Contents

Introducción **2**

Análisis de la serie **3**

Modelos **6**

Verificación de supuestos **6**
Normalidad 6
Varianza constante. 8
Media cero. 9
Independencia. 10

Holt-Winters **12**

Serie 2011-2019 **16**

Verificación de Supuestos **17**

Holt-Winters 2011-2019 **19**

Comparación de predicción **20**

Back Testing **21**

Referencias **22**

Introducción

La base de datos fue generada a partir del acopio y procesamiento de los datos alusivos a los accidentes ocurren de manera nacional. Esta información contribuyó a la planeación, organización del transporte y la prevención de accidentes.

En 1997 inició la etapa de descentralización de actividades con el propósito de que el levantamiento de los datos sea más eficiente con ello eliminar rezagos en el suministro de la información, coadyuvando a la generación y difusión de la Estadística ATUS en forma oportuna.

En la etapa de descentralización, el ámbito regional del INEGI desempeñó el papel principal al asumir el desarrollo de las actividades referentes al levantamiento, procesamiento de la información e integración de bases de datos.

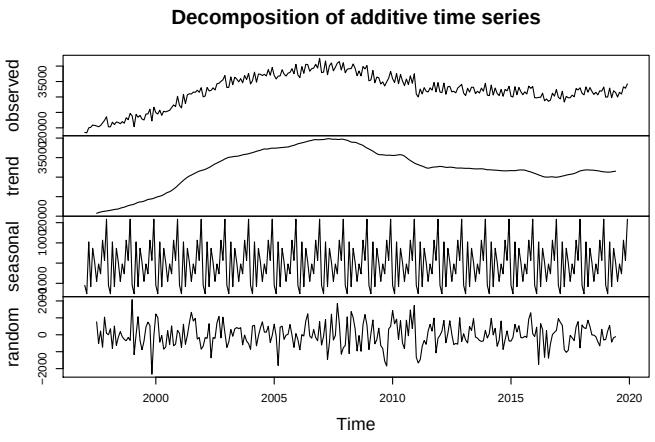
El objetivo de la base de datos es producir información anual sobre la siniestralidad del transporte terrestre a nivel nacional, entidad federativa y municipio, mediante el acopio y procesamiento de datos alusivos a los accidentes ocurridos en zonas federales y no federales, contribuyendo con ello a la planeación y organización del transporte.

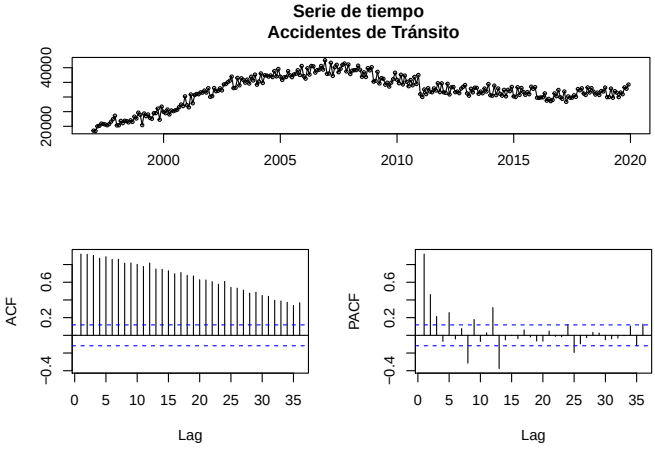
Para fines del proyecto, nos enfocaremos en la serie de tiempo sobre el total de accidentes (Fatales, no fatales, sin daños) que hubo desde 1997 que fue el primer año en almacenar los datos hasta el 2019, ya que el 2020 puede presentar fallas o sesgos, debido a que es el último año de registro.

Análisis de la serie

```
##        Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
## 1997 18515 18407 20004 20175 20915 20773 20609 20355 20754 21579 22435 23593
## 1998 20309 20350 21847 21062 21896 21746 21325 22036 21498 23285 22668 24665
## 1999 23898 20392 24369 23442 24030 22857 22554 24471 24490 26042 22245 26704
## 2000 24939 24484 25726 24053 25389 25166 25431 25658 26558 27421 26894 30219
## 2001 27241 26460 30856 27863 30690 31035 30948 31571 31492 32194 31515 33004
## 2002 30048 30306 33064 31953 32126 32957 32262 34292 34432 35058 35545 36959
## 2003 33045 33139 36509 33978 36388 35813 34882 35937 34584 36927 35647 37641
## 2004 34183 35300 38045 34789 37420 37331 36945 37337 36806 38854 37036 39561
## 2005 36943 35823 36720 36830 38824 37411 36770 37705 37945 39148 37530 40584
## 2006 37143 36260 39837 37662 40582 40468 38334 39168 39337 40591 39359 42531
## 2007 37901 38026 41678 37198 40127 41006 38469 39647 40977 41431 38726 41093
## 2008 37830 38986 39155 39186 40683 39302 36902 38421 36835 39855 39089 40191
## 2009 35235 35498 38587 34625 36433 36213 34187 34876 33558 34882 36091 38282
## 2010 35880 34700 37599 34279 37308 35196 33677 35821 34006 36847 34436 37518
## 2011 30960 30001 32571 30980 33014 31860 31612 32954 32205 34713 31668 34647
## 2012 31612 31403 34514 30869 33325 33422 31650 31846 31310 32891 33480 34089
## 2013 31245 30440 32819 31580 33178 33122 31082 31855 31273 32882 31864 34432
## 2014 30592 30437 33881 30627 32908 31113 30472 32438 30228 32449 31977 33451
## 2015 30210 29997 33329 30731 33080 31711 30737 31662 30948 33513 32777 33371
## 2016 29758 29714 29866 29883 31235 28755 29356 28580 29008 31231 30260 32405
## 2017 29894 29362 31871 28334 30209 29740 29765 30593 30053 32720 32280 32968
## 2018 31188 30690 33270 31543 33028 31825 30833 31522 30688 32017 31879 33225
## 2019 29936 29989 32886 29749 32003 31452 29981 31526 30940 33283 32830 34239
```

Para anlizar nuestra serie de tiempo, podemos descomponerla.





Nuestra Serie presenta una tendencia creciente ede 1997 a 2008 o 2009 mientras que entre 2010 y 2011 hay un cambio drástico, después se estaviliza la serie, también notamos que los ciclos con anuales, aunque no se alcanzan a diferenciar bien.

Antes de trabajar con la Serie de tiempo, vamos tenermos que verificar los supuesto de homocedasticidad y estacionariedad.

Cuyos supuestos se basaron mediante el test de Breusch-Pagan para homocedasticidad, Dickey-Fuller y Kwiatkowski-Phillips-Schmidt-Shin para estacionariedad.

Recordando el contraste de hipótesis para Homocedasticidad con bptest.

H_0 : La varianza es constante (Homocedastisidad) vs H_aLa varianza no es constante (Heterocedasticidad)

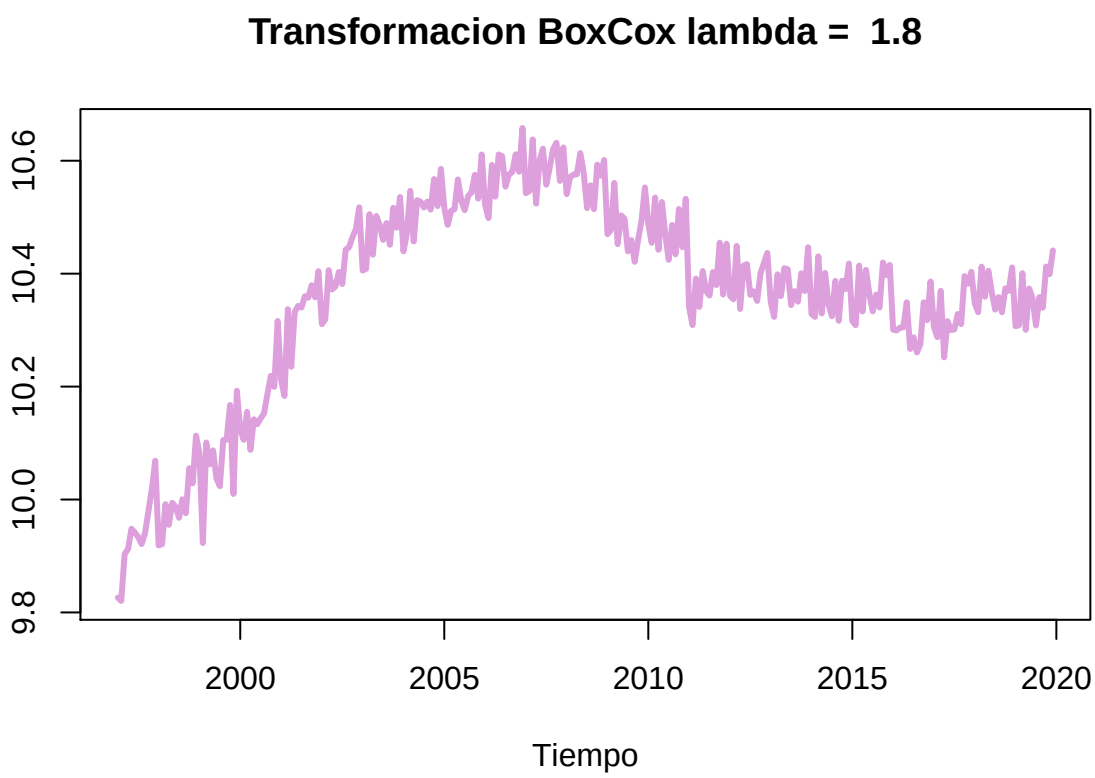
Contraste con adf.test para estacionariedad.

