TRABAJO TÉCNICAS DE PREVISIÓN

Autores: Javier Porcel Marí

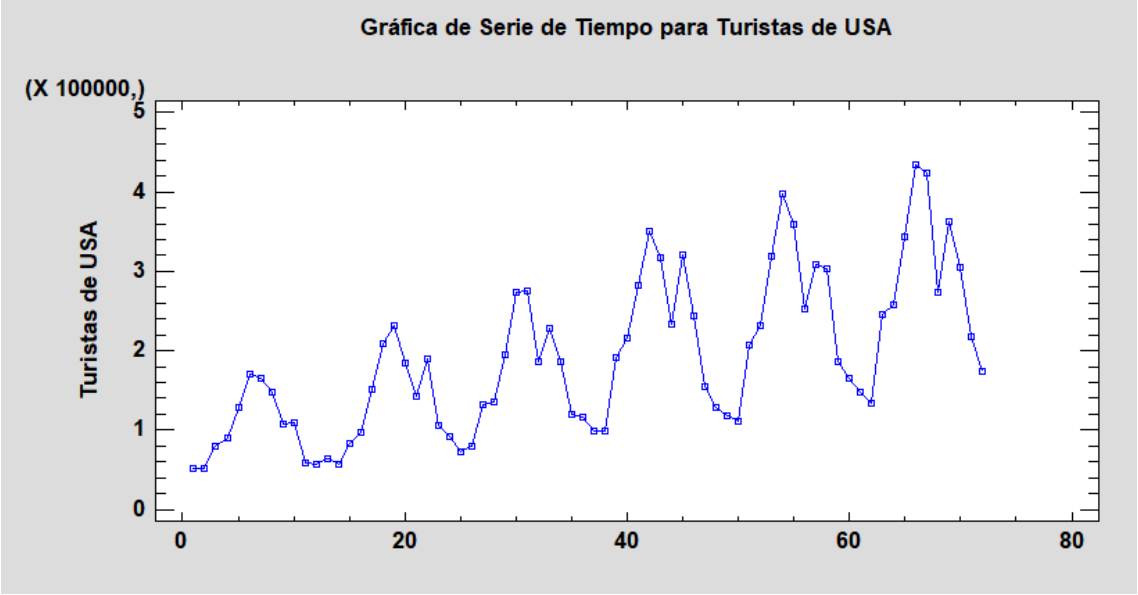
Luis Ribelles García

1. Introducción. Objetivos

1.1. Datos de la serie.

En este trabajo se ha decido trabajar con los datos de turistas procedentes de Estados Unidos que han llegado a España cada mes desde enero 2014 hasta diciembre 2019. En base a estos datos se quiere predecir el número de turistas que llegaran a España en el año 2020, si las condiciones sanitarias, que existían en los años 2014 y 2019 se hubieran mantenido durante el año 2020. Los datos del número de turistas norteamericanos que llegaron a España se han obtenido en el INE para aquellos datos posteriores a septiembre del año 2015 mientras que para los datos anteriores a esa fecha se ha consultado Tour Spain.

La representación gráfica de la serie de turistas norteamericanos que visitan España cada mes es la siguiente:



1.2. Tendencia.

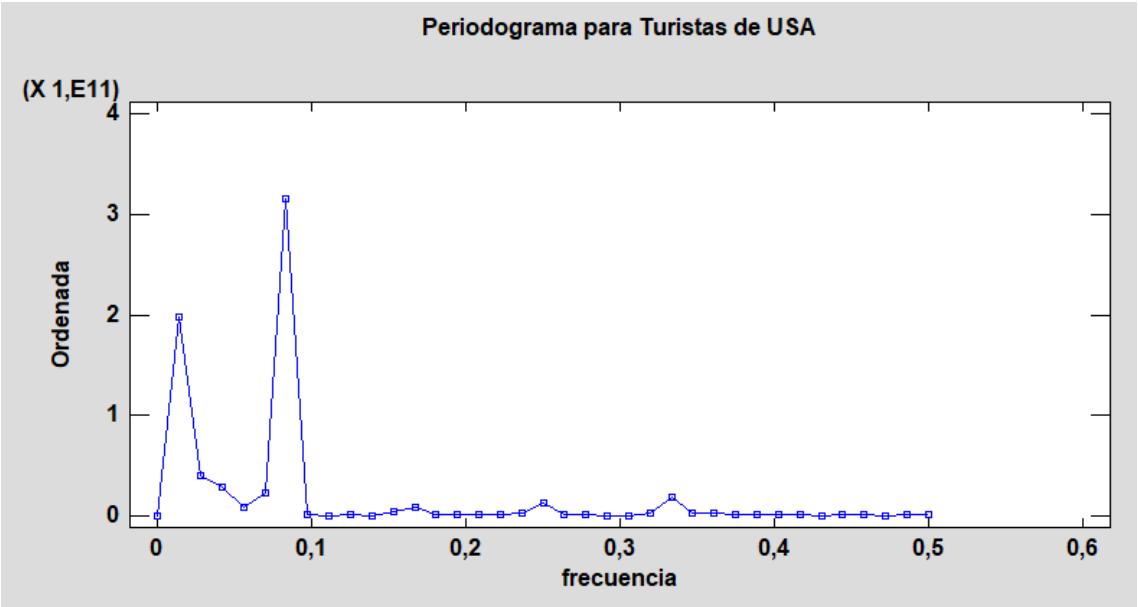
Como se puede observar en el gráfico, la serie mantiene una tendencia lineal creciente durante toda la serie temporal. Puesto que existe tendencia, la media no es constante y por lo tanto se trata de un proceso no estacionario.

1.3. Varianza.

La serie temporal no tiene varianza constante a lo largo de esta. Esto se aprecia en la representación gráfica de los datos crece puesto que a medida que la serie temporal avanza en el tiempo la distancia entre los máximos y los mínimos de cada estacionalidad es mayor lo que indica una varianza creciente. Por lo tanto, esta serie, no se trata de un proceso estacionario y habrá que transformarla para poder trabajar con modelos de predicción ARIMA.

1.4. Estacionalidad.

En la representación gráfica de la serie temporal se puede apreciar una componente estacional que se repite cada 12 meses. Para comprobar que longitud del pedido estacional es de 12 meses se ha utilizado el periodograma, calculando así la duración de dicho periodo.

