corte pertinente pues aparenta marcar un cambio en el comportamiento de las defunciones totales.

Descripción del modelo o metodología

Con la finalidad de realizar un pronóstico de la serie de defunciones totales anuales de Costa Rica, se desea implementar el modelo estadístico que mejor se ajuste a los datos.

Para nuestro estudio en cuestión, se ha optado por realizar una implementación de Modelos de Espacio-Estado.Particularmente, Modelos Dinámicos Lineales (DLM).

Tal como lo establece Petris et al. (2007), estos últimos son una clase de Modelos de Espacio-Estado también llamados Modelos de Espacio-Estado Lineales Gaussianos. Estos modelos son especificados mediante dos ecuaciones, para $t \geq 1$ se tiene:

$$Y_t = F_t\theta_t + v_t,$$
$$\theta_t = G_t\theta_{t-1} + w_t$$

Donde la primer ecuación es llamada ecuación de observación, y la segunda ecuación estado o ecuación del sistema.

Es importante señalar que F_t y G_t son matrices y (v_t), (w_t) son secuencias de ruidos blancos independientes tales que:

$$v_t \sim \mathcal{N}_m(0, V_t),$$
$$w_t \sim \mathcal{N}_p(0, W_t)$$

Los DLM 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.

La estimación y pronóstico se pueden resolver calculando las distribuciones condicionales de las cantidades de interés, dada la información disponible. Para estimar el vector de estados es necesario computar la densidad condicional $p(\theta_t|y_1, .., y_t)$. En particular, nos interesa el problema de filtrado (cuando $s = t$), donde los datos se supone que llegan secuencialmente en el tiempo.

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:

$$a_{t+k} = G_{t+k}m_t$$
$$R_{t+k} = G_{t+k}C_{t+k}G_{t+k}^{'} + W_{t+k}$$

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:

$$f_{t+k} = F_{t+k}a_{t+k}$$
$$Q_{t+k} = F_{t+k}R_{t+k}F_{t+k}^{'} + V_{t+k}$$

La densidad de filtrado de θ_{t+k} con $k \geq 1$ dada la información pasada D_{t+k}, es Gaussiana con media y varianza condicional dadas respectivamente por:

$$m_{t+k} = a_{t+k} + R_{t+k}F_{t+k}^{'}Q_{t+k}^{-1}e_{t+k}$$
$$C_{t+k} = R_{t+k} - R_{t+k}F_{t+k}^{'}Q_{t+k}^{-1}F_{t+k}R_{t+k}$$

Propuesta y justificación modelos DLM

Como se mencionó en Figura 3, la cantidad de defunciones totales siguen una cierta tendencia lineal creciente, en particular para años posteriores a 1980.

Debido a que esta es nuestra variable de interés para realizar un pronóstico, es propicio para nuestro estudio en cuestión la implementación de un modelo con supuesto de linealidad, como se mencionó justamente los DLM siguen este supuesto.

Para llevar a cabo los pronósticos se proponen por tanto tres métodos estadísticos pertenecientes a los DLM, estos son: modelo DLM polinomial de primer orden, modelo DLM polinomial de segundo orden y el modelo ARIMA.

Se propone un modelo DLM de primer orden ya que como establece Mary (2006) los DLM de primer orden son algoritmos recomendados al lidiar con datos anuales debido a que las series de tiempo es corta y no presentan patrones estacionales.Dado que nuestros datos son anuales, este modelo se presenta como un posible candidato.

Por su parte Mary (2006), señala que los DLM de segundo orden son útiles para describir tendencias. Dada la tendencia observada de la serie de defunciones totales sugiere por tanto realizar un modelo polinomial de segundo orden.

Primera implementación

Como una primer implementación se utilizará un modelo ARIMA. Tal como lo menciona Petris et al. (2007) un modelo ARIMA puede ser considerado un DLM, esto ya que es posible representar todo modelo ARIMA (ya sea univariado o mulltivariado) como un DLM.

La escogencia de este modelo al ser un DLM, sigue la misma línea de justificación antes mencionado sobre la elección de modelos DLM para nuestro estudio, siendo este un caso particular de estos.

Sin embargo, es importante mencionar que la escogencia de este modelo como primera implementación también se basa en su simplicidad, y en que dada la bibliografía consultada, se observa que en múltiples investigaciones con temáticas relacionadas a nuestro estudio como el de Adekanmbi et al. (2014) y el estudio por Ordorica (2004), se implementa este tipo de modelo.

No obstante, según Petris et al. (2007) estos modelos proporcionan un enfoque de caja negra para el análisis de datos, ofreciendo la posibilidad de predecir observaciones futuras, pero con una interpretabilidad muy limitada del modelo ajustado.

Por lo que para bitácoras posteriores se describirán en detalle e implementaran dos modelos de mayor complejidad pero con una mejor interpretabilidad, como lo son los DLM polinomiales de primer y segundo orden, antes mencionados.

En cuanto a la implementación del primer intento utilizando ARIMA, como se quiere hacer una comparación entre los modelos, se usará una base de entrenamiento y una de prueba para hacer el diagnóstico y cuyos valores ya son conocidos para poder valorar la eficacia de dicho modelo. Se quiere hacer el pronóstico a dos años, por lo que la base de entrenamiento serán los datos entre 1950 y 2018, y se pronosticará las defunciones totales para el 2019 y 2020.

```
entrenamiento <- base$defunciones[1:69]
serie <- ts( entrenamiento, start=c(1950,1), frequency=1)
```

Primero, se grafica la serie de tiempo de defunciones totales. Para emplear el modelo se requiere estacionareidad.

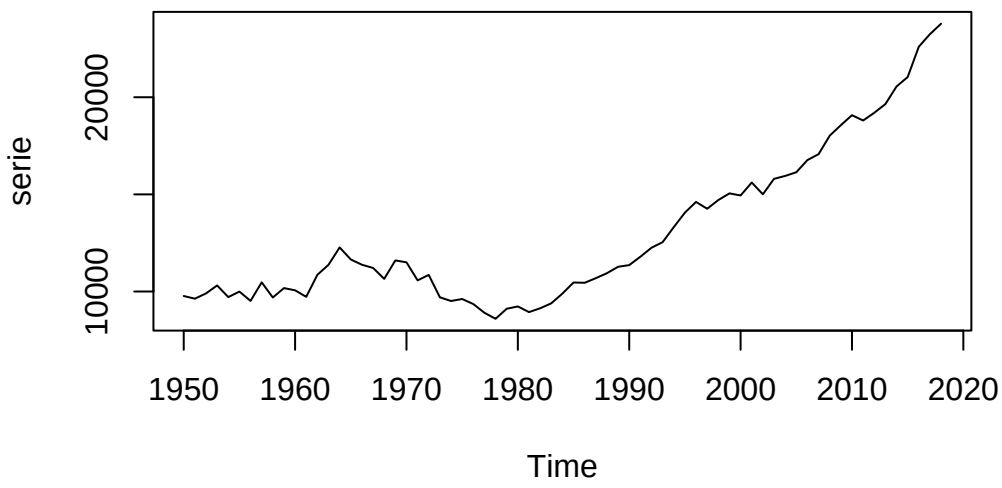


Figura 16: Serie de tiempo de defunciones totales

```
     [,1]  [,2]  [,3] [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11] [,12]
ACF  0.93  0.86  0.80 0.74  0.69  0.64  0.59  0.54  0.49  0.43  0.38  0.34
PACF 0.93 -0.02 -0.04 0.06 -0.02 -0.01 -0.02 -0.03 -0.08  0.00 -0.03  0.01
     [,13] [,14] [,15] [,16] [,17] [,18] [,19]
ACF    0.3  0.26  0.23  0.20  0.17  0.13   0.1
PACF   0.0  0.00  0.02 -0.05  0.02 -0.09   0.0
```