$$H_0 : \text{La serie no es estacionaria } vs \ H_a : \text{La serie es estacionaria}$$

Contraste con kpss.test para estacionariedad.

$$H_0 : \text{La serie es estacionaria } vs \ H_a : \text{La serie no es estacionaria}$$

Mediante estos contrastes de hipótesis se hizo una transformación Box-Cox con el método *loglik*.



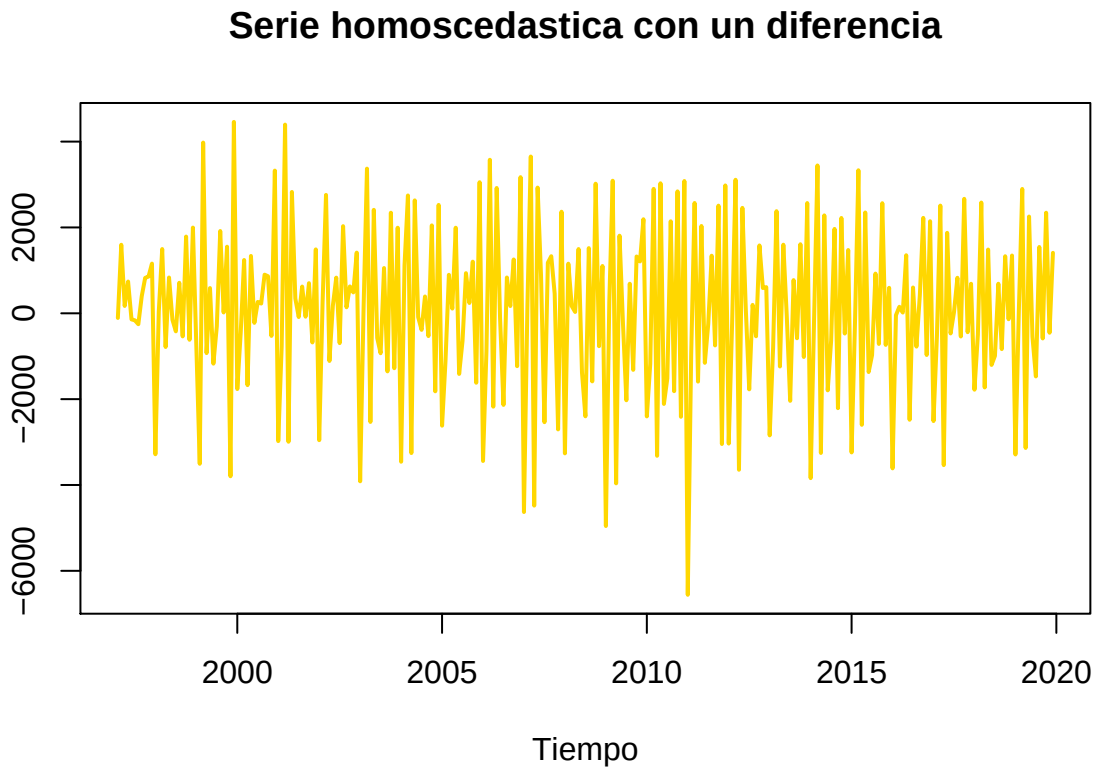

```
## [1] TRUE
```

```
##
##  studentized Breusch-Pagan test
##
## data:  datos.ts1 ~ T1
## BP = 77.338, df = 1, p-value < 2.2e-16
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  datos.ts1
## Dickey-Fuller = -2.5922, Lag order = 6, p-value = 0.3265
## alternative hypothesis: stationary
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  datos.ts1
## KPSS Level = 1.415, Truncation lag parameter = 5, p-value = 0.01
```

Y notamos que nuestra serie no cumple homocedasticidad ni estacionariedad. Por esta razón aplicamos una diferencia y notamos los resultados.




```
## [1] TRUE
```

```
##
##  studentized Breusch-Pagan test
##
## data:  diff(datos_st) ~ T1
## BP = 0.35721, df = 1, p-value = 0.5501
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff(datos_st)
## Dickey-Fuller = -6.0605, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  diff(datos_st)
## KPSS Level = 0.31696, Truncation lag parameter = 5, p-value = 0.1
```

Con una diferencia podemos notar que nuestra serie cumple el supuesto de homocedasticidad y estacionariedad.

El método usado en este estudio, será Box-Jenkins, este método consta de tres etapas:

- **Primera etapa:** Identificación de los parámetros d,p,q y D,P,Q
- **Segunda etapa:** Estimación de los coeficientes.
- **Tercera etapa:** Verificación de los supuestos.

Después de la verificación de supuestos hacemos la predicción. En éste proyecto, nuestra predicción será de dos años.

Modelos

Propusimos 4 modelos con Box-Jenkings.

El primero modelo que se propuso fue un ARMA(11,23), esto se hizo guiandonos por los gráficos de autocorrelación ACF y PACF. Otro modelo que se propuso, fue una SARIMA(0,1,1)(0,1,1)[12], esto fue mediante la función autoarima del paquete *forecast*. Mientras que los otros dos modelos, se construyeron basandonos en el autoarima y en los supuestos, donde el tercer modelo es un SARIMA(0,0,1)(0,2,2)[12] y SARIMA(6,0,3)(0,2,3)[12]

Verificación de supuestos

Normalidad

Para todos los ajustes se verificó normalidad mediante Anderson-Darl.ing y Jarque-Bera

Modelo 1

```
##
##  Anderson-Darling normality test
##
## data:  primer_ajuste$residuals
## A = 0.76791, p-value = 0.04544
```

```
##
##  Jarque Bera Test
##
## data:  primer_ajuste$residuals
## X-squared = 15.224, df = 2, p-value = 0.0004945
```

Notamos que aunque AD nos dice que por poco la cumple, además de que este es un modelo excesivamente grande, por lo tanto no cumple normalidad.

Modelo 2

```
##
##  Anderson-Darling normality test
##
## data:  segundo_ajuste$residuals
## A = 1.2689, p-value = 0.002631
```

```
##
##  Jarque Bera Test
##
## data:  segundo_ajuste$residuals
## X-squared = 16.413, df = 2, p-value = 0.0002729
```

Para este modelo, tenemos las situación de que tampoco cumple normalidad, el p-value disminuyó, pero no fue tanto comparado a que los parámetros son pequeños.

Modelo 3

```
##
##  Anderson-Darling normality test
##
## data:  tercer_ajuste$residuals
## A = 1.489, p-value = 0.000755
```

```
##
##  Jarque Bera Test
##
## data:  tercer_ajuste$residuals
## X-squared = 5.7908, df = 2, p-value = 0.05528
```