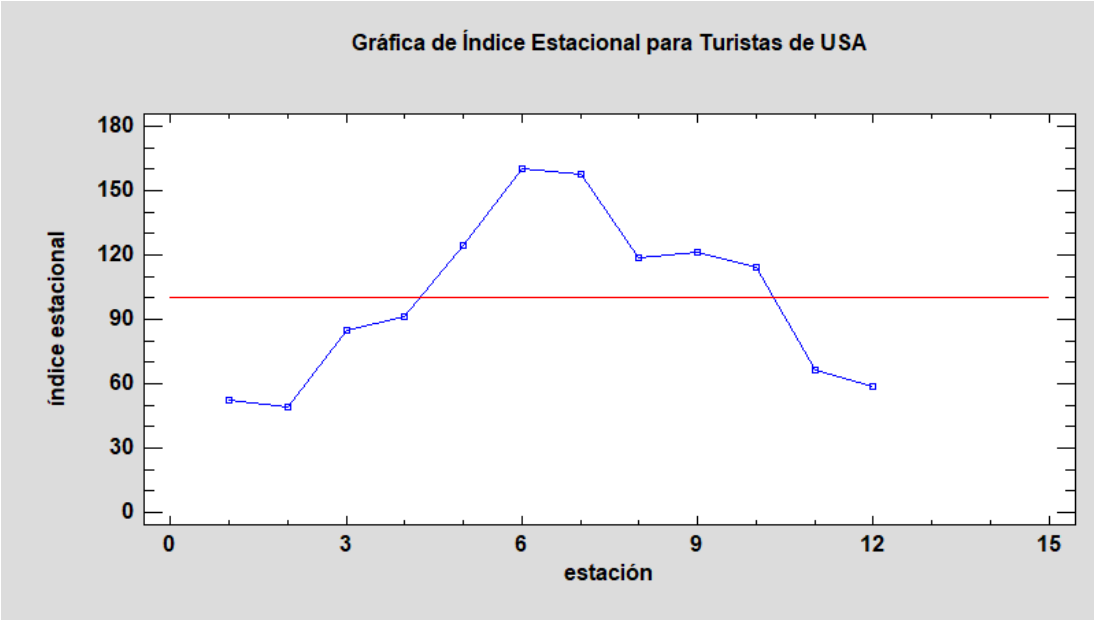


Como se puede apreciar en el periodograma el mayor pico se da cuando la frecuencia es igual a 0,083333. Por lo tanto, utilizamos el valor de la frecuencia en el pico para calcular la duración de periodo estacional.

$$Duración\ de\ la\ estacionalidad = \frac{1}{f} = 12$$

Por lo tanto, la duración de periodo estacional es de 12 meses.

Para estudiar cómo se comporta esta componente estacional se ha representado la gráfica de índice estacional:



Como se puede observar en el gráfico anterior los meses de verano son los meses en los que vienen mayor número de turistas norteamericanos a visitar España. Mientras que los meses invernales son los en los que menor número de turistas procedentes de los Estados Unidos, visitan nuestro país.

2. Modelos deterministas.

2.1. Preselección de modelos válidos.

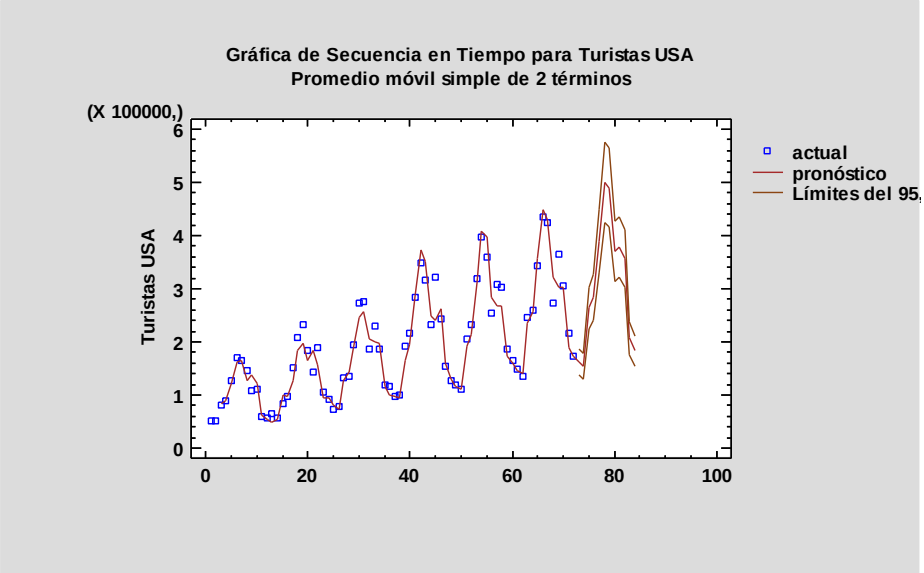
En este apartado se trata de hacer predicciones mediante cuatro modelos deterministas: modelo de medias móviles, modelo de suavizado exponencial simple, modelo de Holt y modelo de Holt-Winters.

Aunque se trabajará con los cuatro modelos, es de esperar que el modelo de Holt-Winters proporcione mejores resultados que los otros modelos pues este modelo se ajusta bien a datos que presentan una tendencia y una estacionalidad marcada al igual que los datos que se tratan en este trabajo, como se ha explicado en la sección anterior.

2.2. Análisis de los modelos.

2.2.1. Modelo de medias móviles.

Se obtiene el siguiente gráfico al aplicar el modelo de medias móviles simple de segundo orden multiplicativo (se ha escogido esta configuración porque retorna un RMSE más bajo):

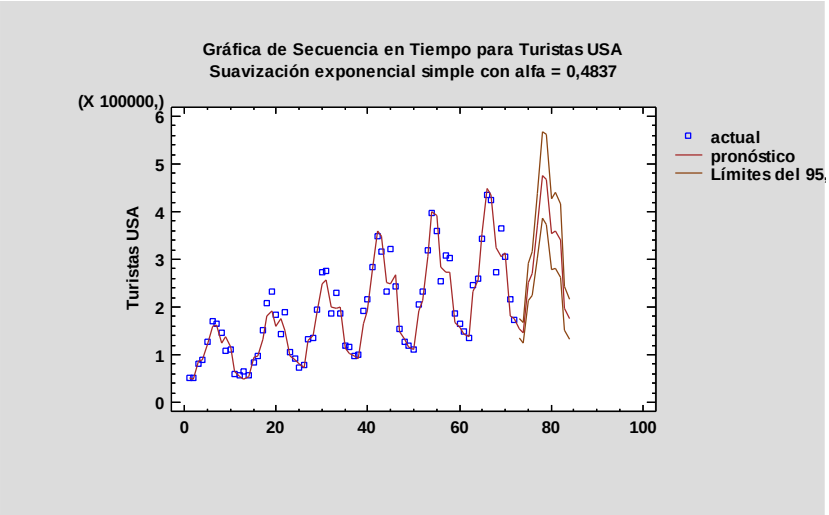


Lo que se traduce en la siguiente tabla de pronósticos:

Periodo	*Pronóstico*	*Límite en 95,0% Inferior*	*Límite en 95,0% Superior*
73,0	162784,	138093,	187476,
74,0	153854,	130517,	177191,
75,0	263765,	223757,	303774,
76,0	283712,	240678,	326746,
77,0	388341,	329436,	447245,
78,0	498995,	423306,	574683,
79,0	490319,	415946,	564691,
80,0	370605,	314391,	426819,
81,0	377852,	320538,	435165,
82,0	356094,	302081,	410107,
83,0	207378,	175922,	238833,
84,0	183162,	155379,	210944,