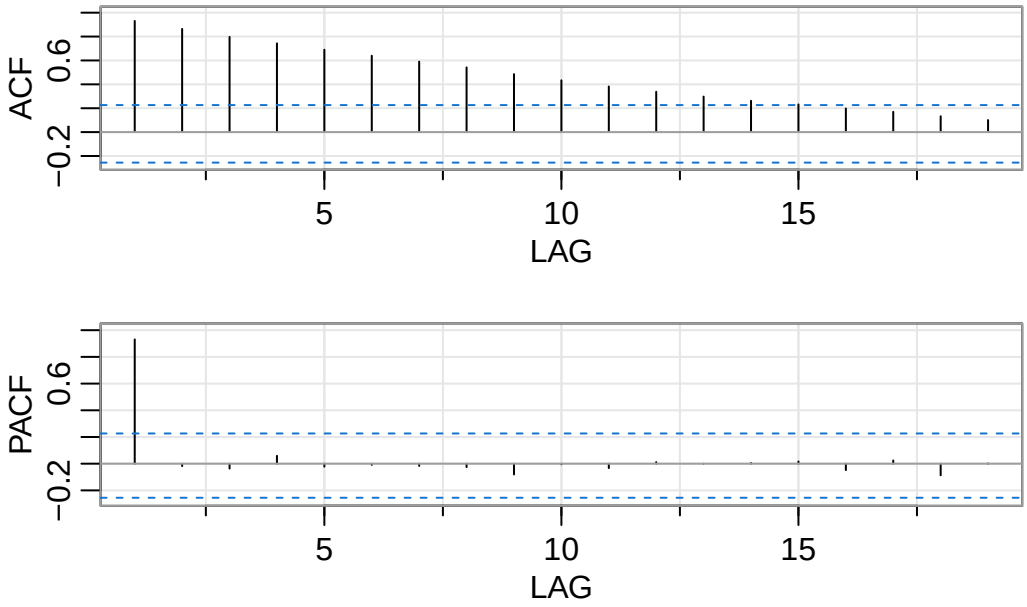


Figura 17: ACF y PACF de la serie de defunciones totales

Se observa de la Figura 24 no estacionareidad muy clara, además de que en la Figura 25 se aprecia que la decadencia de las correlaciones no son suficientemente rápidas. Se lleva a cabo una diferenciación

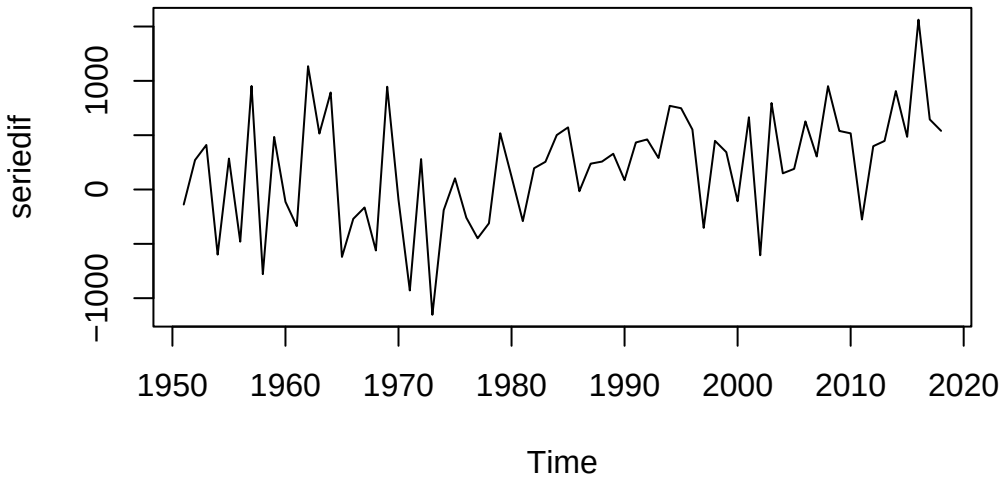


Figura 18: Serie de tiempo de defunciones totales con una diferencia

```
     [,1] [,2] [,3]  [,4] [,5] [,6] [,7]  [,8]  [,9] [,10] [,11] [,12] [,13]
ACF     0 0.29 0.11  0.00 0.28 0.02 0.26 -0.01  0.03  0.18 -0.05  0.07  0.08
PACF    0 0.29 0.12 -0.09 0.23 0.04 0.14 -0.08 -0.07  0.15 -0.03 -0.13  0.11
     [,14] [,15] [,16] [,17] [,18] [,19]
ACF  -0.09  0.11  0.00  0.04 -0.08  0.02
PACF -0.10  0.02  0.08 -0.03 -0.14  0.10
```

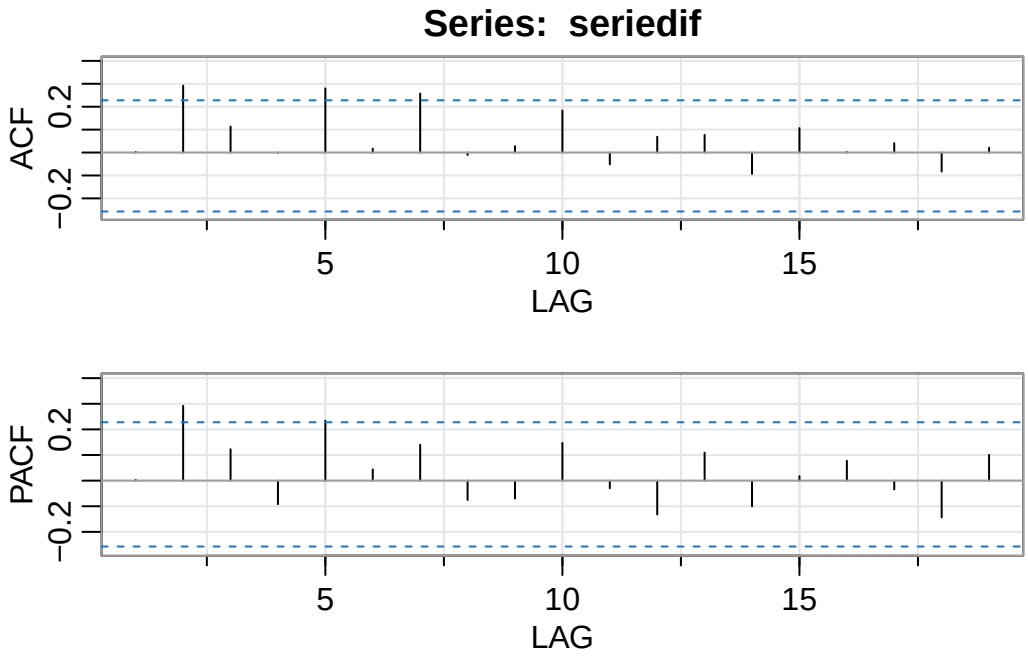


Figura 19: ACF y PACF de la serie de defunciones totales con una diferencia

Se observa de la Figura 18 que aún hay evidencia de no estacionareidad por lo que se lleva a cabo una segunda diferenciación

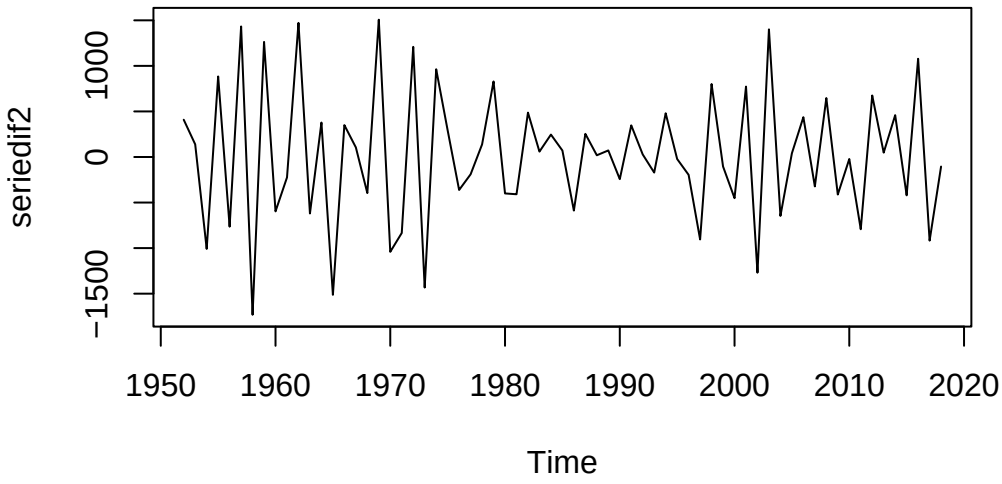


Figura 20: Serie de tiempo de defunciones totales con dos diferencias

```
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6] [,7]  [,8]  [,9] [,10] [,11] [,12]
ACF  -0.65  0.23 -0.03 -0.20  0.27 -0.24 0.25 -0.15 -0.07  0.19 -0.16  0.05
PACF -0.65 -0.33 -0.09 -0.37 -0.15 -0.22 0.01  0.02 -0.20 -0.04  0.09 -0.13
     [,13] [,14] [,15] [,16] [,17] [,18] [,19]
ACF   0.10 -0.20  0.15 -0.06  0.07 -0.09  0.01
PACF  0.08 -0.06 -0.13 -0.01  0.09 -0.15 -0.12
```