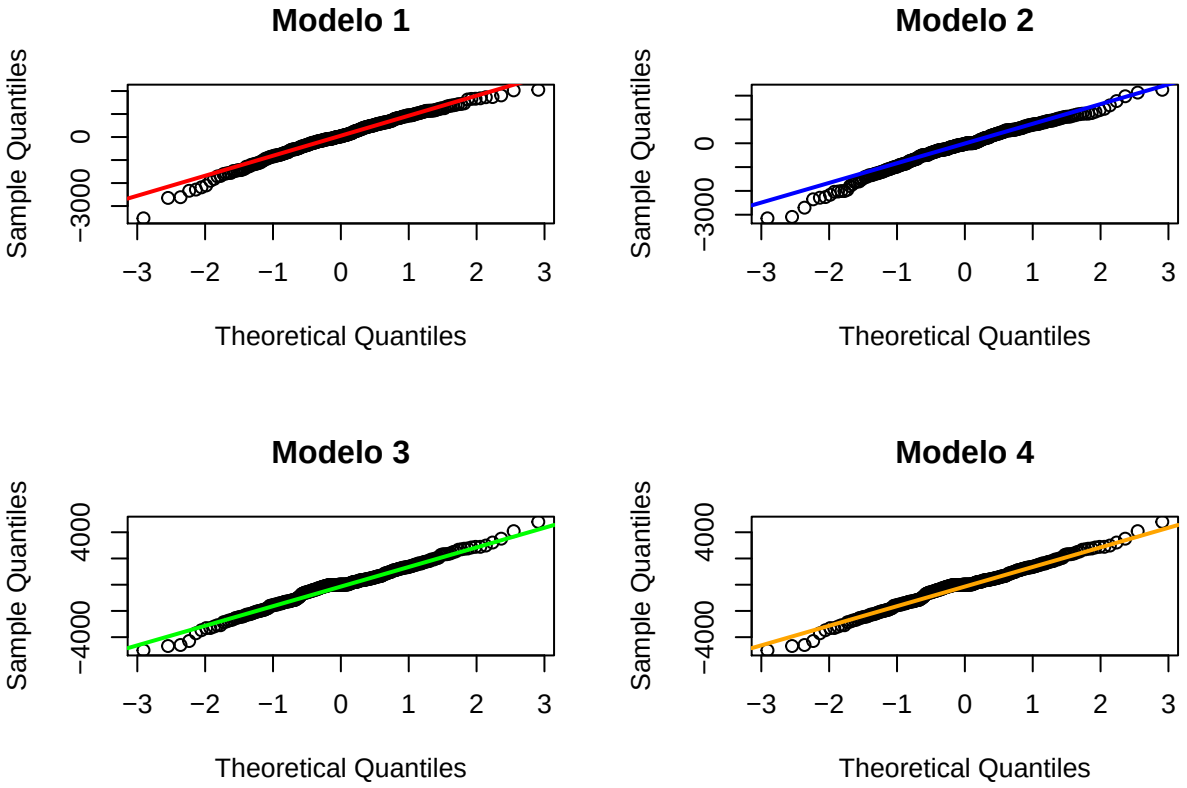
Este modelo diriamos que pasa normalidad, pues Jarque-Bera es un buen indicador de normalidad para las series de tiempo.

Modelo 4

```
##
##  Anderson-Darling normality test
##
## data:  cuarto_ajuste$residuals
## A = 1.3391, p-value = 0.001766
```

```
##
##  Jarque Bera Test
##
## data:  cuarto_ajuste$residuals
## X-squared = 6.2194, df = 2, p-value = 0.04461
```

Si no nos ponemos muy exigentes, diriamos que el modelo también pasa normalidad.



Varianza constante.

La varianza constante se verificó mediante bptest, cuyo contraste de hipótesis es:

H_0 : La varianza es constante (Homocedastisidad) *vs* H_a : La varianza no es constante (Heterocedasticidad)

Modelo 1

```
##
##  studentized Breusch-Pagan test
##
## data:  Y ~ X
## BP = 0.36581, df = 1, p-value = 0.5453
```

Modelo 2

```
##
##  studentized Breusch-Pagan test
##
## data:  Y ~ X
## BP = 0.35784, df = 1, p-value = 0.5497
```

Modelo 3

```
##
##  studentized Breusch-Pagan test
##
## data:  Y ~ X
## BP = 0.0026254, df = 1, p-value = 0.9591
```

Modelo 4

```
##
##  studentized Breusch-Pagan test
##
## data:  Y ~ X
## BP = 0.010297, df = 1, p-value = 0.9192
```

Aquí, todos nuestros modelos tienen varianza constante.

Media cero.

Usamos t.test para verificar si nuestro modelo tiene media cero, donde el contraste es el siguiente.

$$H_0 : \text{La media es igual a cero } vs \ H_a : \text{La media no es igual a cero}$$

Modelo 1

```
##
##  One Sample t-test
##
## data:  primer_ajuste$residuals
## t = 0.31641, df = 274, p-value = 0.7519
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -90.96527 125.80567
## sample estimates:
## mean of x
##   17.4202
```

Modelo 2

```
##
##  One Sample t-test
##
## data:  segundo_ajuste$residuals
## t = -1.6574, df = 275, p-value = 0.09858
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -195.43548   16.77559
## sample estimates:
## mean of x
## -89.32995
```

Modelo 3

```
##
##  One Sample t-test
##
## data:  tercer_ajuste$residuals
## t = -0.76054, df = 275, p-value = 0.4476
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -256.0280  113.3324
## sample estimates:
## mean of x
## -71.34784
```

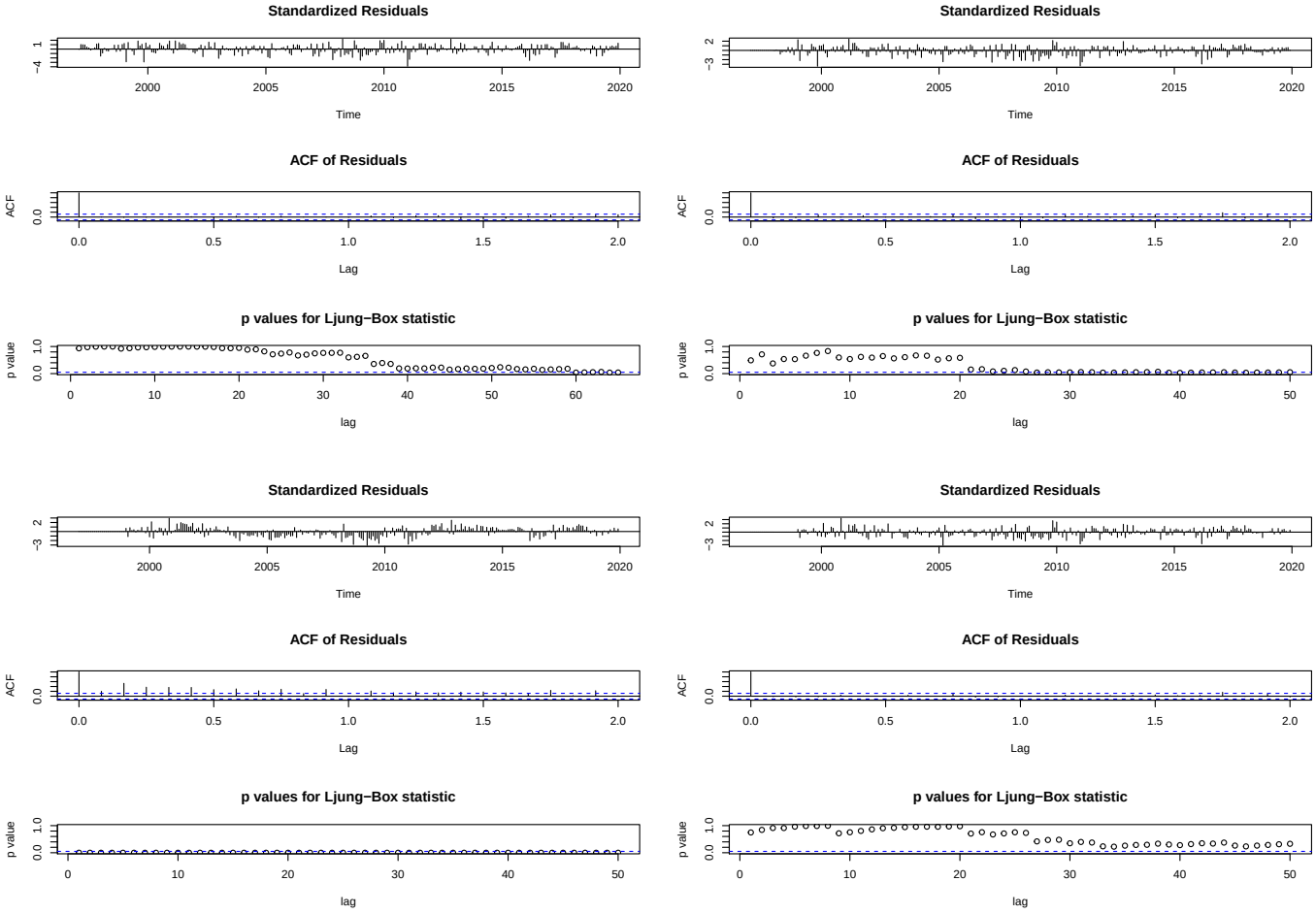
Modelo 4

```
##
##  One Sample t-test
##
## data:  cuarto_ajuste$residuals
## t = -0.46992, df = 275, p-value = 0.6388
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -129.53431   79.61028
## sample estimates:
## mean of x
## -24.96201
```

De nuevo, todos nuestros modelos tienen media cero.