2.2.2. Modelo exponencial simple.

Al utilizar el modelo exponencial simple multiplicativo y con un parámetro α = = 0,4837 resultado de optimizar dicho parámetro se consigue este gráfico:

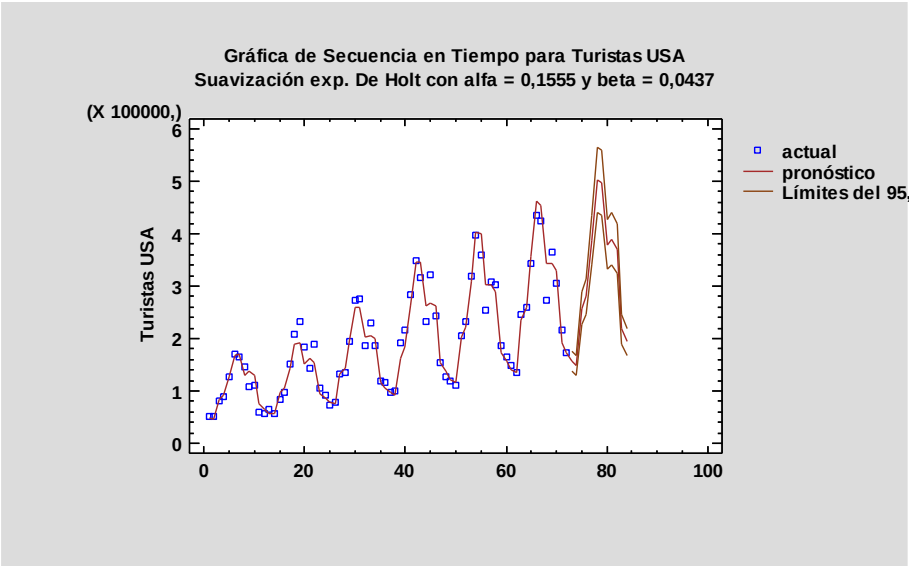


que da lugar a la siguiente tabla de pronósticos:

Periodo	*Pronóstico*	*Límite en 95,0% Inferior*	*Límite en 95,0% Superior*
73,0	155305,	135246,	175363,
74,0	146785,	125726,	167844,
75,0	251646,	212268,	291023,
76,0	270676,	225070,	316282,
77,0	370497,	303919,	437075,
78,0	476067,	385496,	566638,
79,0	467789,	374118,	561461,
80,0	353576,	279410,	427743,
81,0	360490,	281591,	439389,
82,0	339732,	262406,	417057,
83,0	197849,	151152,	244546,
84,0	174746,	132081,	217410,

2.2.3. Modelo de Holt.

Mediante el modelo de Holt multiplicativo con los parámetros optimizados α = = 0,155 y β = 0,0437 se construye esta gráfica:

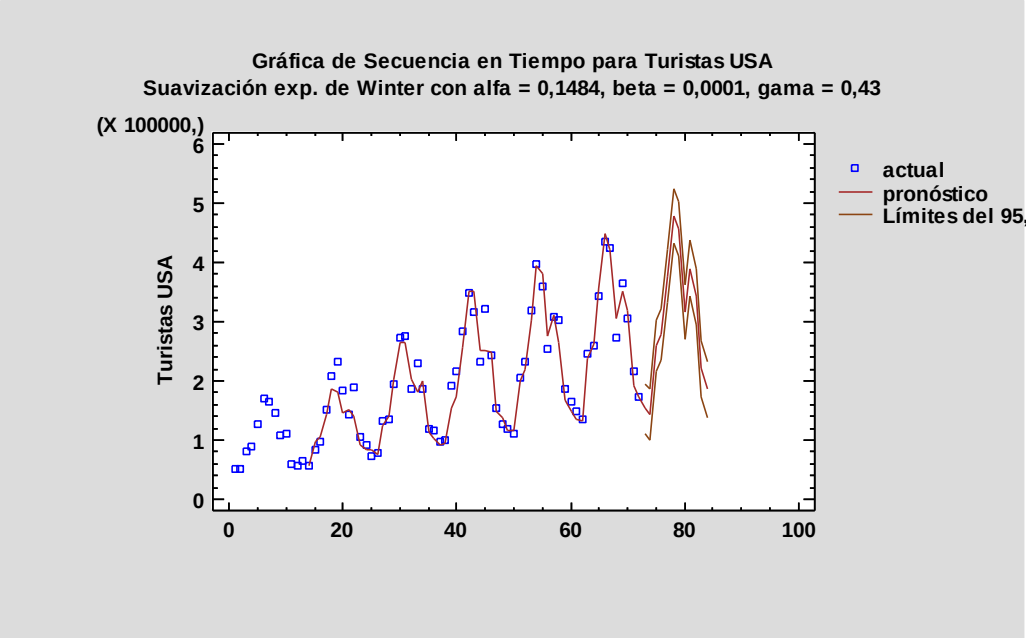


A su vez, se obtiene la siguiente tabla de pronósticos junto con el intervalo de confianza al 95% para cada pronóstico:

Periodo	*Pronóstico*	*Límite en 95,0% Inferior*	*Límite en 95,0% Superior*
73,0	156213,	137413,	175014,
74,0	149042,	131041,	167044,
75,0	257914,	226625,	289203,
76,0	279998,	245853,	314143,
77,0	386787,	339337,	434237,
78,0	501535,	439595,	563476,
79,0	497273,	435403,	559143,
80,0	379231,	331666,	426795,
81,0	390081,	340729,	439434,
82,0	370857,	323501,	418213,
83,0	217861,	189767,	245954,
84,0	194086,	168800,	219372,

2.2.4. Modelo de Holt-Winters.

Por último, se analiza el modelo de Holt-Winters con los parámetros optimizados α = =0,1484, β = 0,0001 y γ = 0,43. A partir de estos parámetros se ha obtenido la siguiente gráfica que nos muestra también el intervalo de confianza al 95% (o como se nombra en la leyenda: límites):



Junto con esta gráfica, el software Statgraphics nos proporciona la siguiente tabla de pronósticos:

Periodo	*Pronóstico*	*Límite en 95,0% Inferior*	*Límite en 95,0% Superior*
73,0	153262,	110677,	195847,
74,0	144373,	101321,	187424,
75,0	258719,	215206,	302232,
76,0	278679,	234709,	322649,
77,0	376448,	332026,	420870,

78,0	477919,	433050,	522789,
79,0	456384,	411070,	501697,
80,0	316973,	271221,	362725,
81,0	390400,	344213,	436588,
82,0	342058,	295439,	388676,
83,0	221201,	174155,	268247,
84,0	186194,	138725,	233664,

2.3. Comparación de modelos.

Para comparar los cuatro modelos que se han analizado en este apartado se compararan las estadísticas del error mediante diversas maneras de medir el mismo (RMSE, MAE, MAPE, …). Sin embargo, de los diversos métodos que son útiles para conocer el error del modelo nos guiaremos por el RMSE. Así pues, estos son los errores para cada modelo:

Modelo	*RMSE*	*MAE*	*MAPE*	*ME*	*MPE*
(A)	24111,1	16016,4	8,24467	3720,11	1,59791
(B)	23457,3	15608,1	8,12654	4859,77	2,31279
(C)	22305,6	14657,7	8,03672	-585,984	-0,703926
(D)	22301,4	15796,6	7,8352	6389,36	3,02737

donde el modelo A corresponde al modelo de medias móviles multiplicativo de segundo orden, el modelo B al modelo de suavizado exponencial simple multiplicativo optimizado, el modelo C al modelo de Holt multiplicativo optimizado y el modelo D al modelo de Holt-Winters optimizado.