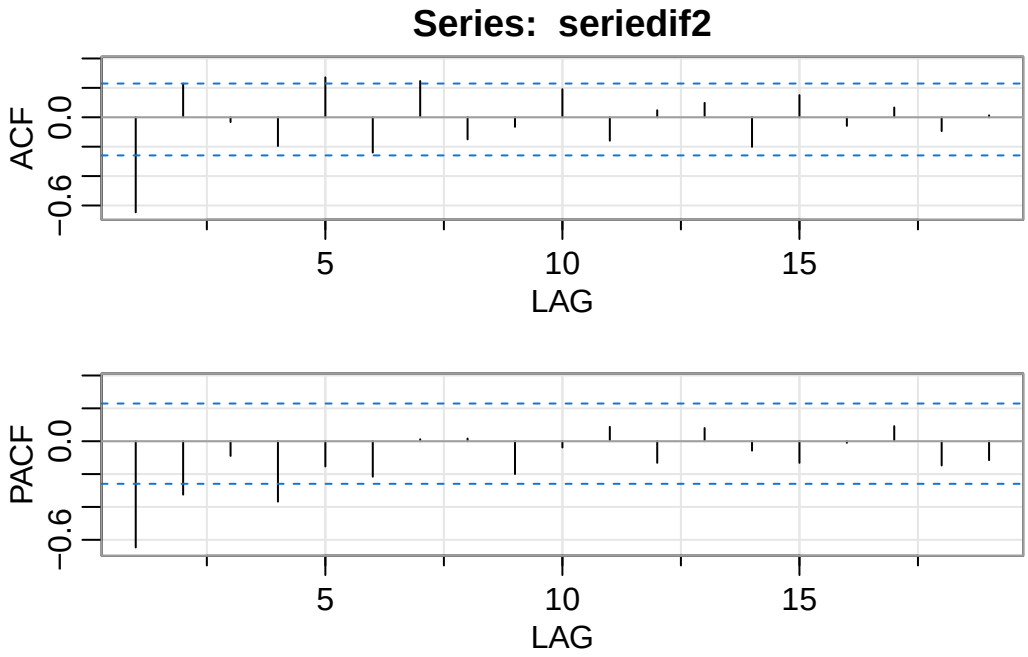


Figura 21: ACF y PACF de la serie de defunciones totales con dos diferencias

En la Figura 20 se aprecia que con dos diferencias ya se logra estacionareidad. Además, de la Figura 21 se nota que tiene una decadencia gradual en las correlaciones y las autocorrelaciones parciales luego del primer rezago por lo que implica un modelo ARMA(1,1). Se lleva a cabo la implementación de este modelo con dos rezagos.

```
modelo <- arima(serie, order=c(1, 2, 1))
```

y se presentan algunos diagnósticos.

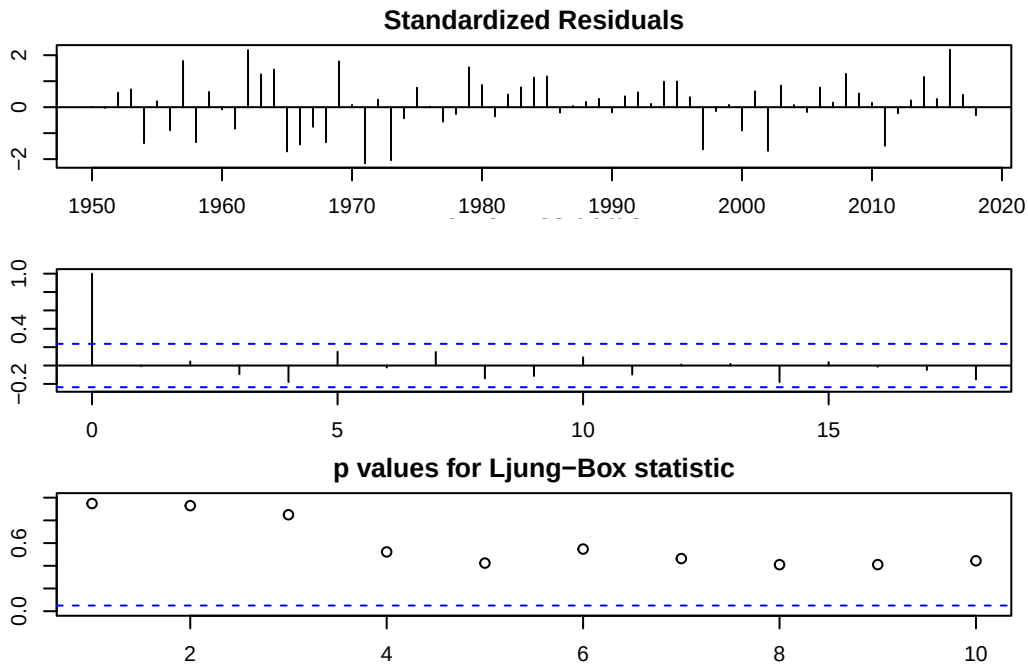


Figura 22: Diagnósticos del modelo ARIMA implementado

En la Figura 22 se aprecia que los residuos no parecen mostrar ningún patrón significante y la varianza aparenta ser constante. El gráfico de ACF confirma que no existe correlación en los residuos. En el último gráfico se muestra que bajo la prueba de Ljung-Box existe evidencia de que no hay ninguna estructura remanente en los residuos, lo que apoya la selección de este modelo. También se puede utilizar la función de auto.arima la cual ajusta modelos ARMA para diferentes escogencias de p y q y los compara utilizando criterios de AIC y BIC para confirmar la elección de este modelo

```
modelo1 <- auto.arima(serie)
summary(modelo1)
```

```
Series: serie
ARIMA(1,2,1)

Coefficients:
          ar1      ma1
      -0.2559  -0.8001
s.e.   0.1398   0.1023

sigma^2 = 252435:  log likelihood = -511.49
AIC=1028.98   AICc=1029.36   BIC=1035.59

Training set error measures:
```

Tabla 7: Pronóstico e intervalos de confianza del modelo ARIMA

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2019	24474.97	23840.77	25109.17	23505.04	25444.90
2020	25125.82	24253.67	25997.98	23791.97	26459.67

```
                    ME     RMSE      MAE       MPE     MAPE      MASE
Training set 57.29695 487.6486 381.1041 0.355309 3.194274 0.8045412
                     ACF1
Training set -0.007621673
```

Finalmente, con este modelo se lleva a cabo el pronóstico a dos años.

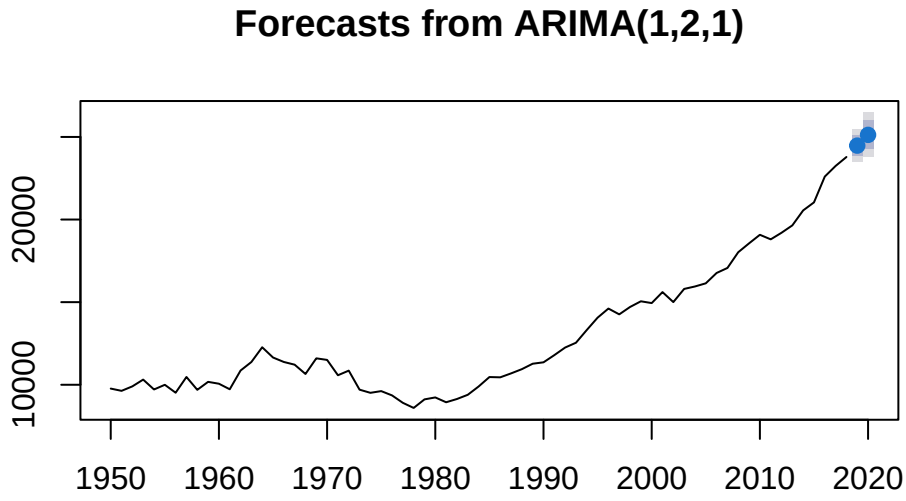


Figura 23: Pronóstico a dos años utilizando el modelo ARIMA

La Tabla 7 resumen los pronósticos e intervalos de confianzo utilizando el modelo ARIMA que se pueden comparar con los valores reales de 24292 para el 2019 y 26209 para el 2020. Es importante recordar que el 2020 fue anómalo por ser el primer año del COVID-19 en Costa Rica lo que ocasionó más muertes. Áún así se aprecia que para ambos años se obtuvieron valores que caen dentro del intervalod de confianza 95%. La diferencia entre el pronóstico y el valor real en valor absoluto es dado por

```
[1] 182.9708
```

```
[1] 1083.177
```

Bitácora 4

Segundo intento de inferencia

ARIMA

En la bitácora anterior se hizo un ajuste de un modelo ARIMA a la serie de defunciones totales. En esta bitácora nuevamente se hará un ajuste similar, y a pesar de que el modelo escogido es el mismo, se hará un análisis más certero sobre los grados implicados por los gráficos de ACF y PACF empíricos, los diagnósticos de los modelos y se hará una comparación entre tres propuestas. Este último elemento fue particularmente faltante en la última bitácora.