Independencia.

Utilizaremos la prueba de Ljung-Box, que comprueba si una serie de observaciones en un período de tiempo específico son aleatorias e independientes. Este supuesto lo optendremos mediante gráficos.



Unicamente el último pasa el supuesto de independencia, que fue el SARIMA(6,0,3)(0,2,3)[12]

Hacemos un cuadro comparativo para comparar los supuestos de los modelos.

Modelo	Normalidad	Variable Cte	Media cero	Independencia
Primer	no	si	si	no
Segundo	no	si	si	no
Tercer	si	si	si	no
Cuarto	si	si	si	si

Por otro lado vamos a comparar los errores de los modelos y sus respectivo AIC y BIC

AJUSTE	AIC	BIC	ME	MAE	RMSE	Parámetros
ARMA(11,23)	4602.3613	4702.0141	17.4202	713.9011	911.496560293074	26
ARIMA(0,1,1)(0,1,1)	4354.8056	4367.522	-89.3299	688.6297	898.253716185352	2
ARIMA(0,0,1))(0,2,3)	4462.5866	4482.2337	-71.3478	1162.1287	1557.32516995708	4
ARIMA(6,0,3))(0,2,3)	4250.5725	4298.455	-24.962	656.7973	881.238883210699	12

El primer modelo tiene muchos parámetros a estimar y la mitad de ellos contienen al cero, por lo que no es una buena propuesta de modelo, el segundo ajuste solamente tiene 2 parámetros a estimar y ninguno contiene al cero, el tercer ajuste tiene 4 parámetros y solamente 1 de ellos contiene al cero, mientras que el cuarto ajuste tiene 12 parámetros a estimar y uno de ellos contiene al cero.

Basandonos en el modelos de AIC nos quedariamos con ARIMA(6,0,3))(0,2,3) además de que pasa todos los supuestos, pero son demaciados parámetros y la mitad de ellos contienen al cero, por lo que no es un buen ajuste, igual que el primer modelo, tiene demasiados parámetros.

Por otro lado, el tercer modelo, es buena opción, aunque estamos sobreajustando el modelo,la razón es la siguiente.

```
## [1] 3901214
```

```
## [1] 12277059
```

```
## [1] 3901214
```

```
## [1] 4262910
```

```
## [1] 1995181
```

```
## [1] 5675175
```

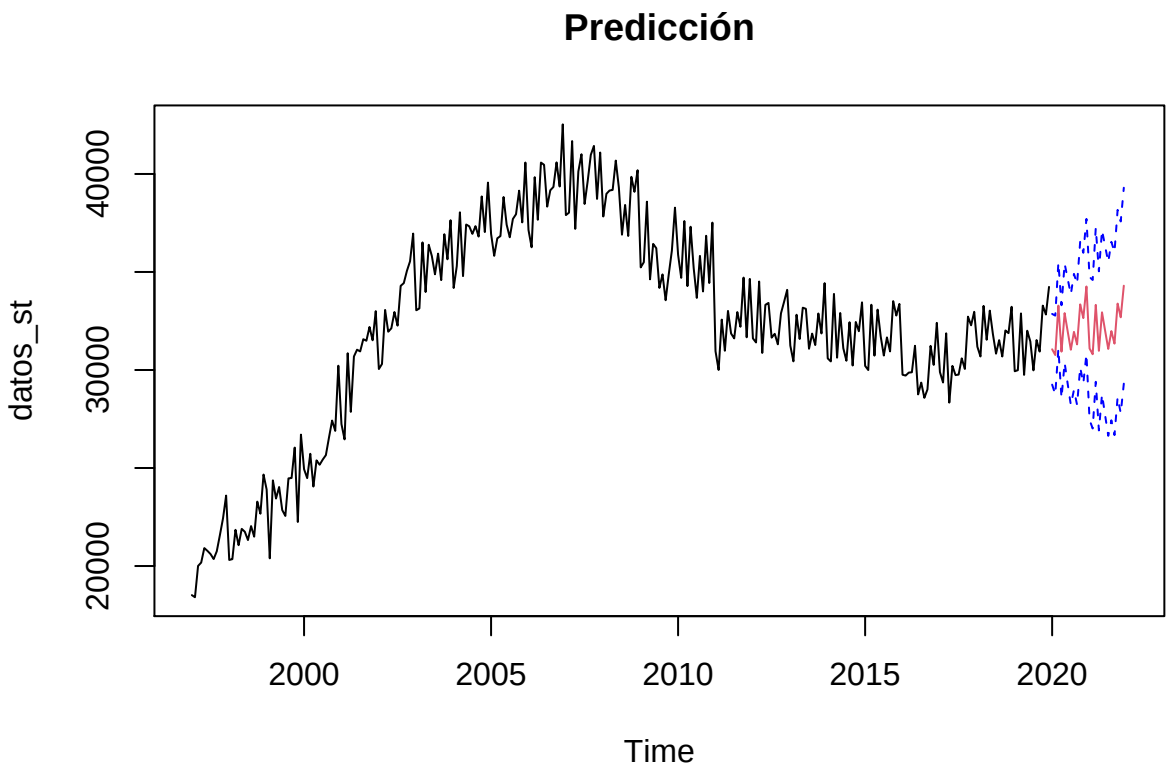
```
## [1] 6094403
```

```
## [1] 6094403
```

Por lo que nos quedariamos con el modelo

##Forecast

Como mencionamos antes, la predicción la haremos por 3 años.



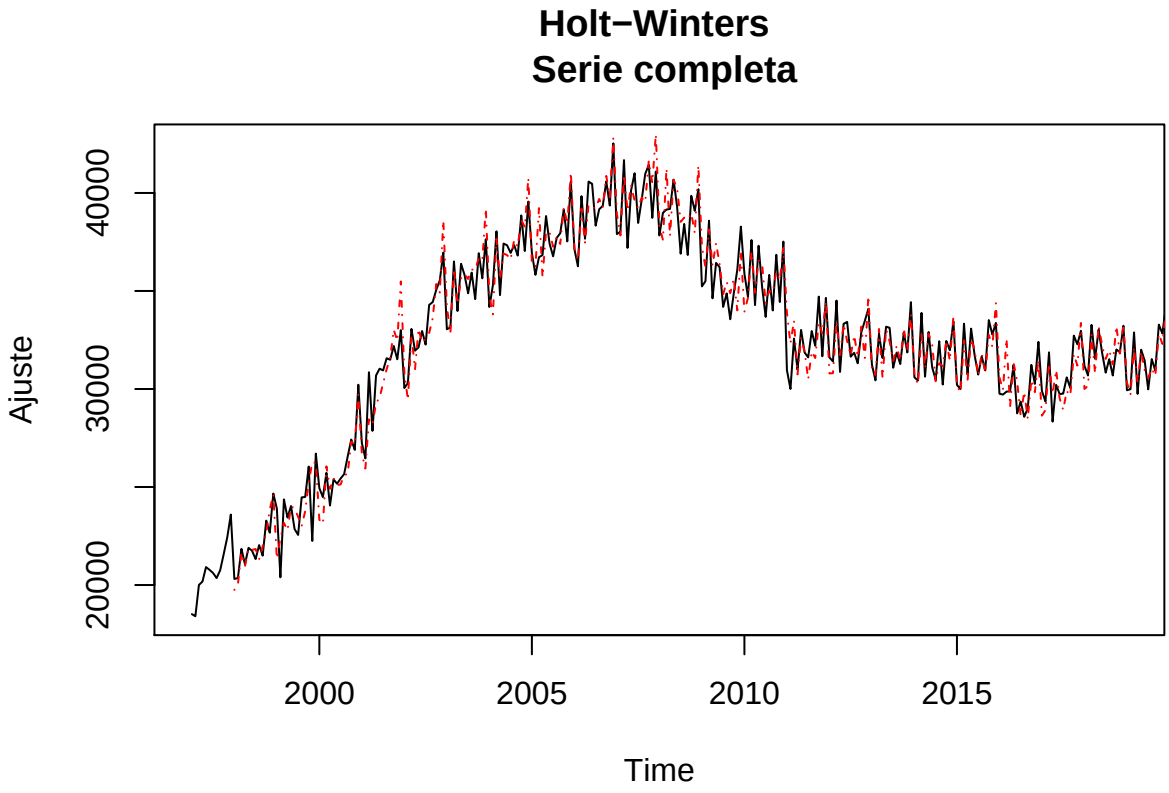

```
##           Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
## 2020 31056.83 30763.37 33281.59 30935.84 32901.47 31960.38 31034.09 31944.45
## 2021 31097.02 30803.56 33321.78 30976.03 32941.66 32000.57 31074.28 31984.64
##           Sep      Oct      Nov      Dec
## 2020 31305.11 33347.03 32647.15 34260.12
## 2021 31345.30 33387.22 32687.34 34300.31
```

Notamos que sus bandas de confianza son muy amplias, por lo que podemos probar con un suavizamiento exponencias y despues podemos comparar resultados.