Tal y como ya se predijo en la preselección de modelos válidos el mejor modelo, siguiendo el criterio de minimizar el RMSE, es el correspondiente al modelo de Holt-Winters pues este modelo actúa de manera adecuada con datos con tendencia y estacionalidad marcada.

Además, se realizan cinco pruebas de validación: prueba de corridas excesivas arriba y abajo (RUNS), prueba de corridas excesivas arriba y abajo de la mediana (RUNM), prueba de Box-Pierce para autocorrelación excesiva (AUTO), prueba para diferencia en medias entre la 1ª mitad y la 2ª mitad (MEDIA) y prueba para diferencia en varianza entre la 1ª mitad y la 2ª mitad (VAR). El modelo de Holt-Winters (modelo D) supera las cinco pruebas tal y como se muestra en esta tabla proporcionada por el Statgraphics:

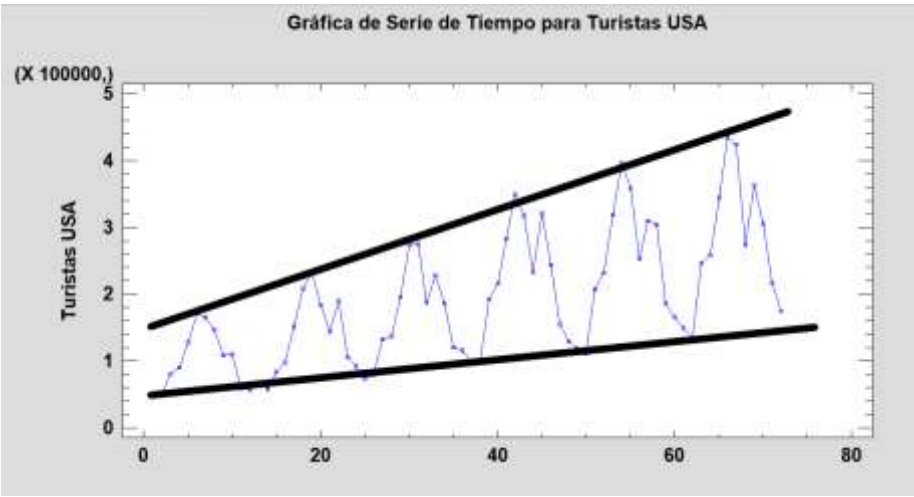
Modelo	*RMSE*	*RUNS*	*RUNM*	*AUTO*	*MEDIA*	*VAR*
(A)	24111,1	**	OK	*	OK	*
(B)	23457,3	OK	OK	*	OK	*
(C)	22305,6	*	OK	*	OK	*
(D)	22301,4	OK	OK	OK	OK	OK

4. Modelos ARIMA

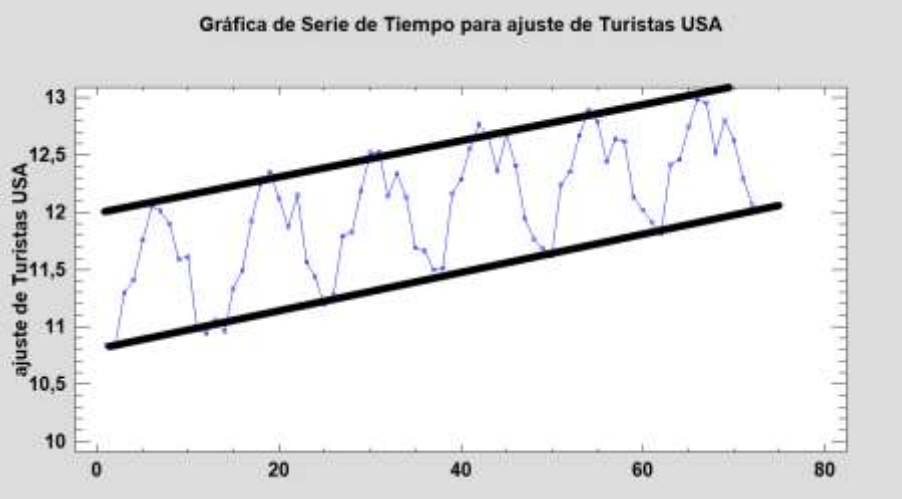
4.1 Metodología Box-Jenkins

En primer lugar, la metodología Box-Jenkins nos indica que se hagan las transformaciones necesarias a la serie temporal para que esta tenga media y varianza constante además de no estacionalidad. Así pues, se aplica un logaritmo neperiano a la serie para conseguir varianza constante, una diferencia regular para alcanzar la media constante y finalmente una diferencia estacional (tomando una estacionalidad de 12 periodos tal y como se ha razonado en el apartado 1) para eliminar la estacionalidad. Estos pasos se ilustran en las siguientes gráficas:

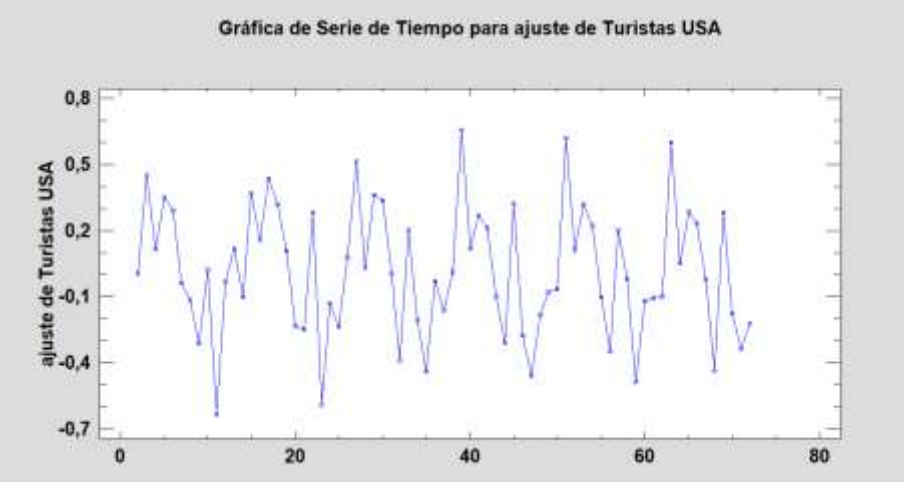
Paso 1: Se detecta que la varianza es creciente en la serie Z(t).