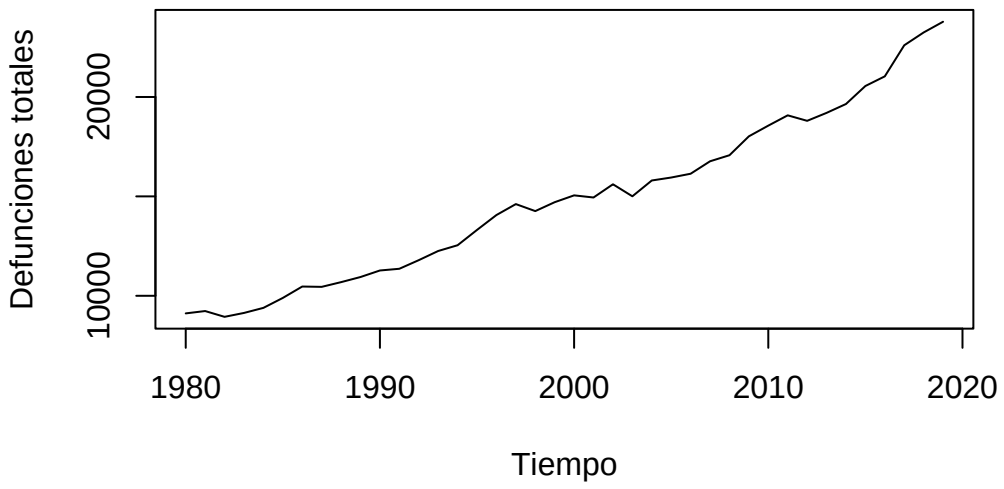


Figura 24: Serie de tiempo de defunciones totales

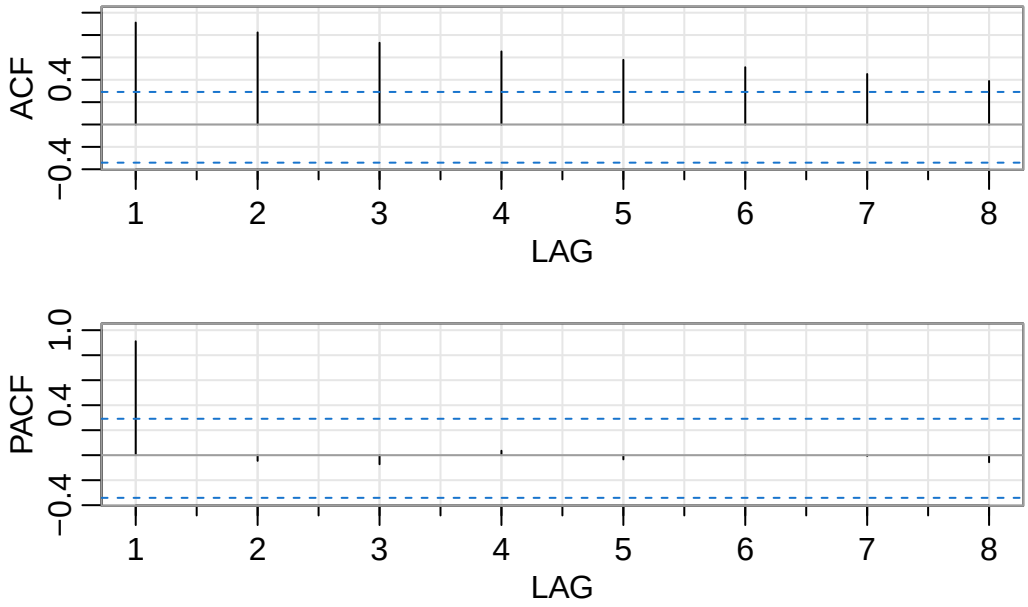


Figura 25: ACF y PACF de la serie de defunciones totales con una diferencia

En la Figura 24 se observa que la serie de defunciones totales no es estacionaria, pues tiene una tendencia creciente muy clara. Esto se confirma con el gráfico en la Figura 25 que muestra que la decadencia de las correlaciones no es suficientemente rápida (lineal). Se lleva a cabo un diferencia para tratar de eliminar la tendencia.

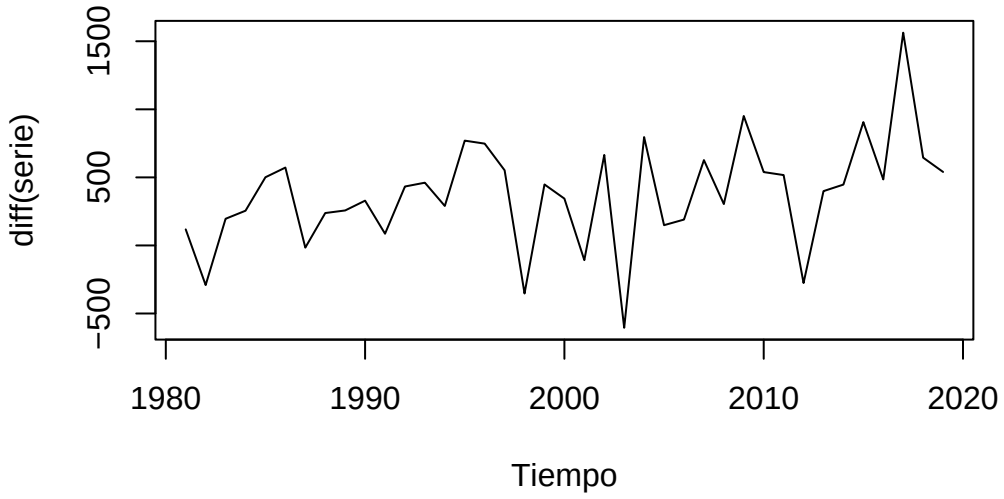


Figura 26: Serie de tiempo de defunciones totales con una diferencia

En la Figura 26 no queda claro que la tendencia ha sido eliminado, de hecho parece ser creciente. Más aún, la varianza no aparenta estable, pues se nota más varianza a partir de 1998.

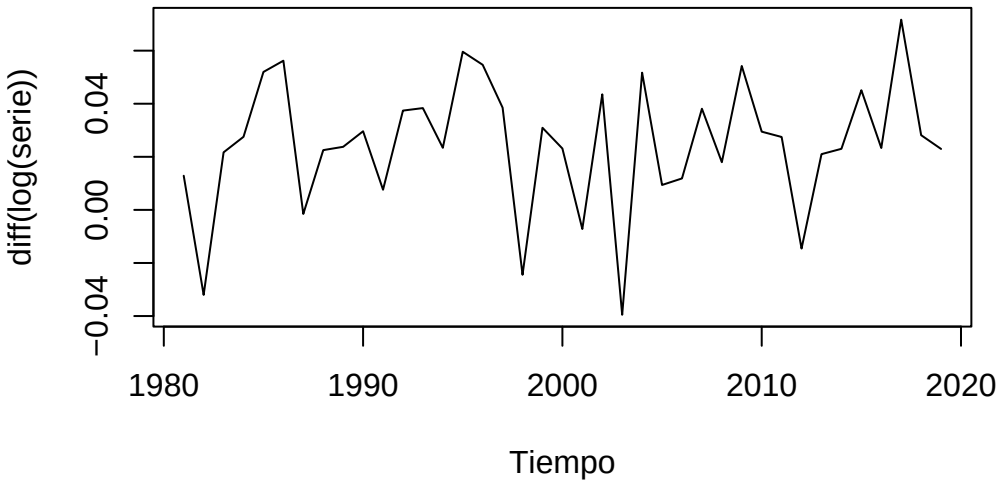


Figura 27: Serie de tiempo de la tasa de crecimiento de defunciones

La serie resultante mostrada en la Figura 27 parece ser estable, estacionaria e interpretable como la tasa de crecimiento porcentual de las defunciones. Esto se confirma al observa el ACF y PACF en la Figura 28