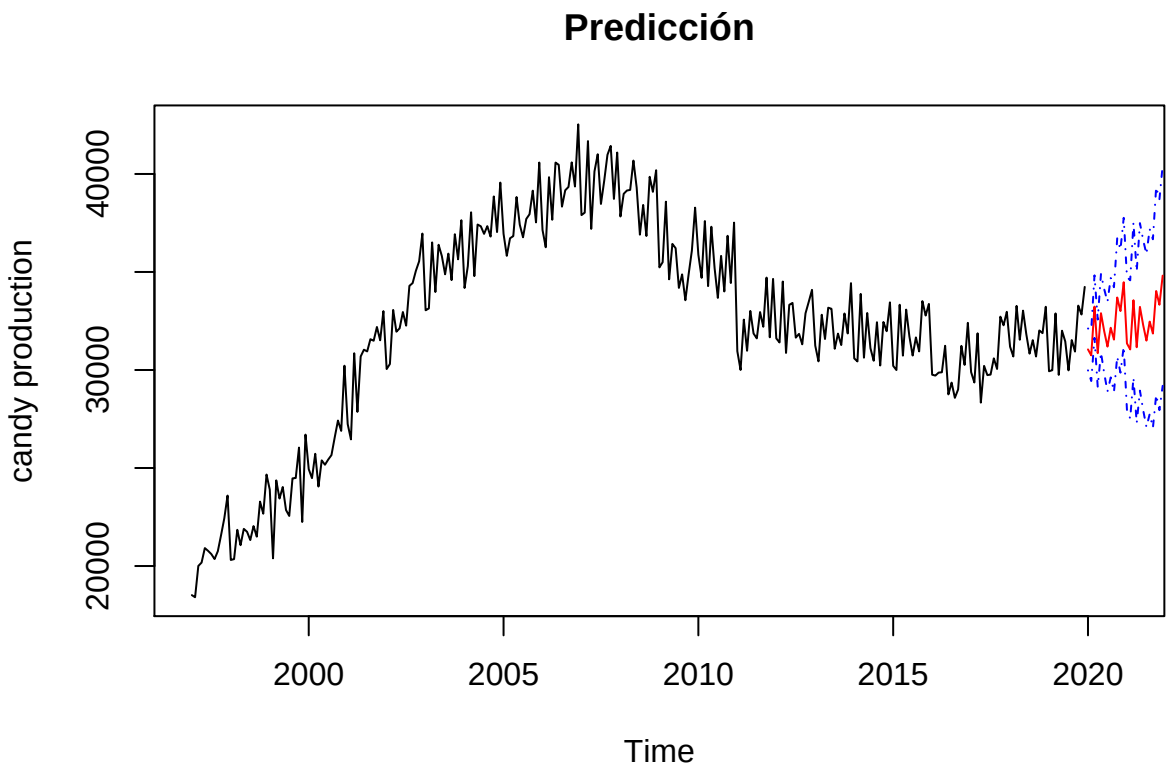
Holt-Winters

Holt-Winters es un método de pronóstico de triple exponente suavizante y tiene la ventaja de ser fácil de adaptarse a medida que nueva información real está disponible. Holt-Winters considera nivel, tendencia y estacionalidad de una serie de tiempo. Este método tienen dos modelos principales, dependiendo del tipo de estacionalidad; modelo multiplicativo estacional y aditivo estacional.

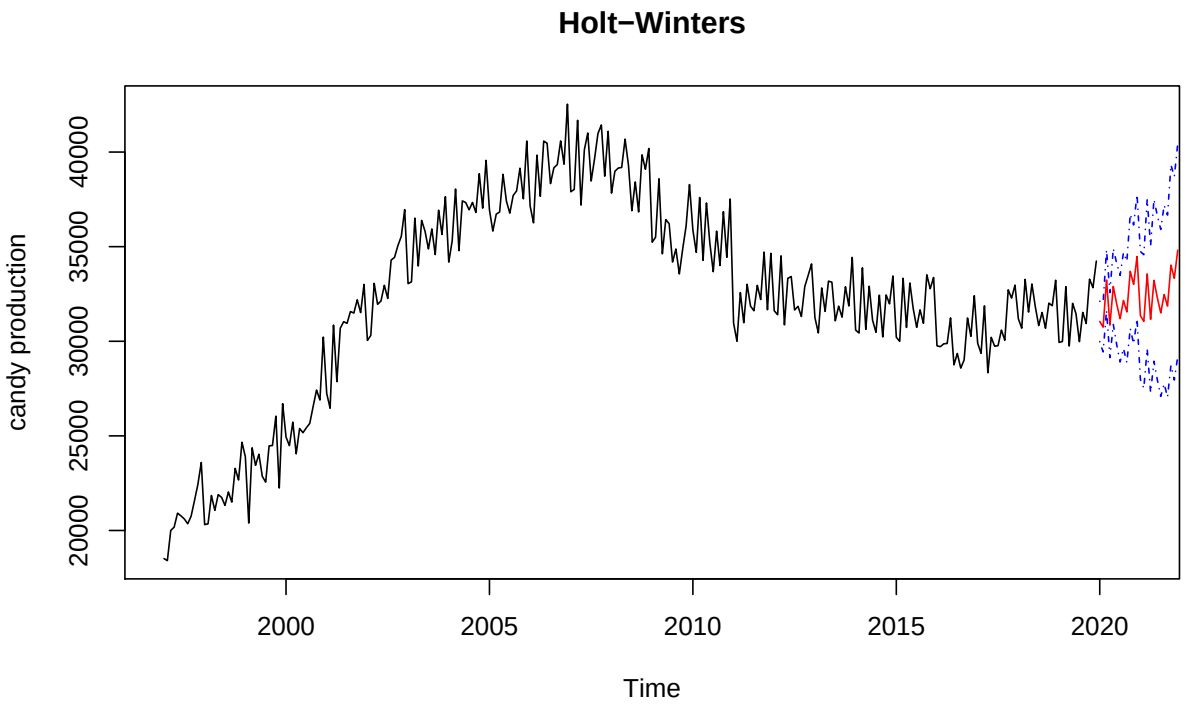
Al ver la serie completa, vemos que tiene cara de un modelo multiplicativo, salvo en los últimos 8 años.

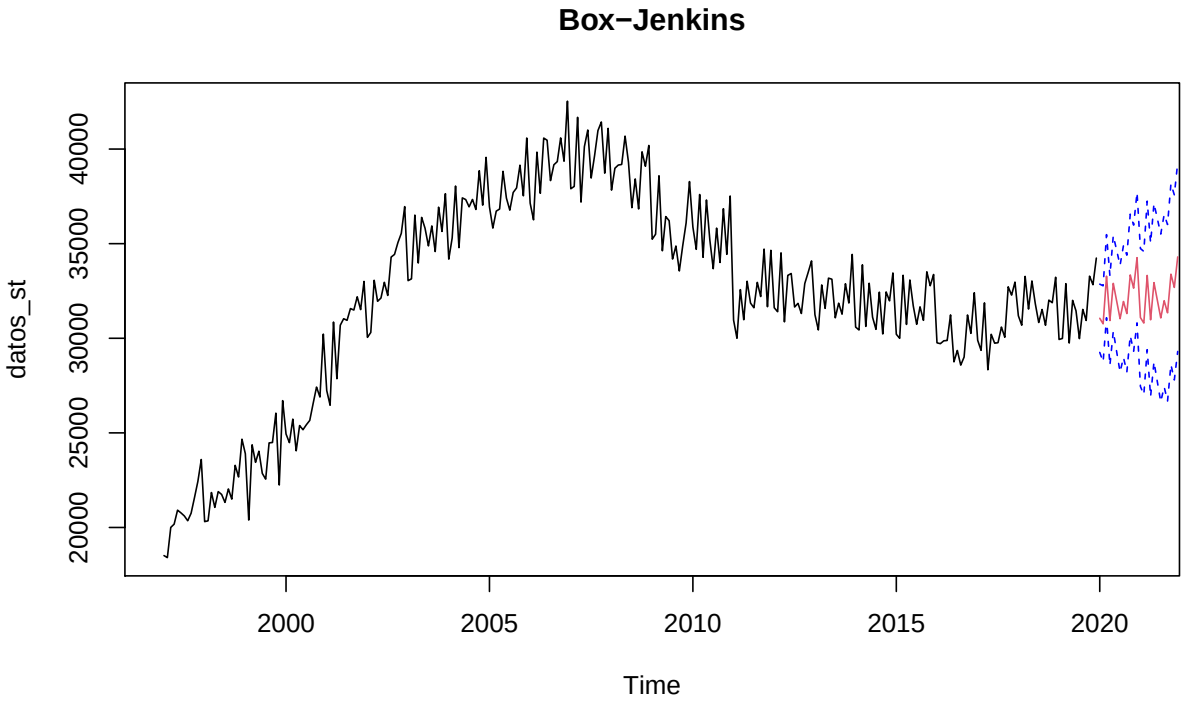


El ajuste es un poco bueno, veamos si las bandas de confianza mejoran con la predicción.