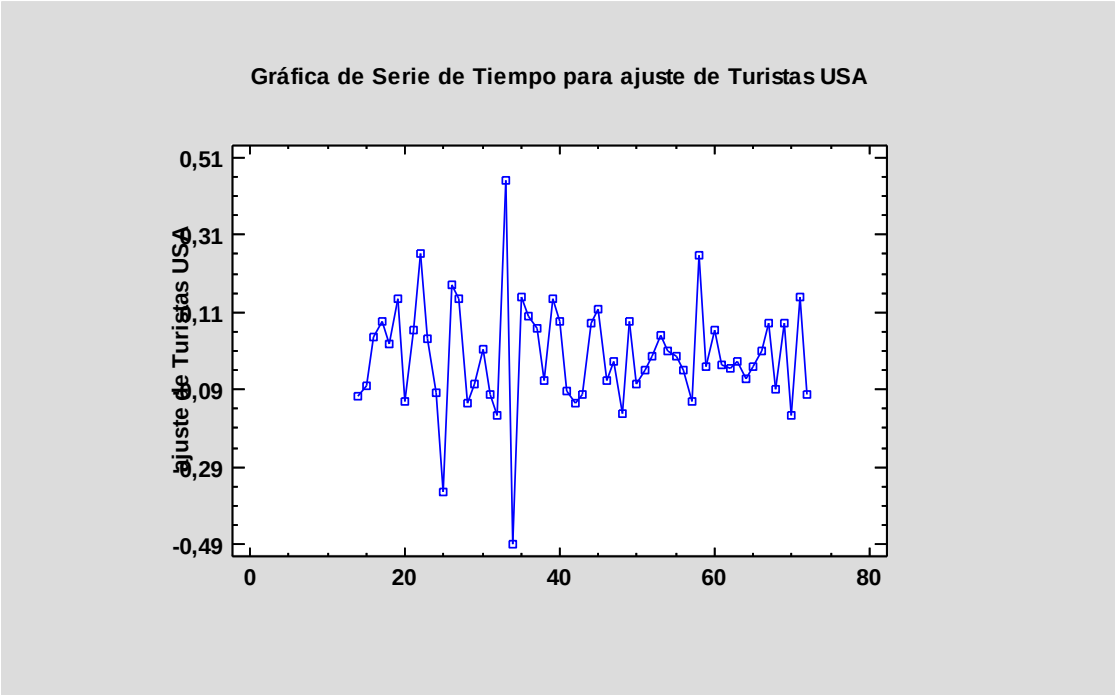
Paso 2: Se consigue estabilizar la varianza mediante W(t) = ln (Z(t)). La media sigue sin ser constante.



Paso 3: Se estabiliza la media mediante $X(t) = \nabla^1 W(t)$. La estacionalidad sigue latente.

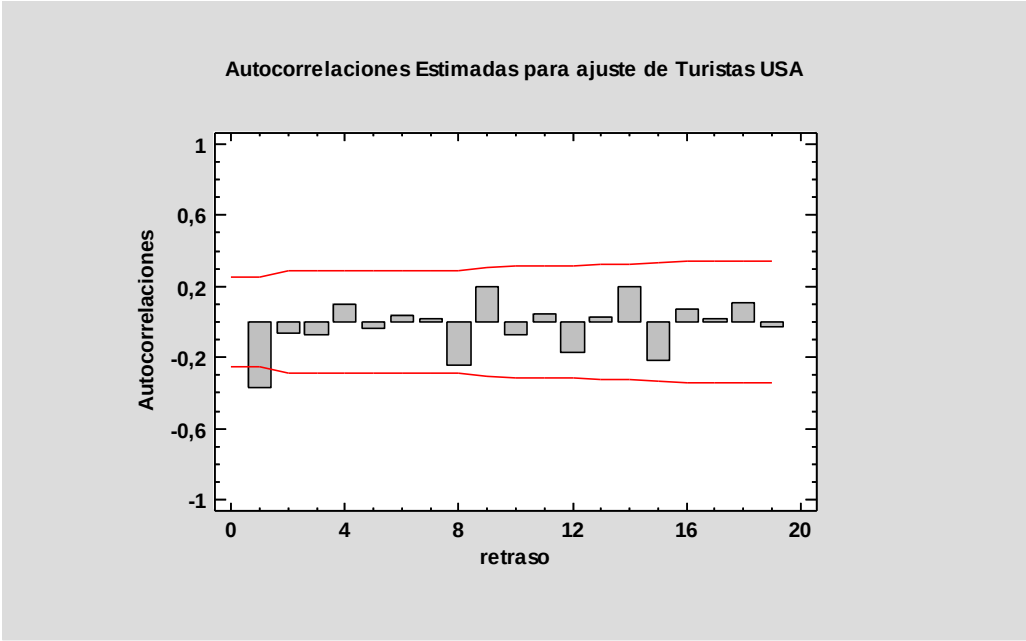


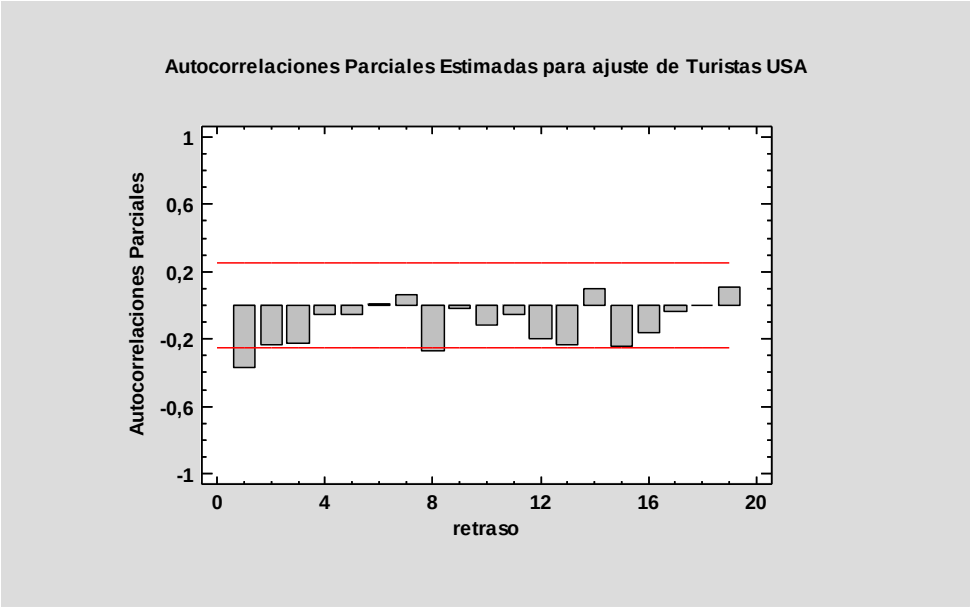
Paso 4: Se elimina la estacionalidad mediante $Y(t) = \nabla^1_{12} X(t)$