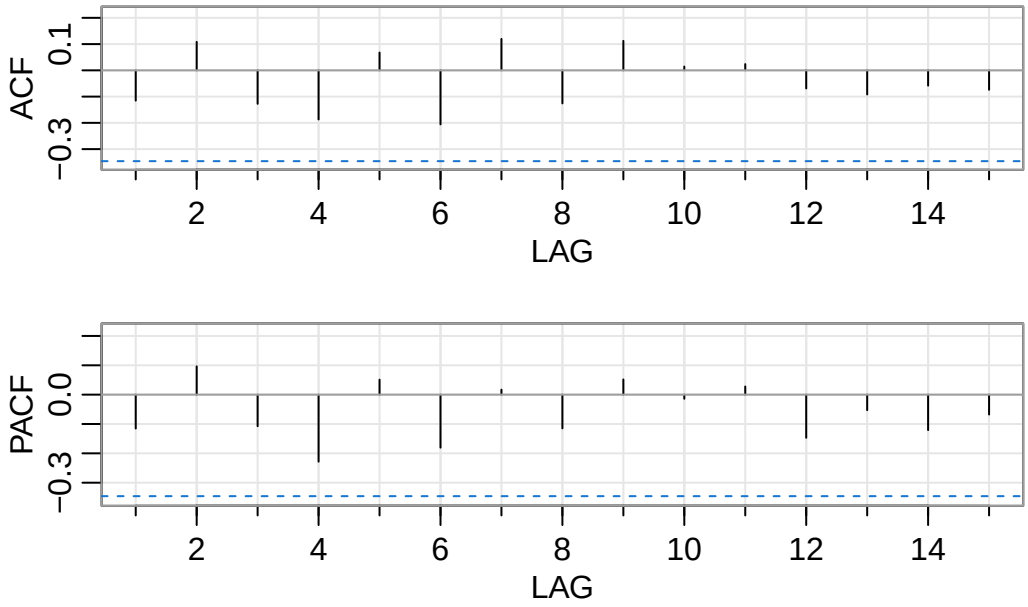


Figura 28: ACF y PACF de la serie de tasa de crecimiento de defunciones

El decrecimiento gradual del ACF y PACF implica un modelo ARMA(1,1) por lo que los modelos propuestos son ARIMA(1,1,0), ARIMA(0,1,1), ARIMA(1,1,1).

```
initial  value -3.713936
iter   2 value -3.719482
iter   3 value -3.719616
iter   4 value -3.719625
iter   5 value -3.719628
iter   5 value -3.719628
iter   5 value -3.719628
final  value -3.719628
converged
initial  value -3.719567
iter   2 value -3.719567
iter   2 value -3.719567
iter   2 value -3.719567
final  value -3.719567
converged
```

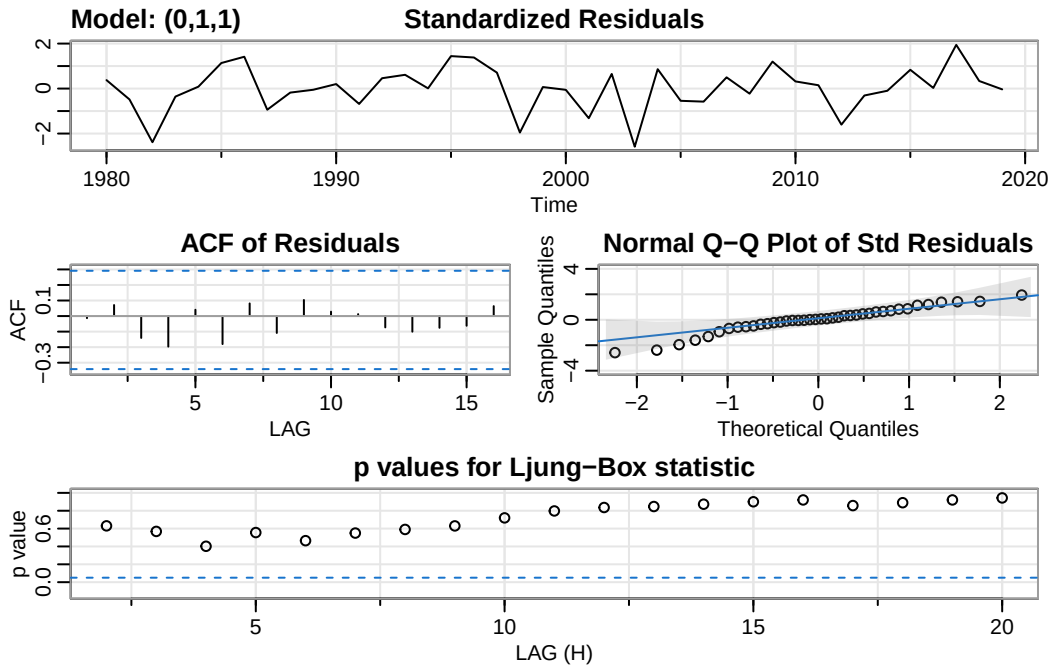


Figura 29: Diagnósticos del modelo ARIMA(0,1,1)

```
initial  value -3.703926
iter   2 value -3.710476
iter   3 value -3.710697
iter   4 value -3.710752
iter   4 value -3.710752
iter   4 value -3.710752
final  value -3.710752
converged
initial  value -3.720446
iter   2 value -3.720503
iter   3 value -3.720512
iter   3 value -3.720512
iter   3 value -3.720512
final  value -3.720512
converged
```

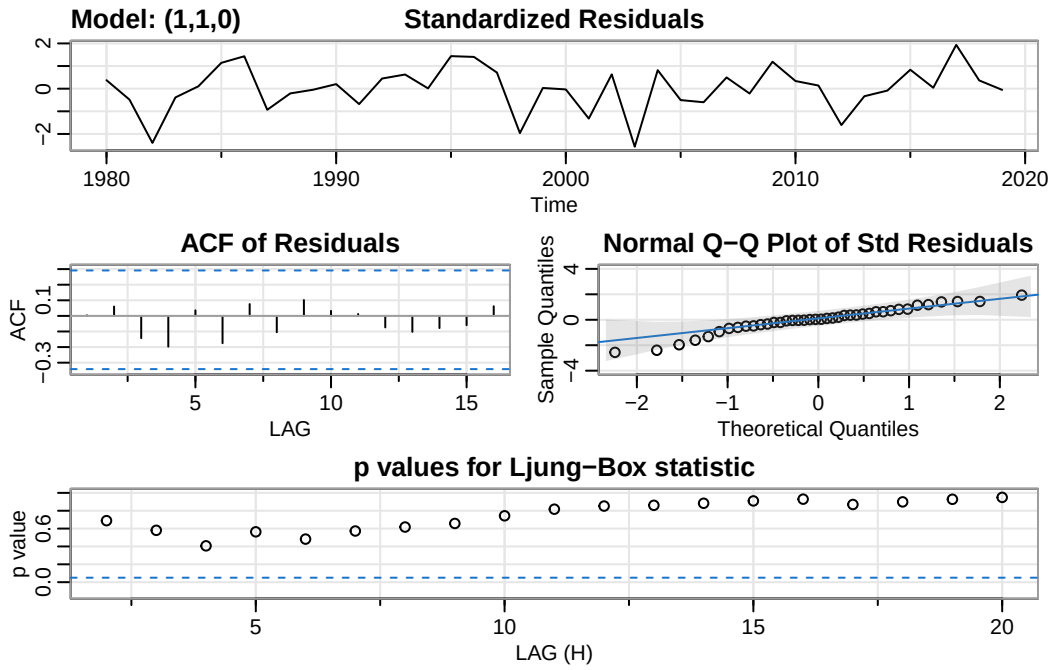


Figura 30: Diagnósticos del modelo ARIMA(1,1,0)

```
initial  value 5.978026
iter   2 value 5.977628
iter   3 value 5.977066
iter   4 value 5.976009
iter   5 value 5.969474
iter   6 value 5.965849
iter   7 value 5.958835
iter   8 value 5.957462
iter   9 value 5.955837
iter  10 value 5.953799
iter  11 value 5.951199
iter  12 value 5.943677
iter  13 value 5.906060
iter  14 value 5.856319
iter  15 value 5.856082
iter  16 value 5.853297
iter  17 value 5.851340
iter  18 value 5.839800
iter  19 value 5.828115
iter  20 value 5.816811
iter  21 value 5.811234
iter  22 value 5.804361
```

```
iter  23 value 5.799263
iter  24 value 5.799179
iter  25 value 5.798691
iter  25 value 5.798691
iter  26 value 5.797929
iter  27 value 5.797924
iter  28 value 5.797562
iter  29 value 5.797507
iter  30 value 5.797293
iter  31 value 5.797217
iter  32 value 5.797064
iter  33 value 5.796980
iter  34 value 5.796970
iter  35 value 5.796786
iter  36 value 5.796716
iter  37 value 5.796583
iter  38 value 5.796507
iter  39 value 5.796501
iter  40 value 5.796305
iter  41 value 5.796248
iter  42 value 5.796122
iter  43 value 5.796056
iter  44 value 5.796051
iter  45 value 5.795891
iter  46 value 5.795837
iter  47 value 5.795834
iter  48 value 5.795713
iter  49 value 5.795656
iter  50 value 5.795652
iter  51 value 5.795516
iter  52 value 5.795466
iter  53 value 5.795464
iter  54 value 5.795357
iter  55 value 5.795306
iter  56 value 5.795302
iter  57 value 5.795198
iter  58 value 5.795150
iter  59 value 5.795147
iter  60 value 5.795049
iter  61 value 5.795004
iter  62 value 5.795000
iter  63 value 5.794912
iter  64 value 5.794868
iter  65 value 5.794865
iter  66 value 5.794788
iter  67 value 5.794746
```

```
iter  68 value 5.794743
iter  69 value 5.794677
iter  70 value 5.794636
iter  71 value 5.794633
iter  72 value 5.794530
iter  73 value 5.794499
iter  74 value 5.794497
iter  75 value 5.794406
iter  76 value 5.794376
iter  77 value 5.794374
iter  78 value 5.794301
iter  79 value 5.794271
iter  80 value 5.794269
iter  81 value 5.794212
iter  82 value 5.794182
iter  83 value 5.794180
iter  84 value 5.794135
iter  85 value 5.794106
iter  86 value 5.794103
iter  87 value 5.794033
iter  88 value 5.794010
iter  89 value 5.794008
iter  90 value 5.793954
iter  91 value 5.793931
iter  92 value 5.793929
iter  92 value 5.793929
iter  93 value 5.793880
iter  94 value 5.793863
iter  95 value 5.793862
iter  95 value 5.793862
iter  96 value 5.793819
iter  97 value 5.793799
iter  98 value 5.793797
iter  98 value 5.793797
iter  99 value 5.793759
iter 100 value 5.793742
final  value 5.793742
stopped after 100 iterations
initial  value 5.970639
iter   2 value 5.970601
iter   3 value 5.970586
iter   4 value 5.970583
iter   5 value 5.970495
iter   6 value 5.967106
iter   7 value 5.967022
iter   8 value 5.967005
```

```
iter   9 value 5.966025
iter  10 value 5.965722
iter  11 value 5.965535
iter  12 value 5.965485
iter  13 value 5.965401
iter  14 value 5.965222
iter  15 value 5.965108
iter  16 value 5.965080
iter  17 value 5.965073
iter  18 value 5.965073
iter  19 value 5.965071
iter  20 value 5.965065
iter  20 value 5.965065
iter  20 value 5.965065
final  value 5.965065
converged
```