Podemos notar que las bandas de confiansa mejoran, entonces la mejor opción es el ajuste con Holt-Winters,



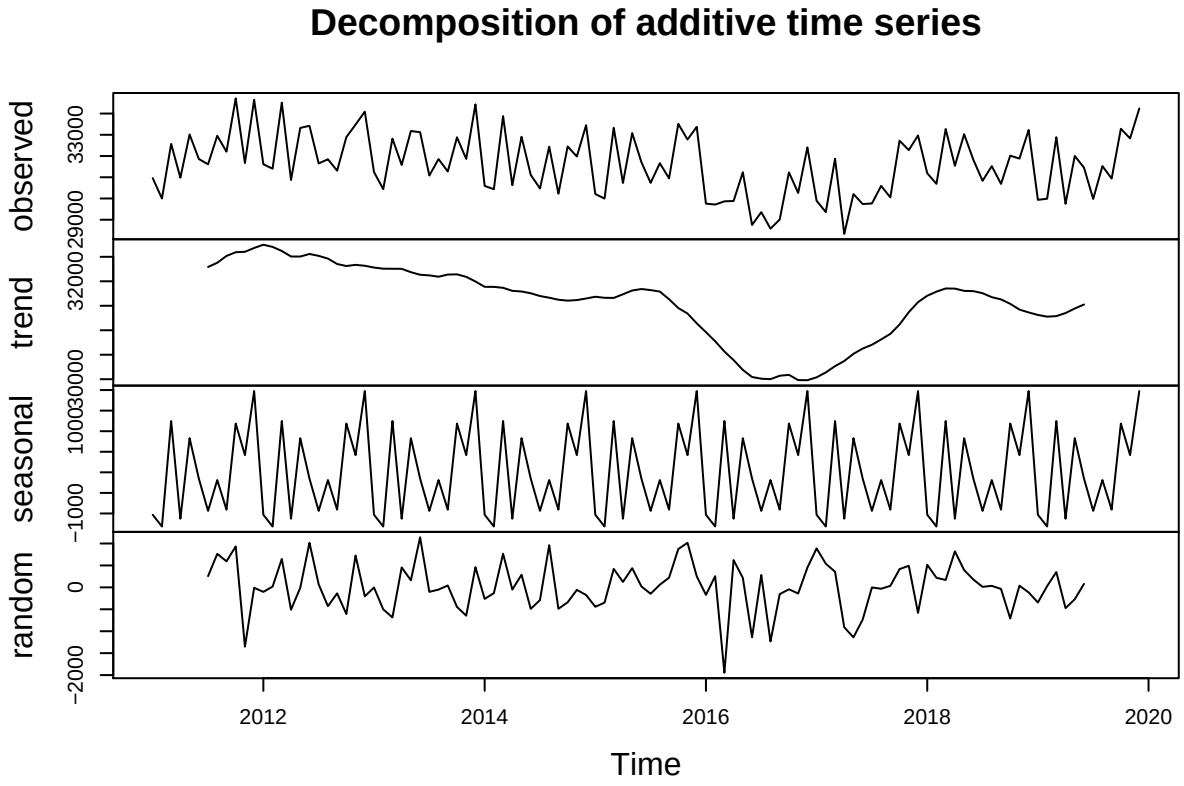


Notamos que nuestra Serie de tiempo tiene un cambio de tendencia a finales de 2010, pues de una tendencia creciente, pasamos a una tendencia decreciente, debido a cambiso demográficos o políticos.

Por ello, analizamos nuestra serie de tiempo a partir de el 2011, ya que consideramos que sigue un proceso constante.

Serie 2011-2019

```
##        Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
## 2011 30960 30001 32571 30980 33014 31860 31612 32954 32205 34713 31668 34647
## 2012 31612 31403 34514 30869 33325 33422 31650 31846 31310 32891 33480 34089
## 2013 31245 30440 32819 31580 33178 33122 31082 31855 31273 32882 31864 34432
## 2014 30592 30437 33881 30627 32908 31113 30472 32438 30228 32449 31977 33451
## 2015 30210 29997 33329 30731 33080 31711 30737 31662 30948 33513 32777 33371
## 2016 29758 29714 29866 29883 31235 28755 29356 28580 29008 31231 30260 32405
## 2017 29894 29362 31871 28334 30209 29740 29765 30593 30053 32720 32280 32968
## 2018 31188 30690 33270 31543 33028 31825 30833 31522 30688 32017 31879 33225
## 2019 29936 29989 32886 29749 32003 31452 29981 31526 30940 33283 32830 34239
```



La serie ya modificada tiene una tendencia un poco más constante, los ciclos son anuales, aunque no se notan mucho.

Con Box-Jenkins elegeimos el modelo que nos arrojó autoarima, entonces aqui vamos implementar la función autoarima para después comparar con Holt Winters.

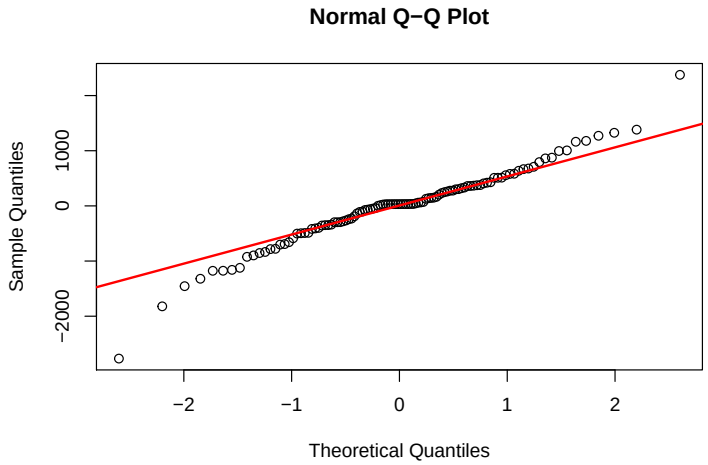
	2.5 %	97.5 %
ar1	-0.6279541	-0.1281756
ar2	0.3972884	0.7802258
ar3	0.3242364	0.6759107
ma1	0.5631467	1.0602687
sar1	-0.9658740	-0.5629042
sar2	-0.5633209	-0.1557617

Notamos que cambia considerablemente el ajuste.

Nuestro modelo a comparar es SARIMA(3,0,1)(2,1,0)[12]

Verificación de Supuestos

Normalidad




```
##
##  Anderson-Darling normality test
##
## data:  fi1$residuals
## A = 1.1968, p-value = 0.003846
```

```
##
##  Jarque Bera Test
##
## data:  fi1$residuals
## X-squared = 29.778, df = 2, p-value = 3.418e-07
```

No pasa normalidad

Varianza constante

```
##
##  studentized Breusch-Pagan test
##
## data:  Y ~ X
## BP = 0.12463, df = 1, p-value = 0.7241
```