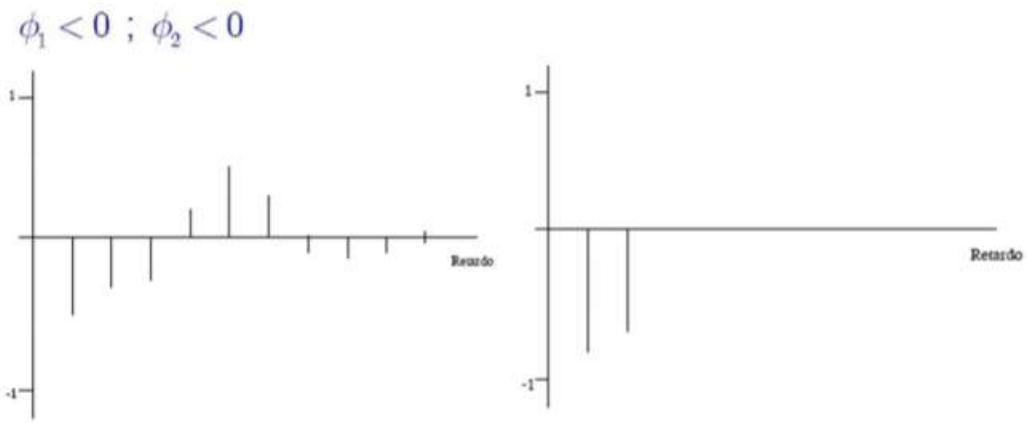
Con estas transformaciones se obtiene la gráfica superior que confirma que se han alcanzado los tres objetivos.

Una vez se ha estabilizado la varianza y la media, además de eliminar la estacionalidad se procede a analizar las funciones de autocorrelación simple y parcial por tal de dislumbrar algún modelo ARIMA tentativo. Estas son las funciones de autocorrelación que nos proporciona el software:

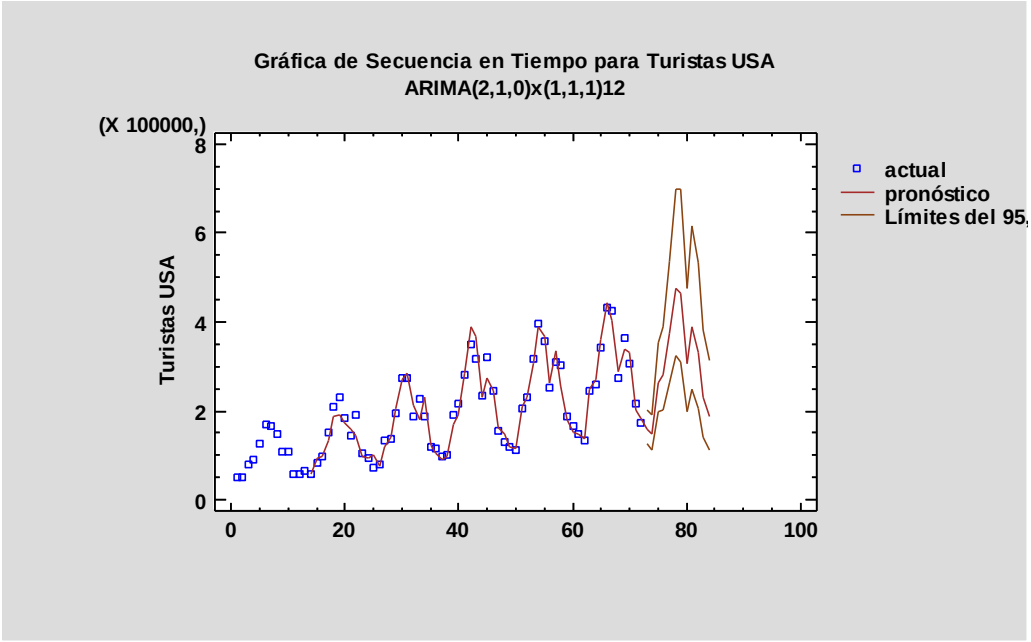




A partir de estas funciones de autocorrelaciones y comparándolas con las teóricas correspondientes se escoge el modelo ARIMA tentativo ARIMA (2, 1, 0)x$(1, 1, 1)_{12}$. De hecho, estas serían las funciones de autocorrelación teóricas para AR (2) cuando los dos parámetros son negativos en las que se observa cierta similitud con las funciones de autocorrelación de nuestra serie:



Así pues el software nos devuelve los valores AR(1) = -0,458274, AR(2) = -0,239071, SAR(1) = 0,629296 y SMA(1) = 0,989524 con una media y una constante no significativas, y además nos proporciona la siguiente gráfica:



Para finalizar se muestra la tabla de pronósticos que nos proporciona el ARIMA seleccionado para los próximos 12 periodos:

Periodo	*Pronóstico*	*Límite en 95,0% Inferior*	*Límite en 95,0% Superior*
73,0	160009,	126335,	202660,
74,0	147533,	113130,	192397,
75,0	264228,	196722,	354899,
76,0	280288,	201434,	390012,
77,0	375303,	262426,	536733,
78,0	475739,	324131,	698260,
79,0	464740,	308859,	699295,
80,0	307839,	199915,	474025,
81,0	390233,	247915,	614249,
82,0	331480,	206214,	532840,
83,0	232152,	141551,	380743,
84,0	188413,	112686,	315029,

5. Validación del modelo escogido.

5.1. Análisis de la significación de los parámetros.

Para comprobar la significación de los parámetros del modelo escogido, se ha realizado la predicción utilizando el software Statgraphics, del cual se ha extraído la siguiente tabla que evalúa la significación de los parámetros.