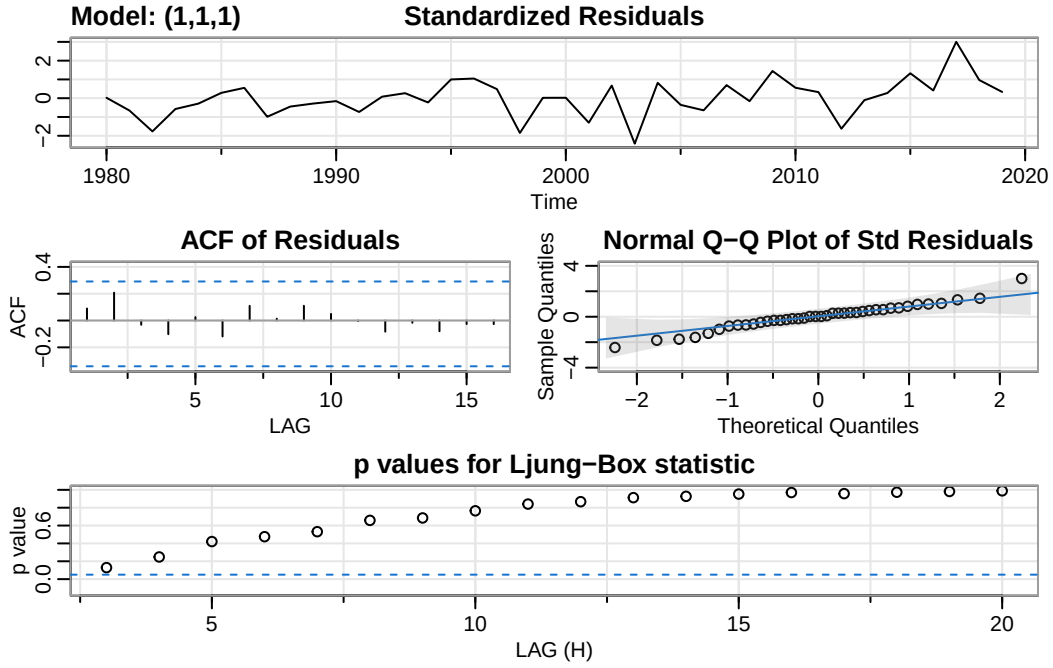


Figura 31: Diagnósticos del modelo ARIMA(1,1,1)

Se observa en los diagnósticos que los tres modelos propuestos logran aprobar contundentemente el supuesto de independencia de los residuos mostrado en los gráficos de ACF que ningún rezago tiene significancia. Aún más evidencia para este supuesto es proporcionado por el valor p de la pruebga de Ljung-Box donde se observa un no rechazo de la independencia en los tres modelos. Al observa el qqplot de los residuos en los modelos MA Y AR se observa que la cola izquierda podría ser un problema

Tabla 8: Resumen de diagnósticos de los modelos propuestos

Modelo	Estadístico W	Valor p Shapiro-Wilks	Estadístico Q	Valor p Ljung-Box	AIC	BIC	AICc
ARIMA(0,1,1)	0.96	0.11	0.01	0.93	-4.45	-4.32	-4.44
ARIMA(1,1,0)	0.95	0.11	0.00	0.98	-4.45	-4.32	-4.44
ARIMA(1,1,1)	0.97	0.29	0.35	0.56	14.97	15.14	14.99

Tabla 9: Intervalos de predicción del modelo ARIMA(1,1,0)

Año	predicción	Inf80	Sup80	Inf95	Sup95
2019	24383.78	23638.63	25152.42	23253.43	25569.08
2020	24991.33	23975.85	26049.81	23455.10	26628.17

para la hipótesis de normal, pues estos están al borde de desviarse significativamente de los cuantiles teóricos de la normal.

La Tabla 8 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 los tres modelos, lo que confirma que los tres pueden ser viables. Se observa que los modelos ARIMA(1,1,0) y ARIMA(0,1,1) obtuvieron los mejores resultados en cuestión de medidas de bondad de ajuste por lo que se escoge el modelo autoregresivo para hacer los pronósticos.

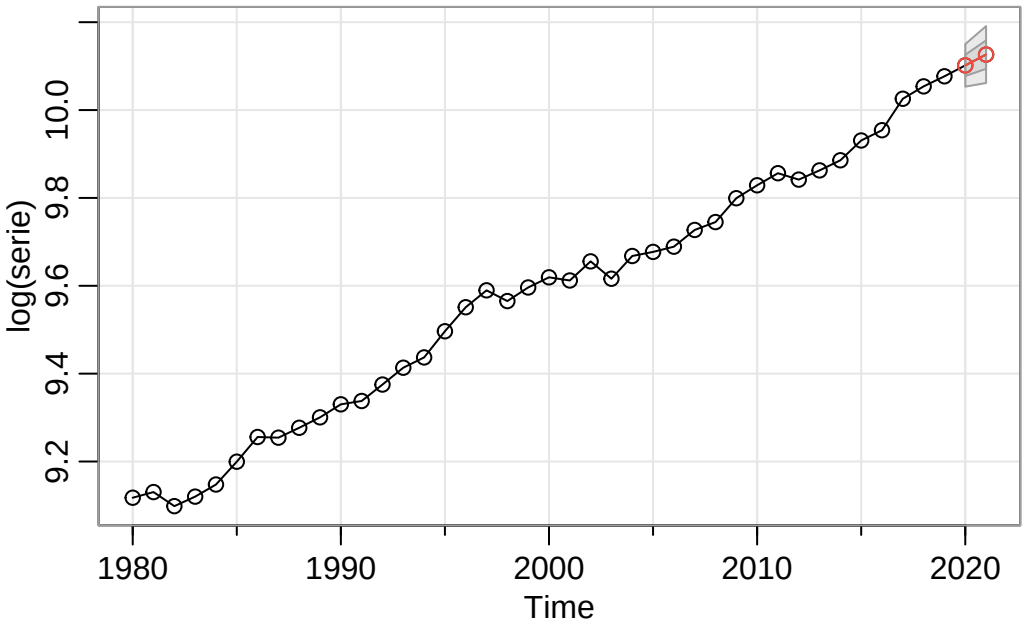


Figura 32: Pronóstico del modelo ARIMA(1,1,0) de la tasa de defunciones

La Tabla 9 muestra el valor pronosticado para las defunciones totales en los años 2019 y 2020 y los intervalos de confianza al 80% y 95%. Al comparar con los valores reales de 24292 para el 2019 y 26209 para el 2020

```
[1]   91.77885 1217.67310
```

Se observa que el valor real para el 2019 es consistente con los intervalos de predicción a ambos niveles de confianza. Para el 2020, se obtiene que el valor real se sale del intervalo al 80%, pero aún así es consistente con el de 95% de confianza. Esto es esperable pues este fue el primer año de pandemia que ocasionó más muertes de que las que se hubiesen dado en caso contrario.