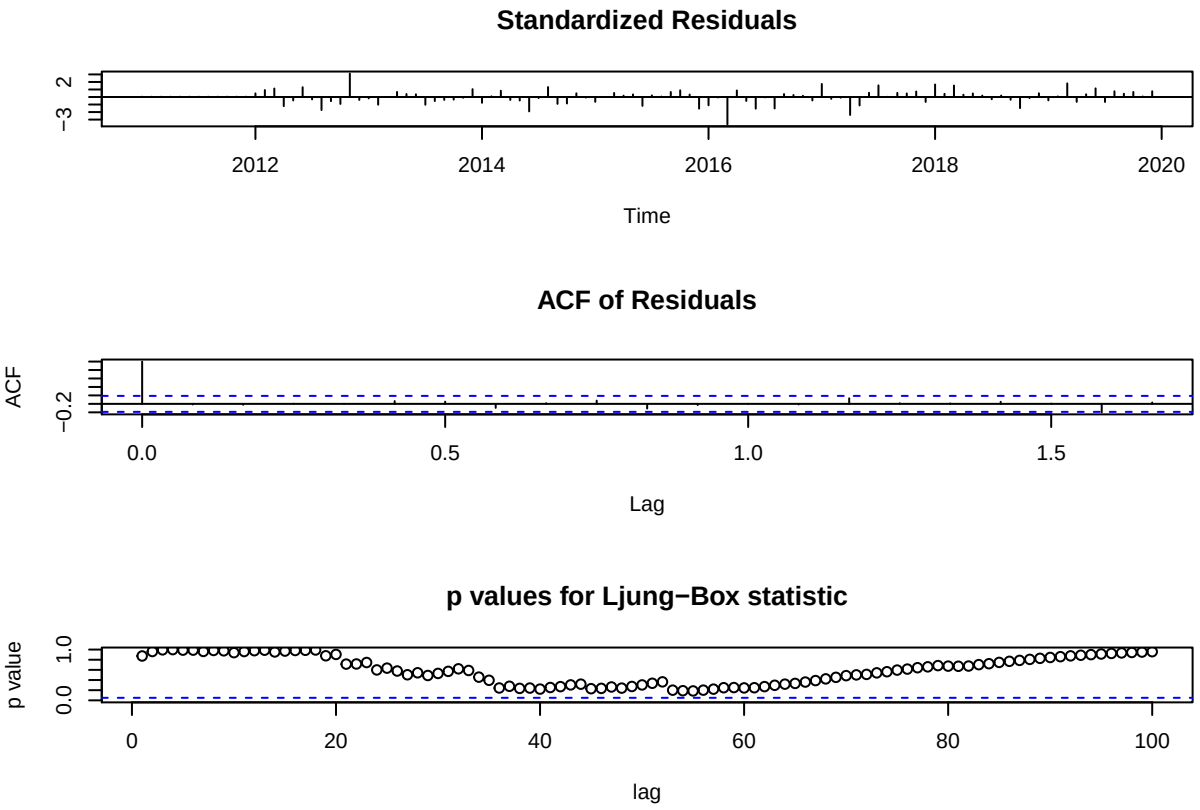
Tiene varianza constante

Media cero

```
##
##  One Sample t-test
##
## data:  fi1$residuals
## t = -0.086655, df = 107, p-value = 0.9311
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
```

```
##  -139.2470  127.5832
## sample estimates:
## mean of x
## -5.831899
```

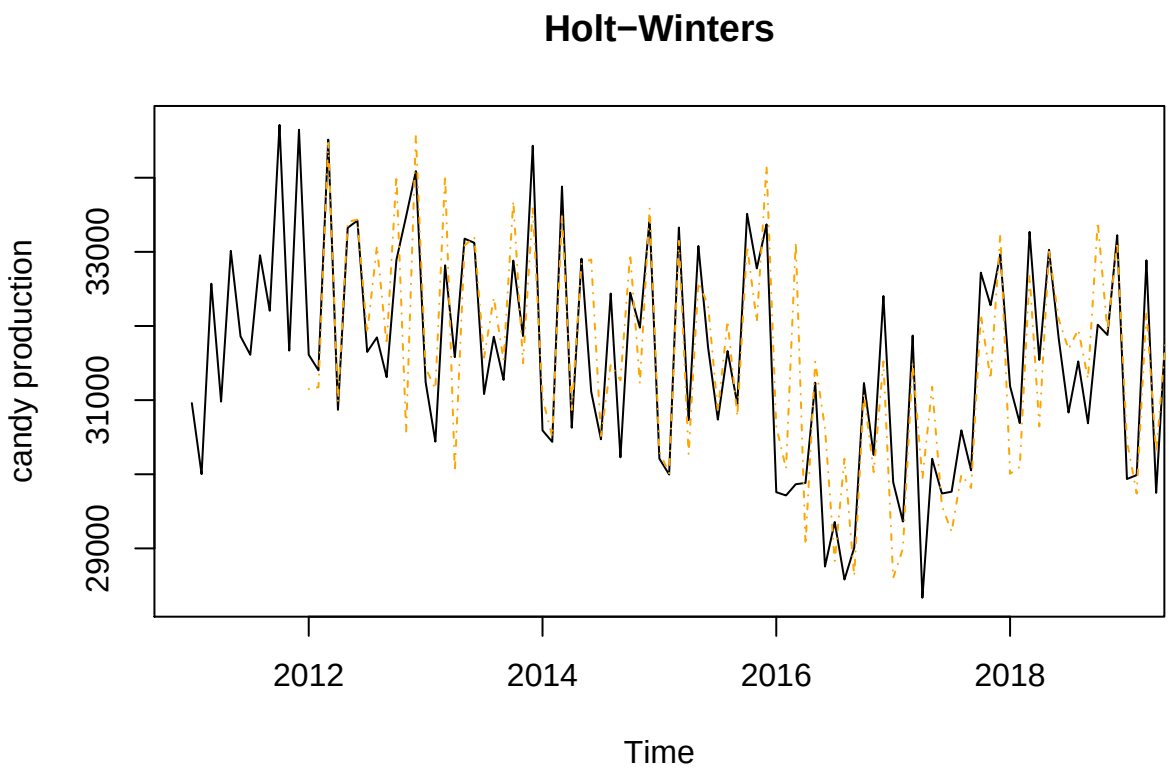
Tiene media cero



Independencia
Cumple independencia.

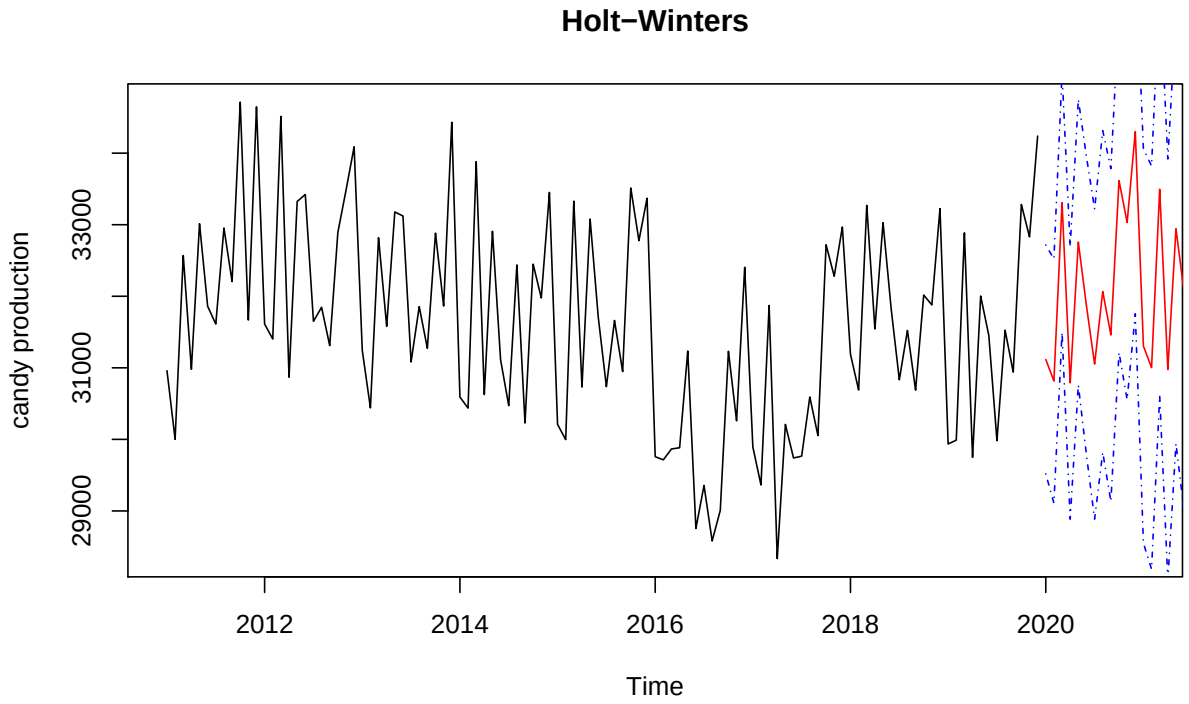
Nuestro modelo nos arroja que tenemos un ruido blanco no gaussiano.