Resumen de Modelo ARIMA

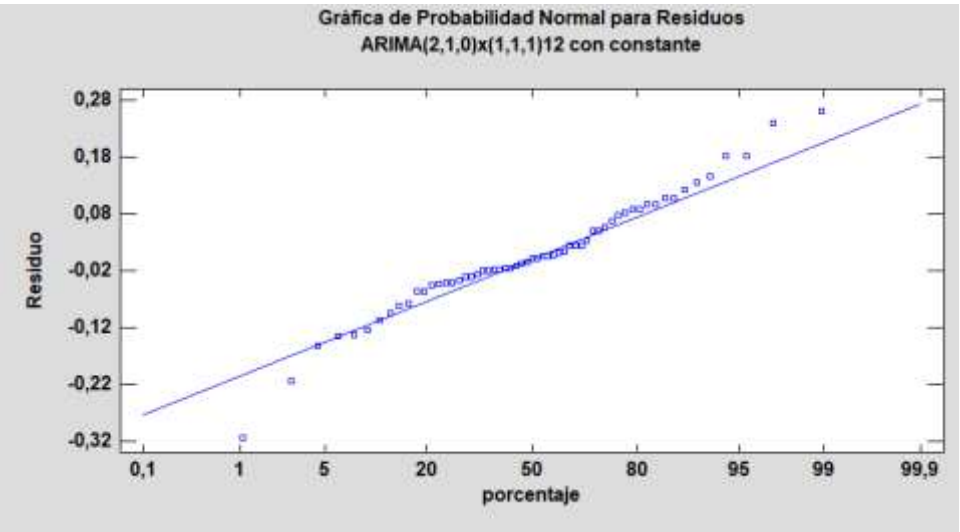
Parámetro	*Estimado*	*Error Estd.*	*t*	*Valor-P*
AR(1)	-0,487252	0,131888	-3,69444	0,000516
AR(2)	-0,210815	0,130676	-1,61327	0,112516
SAR(1)	0,655836	0,150081	4,36988	0,000057
SMA(1)	1,00272	0,0600396	16,701	0,000000
Media	-0,00380878	0,00690506	-0,551592	0,583503
Constante	-0,0022259			

Como se puede observar los parámetros AR (1), SAR (1) y SMA (1) son significativos para un nivel de confianza del 90%, puesto que su p-valor es inferior a 0,1. En el caso del parámetro autorregresivo AR (2) es casi significativo para un nivel de confianza del 90%, a causa de que su p-valor es 0,11 y puesto a que su selección está en concordancia con la F.AS y la F.A.P, se va a mantener en nuestro modelo. Sin embargo, la media al tener un p-valor de 0,58, no es significativa y por lo tanto tampoco lo es la constante.

5.2. Análisis de los residuos: ¿Ruido blanco?

5.2.1. Media nula y distribución normal.

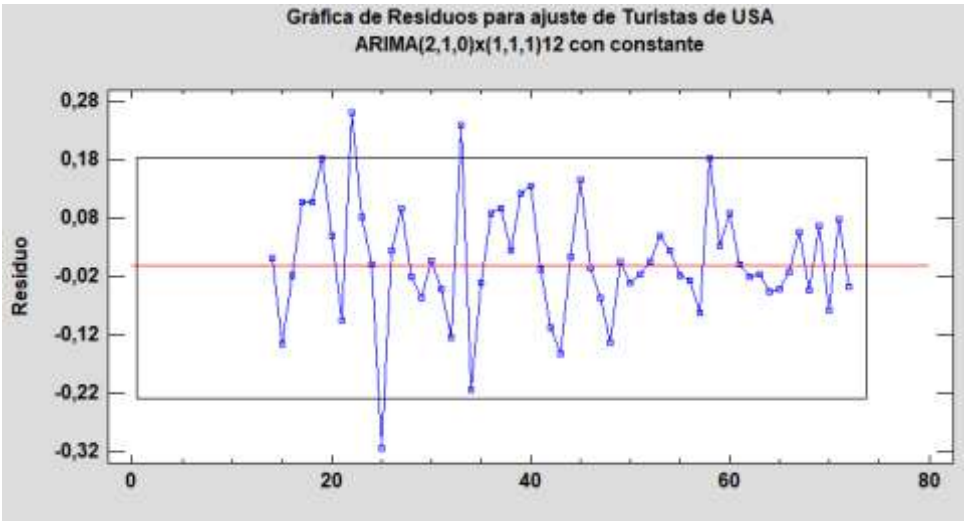
Con el objetivo de comprobar si los residuos tienen media nula y siguen una distribución normal se ha realizado el papel probabilístico normal para los residuos.



Como se puede observar en el papel probabilístico normal, los residuos se ajustan aproximadamente a una recta, por lo que se puede concluir que siguen una distribución normal. Adicionalmente como en el percentil 50 el valor de los residuos es aproximadamente 0, la media de los mismos es nula.

5.2.2. Varianza constante

Se ha representado una gráfica de los residuos frente al tiempo para comprobar si la varianza de los residuos es constante.



Como se observa en la gráfica de residuos frente al tiempo estos fluctúan dentro de una banda rectangular lo que nos permite aceptar la varianza permanece constante.

5.2.3. ¿Están incorrelacionados?

Para comprobar que los residuos si los residuos están Inter correlacionados o son independientes entre sí se ha realizado los siguientes test de aleatoriedad de los residuos, utilizando el software de Statgraphics;

Prueba de Aleatoriedad de residuos
Variable de datos: Turistas de USA
Modelo: ARIMA(2,1,0)x(1,1,1)12 con constante
Ajuste matemático: Log natural

(1) Corridas arriba o abajo de la mediana
Mediana = 0,000357983
Número de corridas arriba o abajo de la mediana = 28
Número esperado de corridas = 30,0
Estadístico z para muestras grandes = 0,397421
Valor-P = 0,691054

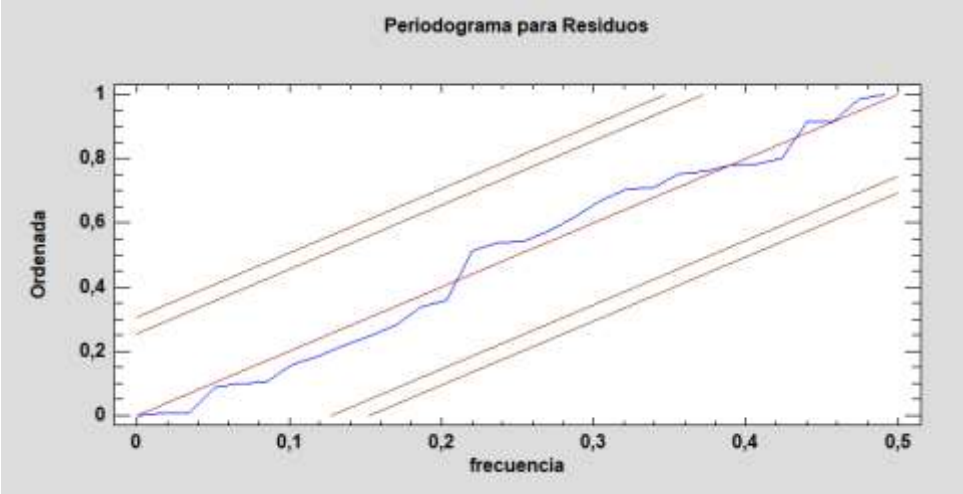
(2) Corridas arriba y abajo
Número de corridas arriba y abajo = 35
Número esperado de corridas = 39,0
Estadístico z para muestras grandes = 1,09769
Valor-P = 0,27234

(3) Prueba Box-Pierce
Prueba basada en las primeras 19 autocorrelaciones
Estadístico de prueba para muestras grandes = 15,5944
Valor-P = 0,409509

Como en los tres casos el p-valor es superior a 0,05, no se puede rechazar la hipótesis nula de que los residuos son aleatorios y no están Inter correlacionados entre sí.