DLM polinomial de orden 1

Este modelo está 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. 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).

```
     [,1]
[1,] 0.38
```

```
         [,1]
[1,] 296550.5
```

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 pronósitcos 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:

$$\begin{aligned} y_{1:t-1} &\sim \mathcal{N}\left(m_{t-1}, R_t = C_{t-1} + W\right) \\ Y_t \mid y_{1:t-1} &\sim \mathcal{N}\left(f_t = m_{t-1}, Q_t = R_t + V\right) \\ \mu_t \mid y_{1:t} &\sim \mathcal{N}\left(m_t = m_{t-1} + K_t e_t, C_t = K_t V\right) \end{aligned}$$

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. En un caso extremo en que $V = 0$, entonces $m_t = y_t$, es decir, el pronóstico a un paso es exactamente la observación más reciente. Es precisamente este fenómeno el que podría estar desembocando en el poco provecho del modelo en términos de sus pronósticos.

```
Time Series:
Start = 2020
End = 2021
Frequency = 1
     Series 1
[1,]    23786
[2,]    23786

$a
Time Series:
Start = 2020
End = 2021
Frequency = 1
     Series 1
[1,]    23786
[2,]    23786

$R
$R[[1]]
          [,1]
[1,] 296550.8

$R[[2]]
          [,1]
[1,] 593101.3


$f
Time Series:
Start = 2020
End = 2021
Frequency = 1
     Series 1
[1,]    23786
[2,]    23786

$Q
```

```
$Q[[1]]
         [,1]
[1,] 296551.2

$Q[[2]]
         [,1]
[1,] 593101.7
```

En cuanto a los diagnósticos del modelo, en la Figura 33 se observa del gráfico de residuos que hay algunos de estos muy cercanos a 2 en valor absoluto, aunque en general no se aprecia con claridad un valor extremo. Por su parte, en el ACF, la proporción de excepciones a la regla empírica es adecuada, siendo una proporción menor al 5% del total de rezagos usados (20), de modo que estos parecen no estar correlacionados. Con la prueba de Ljung-Box en la Figura 34 se ve aún más sustentado que los residuos no estén correlacionados. Sin embargo, hay graves problemas con el supuesto de normalidad. En primera instancia, del histograma presente en la Figura 33, el ajuste normal parece no ser bueno. Esto se constata con el gráfico cuantil-cuantil en la Figura 35, donde se nota un muy mal ajuste en ambas colas. A pesar de que con pruebas como la de Shapiro-Wilk no se rechazaría la hipótesis de normalidad con niveles de significancia usuales del 1% y 5%, teniéndose un valor p de cerca del 8%, del gráfico cuantil-cuantil ya discutido no resulta fundamentada la normalidad de los residuos.

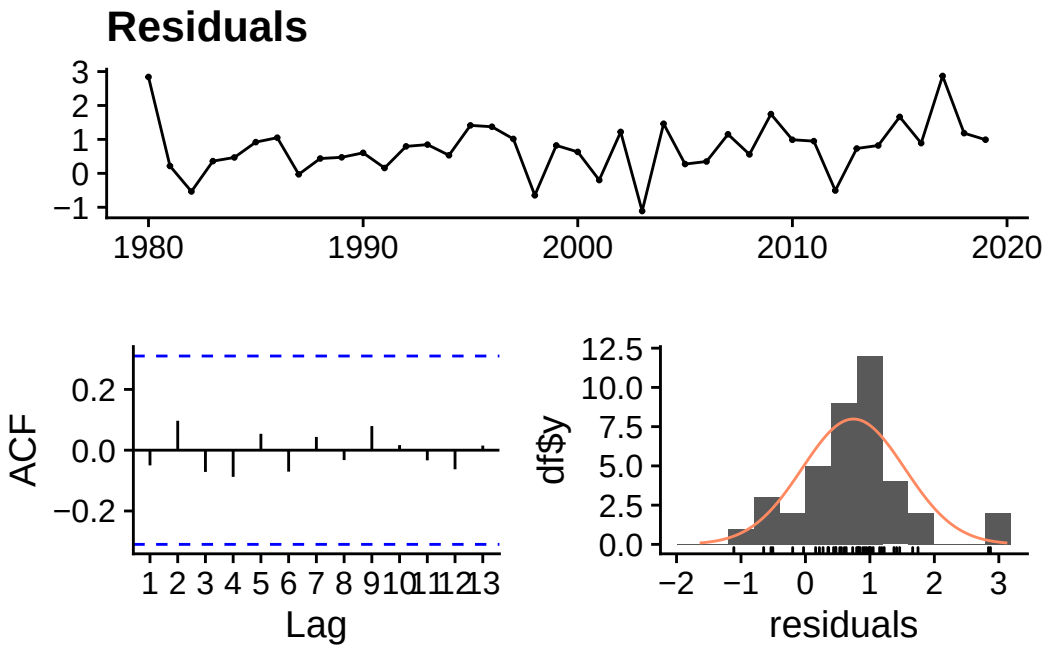


Figura 33: Algunos diagnósticos descriptivos de los residuos para el modelo DLM polinomial de primer orden

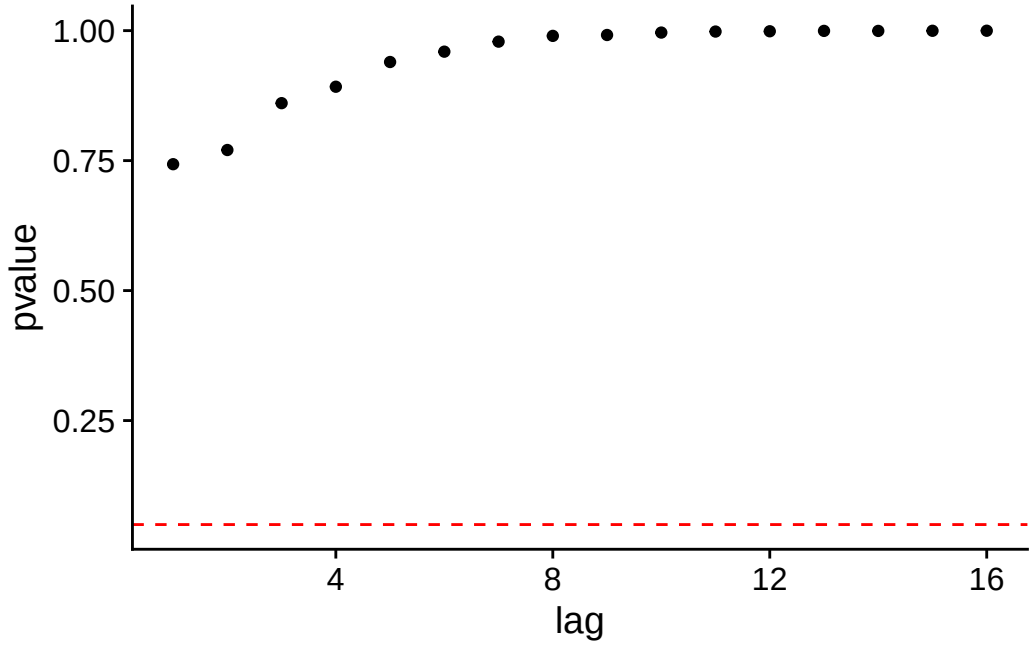


Figura 34: Valores p del estadistico Ljung-Box para el modelo LDM polinomial de orden 1

```
Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```

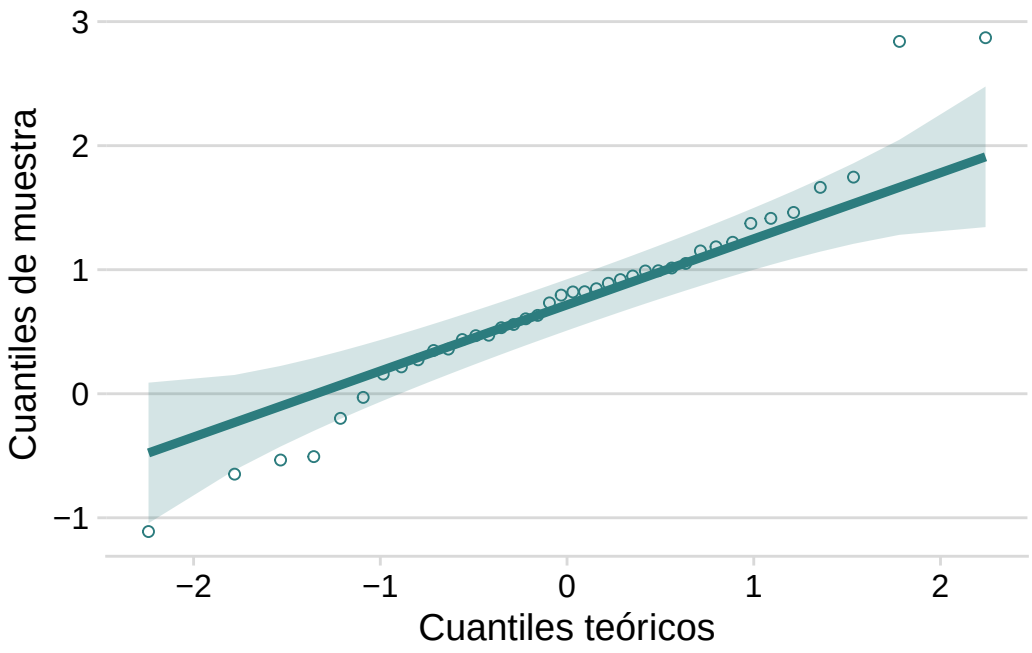


Figura 35: Gráfico cuantil-cuantil de los residuos para el modelo DLM polinomial de primer orden

DLM polinomial de orden 2