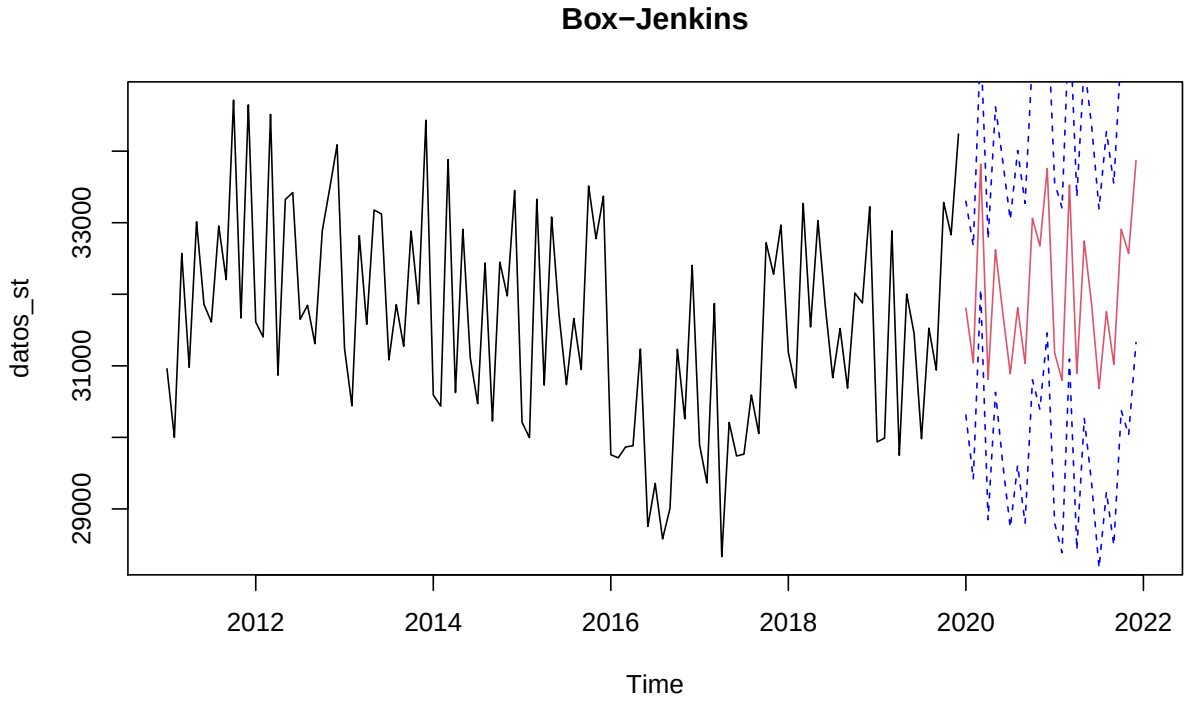
Holt-Winters 2011-2019



Notamos que nuestro ajuste es ligeramente bueno.

Comparación de predicción



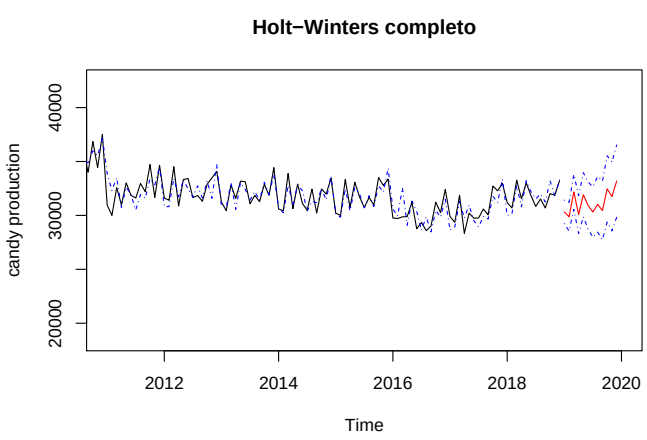


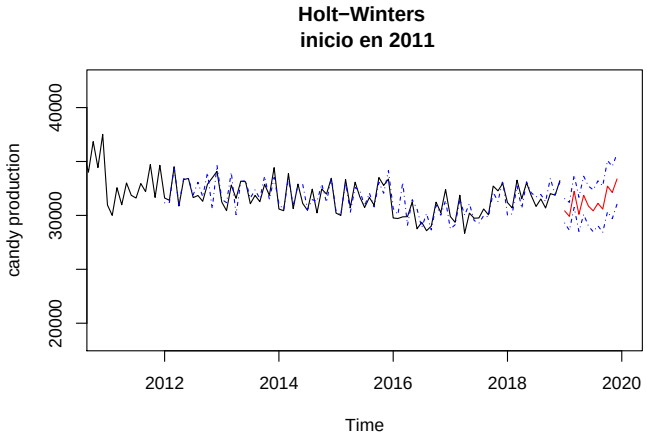
De nuevo pensamos que modelando con Holt-Winters es la mejor opción pero para salir de dudas harémos Back Testing.

Back Testing

Se intentó hacer de back testing para decidir cuál serie de tiempo es más óptima para predecir. Mediante Holt-Winters comparamos los errores cuadráticos de la serie completa vs la serie que comienza en 2011.El año 2019 será nuestra base de prueba.

```
##        Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
## 1997 18515 18407 20004 20175 20915 20773 20609 20355 20754 21579 22435 23593
## 1998 20309 20350 21847 21062 21896 21746 21325 22036 21498 23285 22668 24665
## 1999 23898 20392 24369 23442 24030 22857 22554 24471 24490 26042 22245 26704
## 2000 24939 24484 25726 24053 25389 25166 25431 25658 26558 27421 26894 30219
## 2001 27241 26460 30856 27863 30690 31035 30948 31571 31492 32194 31515 33004
## 2002 30048 30306 33064 31953 32126 32957 32262 34292 34432 35058 35545 36959
## 2003 33045 33139 36509 33978 36388 35813 34882 35937 34584 36927 35647 37641
## 2004 34183 35300 38045 34789 37420 37331 36945 37337 36806 38854 37036 39561
## 2005 36943 35823 36720 36830 38824 37411 36770 37705 37945 39148 37530 40584
## 2006 37143 36260 39837 37662 40582 40468 38334 39168 39337 40591 39359 42531
## 2007 37901 38026 41678 37198 40127 41006 38469 39647 40977 41431 38726 41093
## 2008 37830 38986 39155 39186 40683 39302 36902 38421 36835 39855 39089 40191
## 2009 35235 35498 38587 34625 36433 36213 34187 34876 33558 34882 36091 38282
## 2010 35880 34700 37599 34279 37308 35196 33677 35821 34006 36847 34436 37518
## 2011 30960 30001 32571 30980 33014 31860 31612 32954 32205 34713 31668 34647
## 2012 31612 31403 34514 30869 33325 33422 31650 31846 31310 32891 33480 34089
## 2013 31245 30440 32819 31580 33178 33122 31082 31855 31273 32882 31864 34432
## 2014 30592 30437 33881 30627 32908 31113 30472 32438 30228 32449 31977 33451
## 2015 30210 29997 33329 30731 33080 31711 30737 31662 30948 33513 32777 33371
## 2016 29758 29714 29866 29883 31235 28755 29356 28580 29008 31231 30260 32405
## 2017 29894 29362 31871 28334 30209 29740 29765 30593 30053 32720 32280 32968
## 2018 31188 30690 33270 31543 33028 31825 30833 31522 30688 32017 31879 33225
```






```
## [1] 2305541
```

```
## [1] 1597678
```

	Back Testing
Serie	MSE
1997-2018	2305541
2011-2018	1597678

Notamos que el error cuadático medio es menor en la serie con los datos al inicio del 2011, por ello nuestras suposiciones erran correctas, es mejor predecir con la serie no completa.

La razón por la que decidimos cortar nuestra serie completa, fue porque habia un cambio decreciente, el cual es muy notorio alrededor de 2007-2010, y en 2011 se comenzaba a estabilizar la tendencia.

Por otro lado, una razón social de este cambio se debe a que en 2011 México se adhirió al *Decenio de Acción por la Seguridad Vial 2011-2020* promovido por las Naciones Unidas, creando la *Estrategia Nacional de Seguridad Vial 2001-2020*, cuyo objetivo fue reducir los accidentes fatales y no faltales 50%, promoviendo la participación de las autoridades gubernamentales, teniendo un efecto significativo, ya que vimos que los accidentes no fueron creciendo y se han mantenido con una varianza constante, aunque podría mejorar en un futuro, con nuevas politicas de tránsito.

Referencias

- Albarrán Naranjo Lizbeth. (2021). Series de Tiempo. Facultad de Ciencias, UNAM.
- Paul S.P. Cowpertwait · Andrew V. Metcalfe. (2011). Introductory Time Series with R. Fairview Avenue, N. M2-B876 Seattle, Washington 98109 USA Giovanni Parmigiani: Springer.
- Caminos y Puentes Federales. (2021). Decenio de acción para la seguridad vial. noviembre 11,2021, de Gobierno de México Sitio web: https://www.gob.mx/capufe/articulos/decenio-de-accion-para-la-seguridad-vial-265479#:~:text=Derivado%20de%20ello%2C%20en%20marzo,accidentes%20vehiculares%20en%20el%20mundo