```
NULL
```

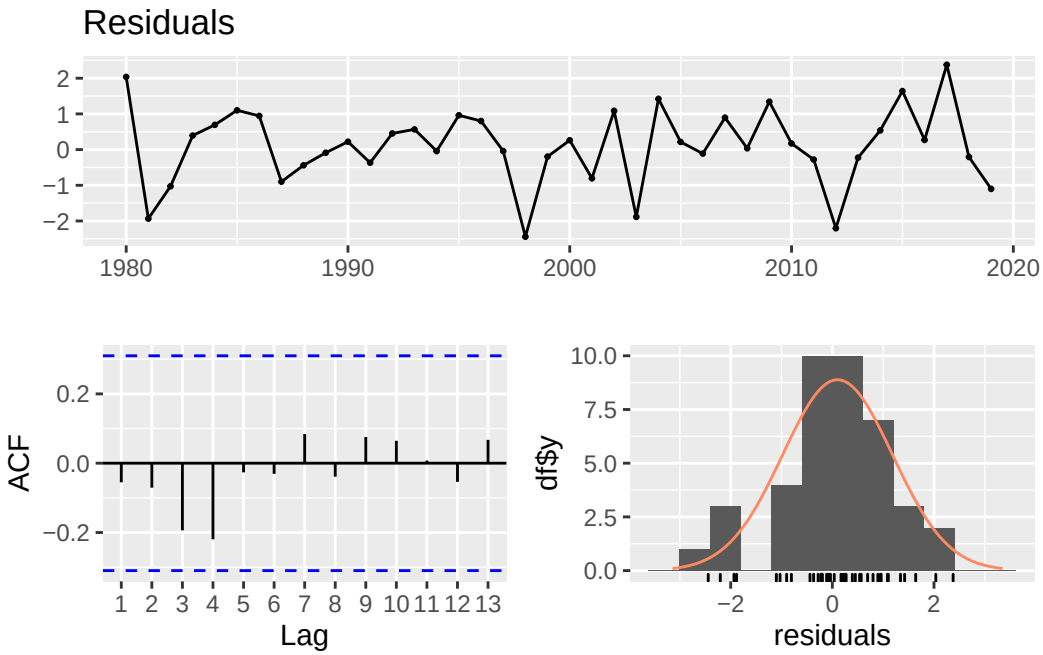


Figura 36: Diagnóstico de residuos para el modelo DLM polinomial de segundo orden

De los diagnósticos de los residuos Figura 36 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.

```
Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```

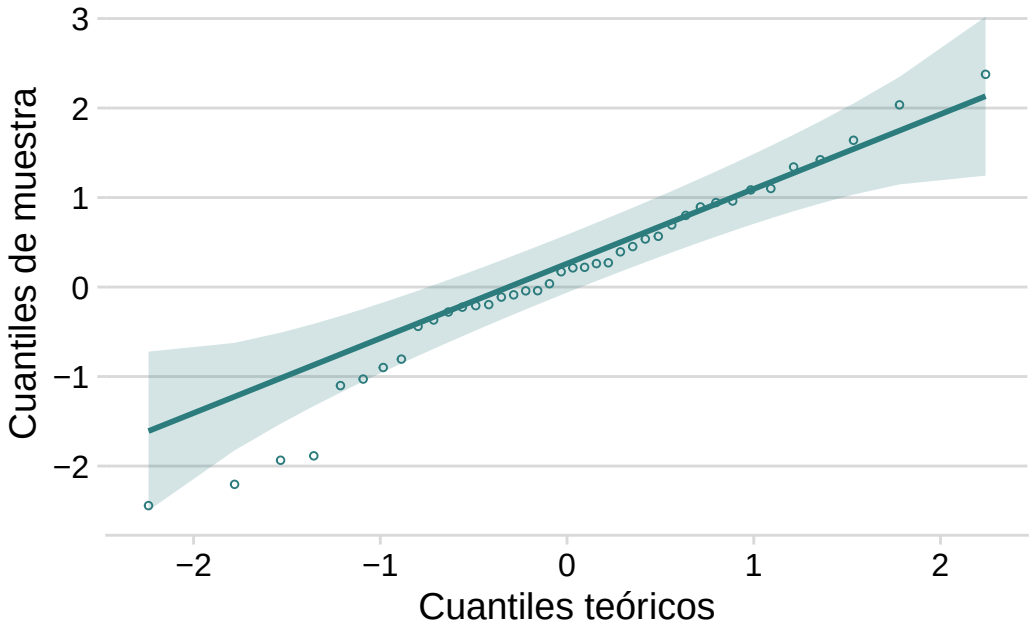


Figura 37: Gráfico cuantil-cuantil de los residuos para el modelo DLM polinomial de segundo orden

Tabla 10: Resumen de diagnósticos de los modelos DLM polinomiales

Modelo	Estadístico W	Valor p Shapiro-Wilks	Estadístico Q	Valor p Ljung-Box	AIC	BIC	AICc
DLM polinomial orden 2	0.97	0.42	5.54	0.94	-21.1	-16.03	-20.43

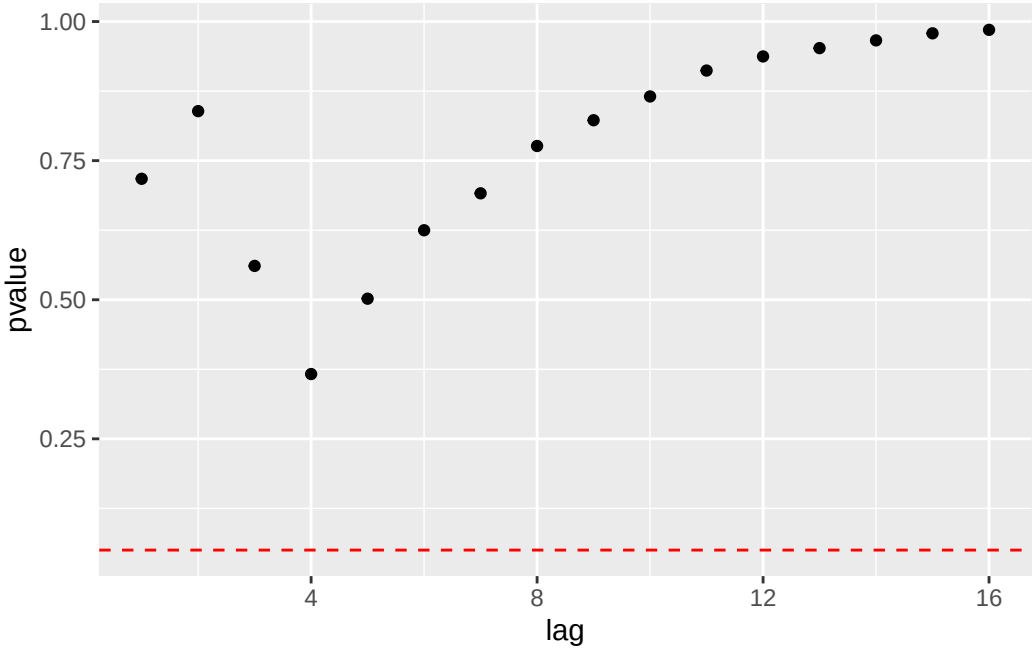


Figura 38: Valores p del estadistico Ljung-Box

El grafico Figura 38 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.

Análogamente a los modelos ARIMA , la Tabla 10 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 9 indicando que el modelo DLM polinomial de orden 2 mediante estos criterios es mejor modelo que los ARIMA.

En Tabla 11 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 442 y 644 defunciones aproximadamente.

```
[1] 409.9398 719.1865
```

Tabla 11: 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

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%, 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 de los modelos

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.

UVE Final

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.

Brownlee, J. (2020). *Introduction to time series forcasting with Python* . eBook.

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*.

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.

Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.".

Wilson, T., & Rees, P. (2021). A brief guide to producing a national population projection. *Australian Population Studies*, *5*(1), 77-100.