5.2.4. Conclusión.

Para validar los análisis realizados anteriormente y comprobar que los residuos son ruido blanco se ha representado el periodo grama para los residuos.



Como se ha podido comprobar en los análisis de los residuos anteriormente realizados estos se distribuyen normalmente, tienen media nula, varianza constante y son

independientes, por lo tanto, se puede concluir que los residuos son ruido blanco. Esta conclusión está en concordancia con la representación del periodo grama de residuos en el cual se puede ver que estos fluctúan alrededor de la bisectriz principal sin sobrepasar los límites, por lo cual los residuos se comportan como ruido blanco. Debido a que los residuos son ruido blanco el modelo seleccionado es válido para realizar predicciones.

5.3. Comparación de modelos alternativos.

Con el objetivo se de comparar el modelo ARIMA, anteriormente seleccionado, se han planteado diferentes modelos alternativos y se ha calculado su RMSE. Se ha realizado la siguiente tabla donde se comparan el RMSE de los modelos alternativos con el modelo seleccionado.

Modelo	RMSE	Orden
ARIMA (2, 1 ,0) x (1, 1 ,1) $_{12}$	21568,7	1º
ARIMA (1,1,0) x $(1,1,0)_{12}$	27036,1	5º
ARIMA (1,1,0) x $(0,1,0)_{12}$	26910,7	4º
ARIMA (2, 1 ,0) x (0, 1 ,0) $_{12}$	26886,5	3º
Suavización exp. de Winter con alfa = 0,1485, beta = 0,0001, gama = 0,4305	22301,3	2º

Como se puede observar en la tabla el mejor modelo es el ARIMA (2, 1 ,0) x (1, 1 ,1)12 puesto que con un RMSE igual 21568,7 es el que tiene menor RMSE de los cinco modelos comparados y por lo tanto es el mejor modelo para predecir nuestra serie.

4. Explotación del modelo escogido.

4.1 Desarrollo de la ecuación explicita de predicción

Se busca la ecuación explícita de la predicción ARIMA (2, 1, 0) x $(1, 1, 1)_{12}$.

La ecuación de predicción de ARIMA (2, 1, 0) x $(1, 1, 1)_{12}$ en notación compacta es la siguiente;

$$\phi_{p=2}(B)\Phi_{P=1}(B^{s=12})\nabla^{d=1}\nabla_{s=12}^{D=1}W_t = \theta_{q=0}(B)\Theta_{Q=1}(B^{s=12})a_t$$

Se desarrolla la ecuación hasta obtener la ecuación de predicción explicita;

$$(1-\phi_1 B-\phi_2 B^2)(1-\Phi_1 B^{12})(1-B)(1-B^{12})W_t=(1-\Theta_1 B^{12})a_t$$

$$\downarrow$$

$$(1-\Phi_1 B^{12}-\phi_1 B+\phi_1\Phi_1 B^{13}-\phi_2 B^2+\phi_2\Phi_1 B^{14})(1-B^{12}-B+B^{13})W_t=$$

$$=-\Theta_1 a_{t-12}$$

$$\downarrow$$

$$\dots$$

$$\downarrow$$

$$Z_t=\exp(W_t)=\exp((\phi_1+1)W_{t-1}-(\phi_1-\phi_2)W_{t-2}-\phi_2 W_{t-3}+(\Phi_1+1)W_{t-12}-$$

$$-(1+\ \phi_1\Phi_1+\phi_1+\Phi_1)W_{t-13}-(\phi_2\Phi_1-\phi_1\Phi_1-\phi_1+\phi_2)W_{t-14}+(\phi_2\Phi_1+$$

$$+\phi_2)W_{t-15}-\Phi_1 W_{t-24}+(\phi_1\Phi_1+\Phi_1)W_{t-25}+(\phi_2\Phi_1-\phi_1\Phi_1)W_{t-26}-$$

$$-\phi_2\Phi_1 W_{t-27}-\Theta_1 a_{t-12})$$

4.2 Comprobación manual de las predicciones obtenidas con el ordenador.

Se obtiene la siguiente tabla al realizar una comprobación manual de las predicciones que nos devuelve el software StatGraphics, mediante la sustitución de los distintos valores en la ecuación que se ha hallado anteriormente:

Periodo	Wt	Zt	StatGraphics
73	11,9830	160010,3	160009
74	11,9018	147532,7	147533
75	12,4845	264207,5	264228
76	12,5435	280272,0	280288
77	12,8354	375263,0	375303
78	13,0725	475674,9	475739
79	13,0492	464718,4	464740
80	12,6373	307824,1	307839
81	12,8745	390234,7	390233
82	12,7112	331455,3	331480
83	12,3552	232154,6	232152
84	12,1463	188396,8	188413

Las pequeñas diferencias se deben al redondeo de decimales a pesar de que se han utilizado cuatro decimales para realizar los cálculos.

5. Conclusiones

Como conclusión en este trabajo cabe mencionar que se ha conseguido realizar predicciones para el año 2020 de la serie de Turistas residentes en Estados Unidos que visitan España cada mes, utilizando las diferentes metodologías aprendidas en la asignatura a lo largo del curso.

Las primeras predicciones se han realizado utilizado los siguientes métodos de predicción: medias móviles, suavizado exponencial simple, método de Holt y el método de Holt-Winter. De los diferentes modelos planteados para los diferentes métodos, el modelo que mejores predicciones hacía se trataba del método de Holt-Winter alfa = 0,1485, beta = 0,0001, gama = 0,4305.

Posteriormente se ha utilizado la metodología Box-Jenkins para seleccionar un modelo ARIMA adecuado para realizar predicciones. Puesto que la serie se trata de una serie no estacionaria se ha tenido que transformar a una serie estacionaria y observando la F.A.S y la F.A.P se ha seleccionado el siguiente modelo ARIMA $(2, 1, 0) \times (1, 1, 1)_{12}$.

Una vez validado el modelo utilizando la metodología Box-Jenkins se han realizado predicciones con el mismo y se ha comparado este modelo con diferentes modelos alternativos. Como se puede observar en la tabla de predicciones, según las predicciones realizadas para el año 2020 por nuestro modelo de predicción, se espera que la llegada de turistas Norte americanos a España crecerá durante el año 2020 moderadamente.

6. Referencias bibliográficas

Predicción en el dominio del tiempo: análisis de series temporales para ingenieros. (García Díaz, Juan Carlos)

Análisis de series temporales (Peña, Daniel)

Series temporales (González Velasco, Miguel | Puerto García, Inés María del | Universidad de Extremadura)

Forecasting: methods and applications (Makridakis, Spyros | Wheelwright, Steven C | Hyndman, Rob J)

Introduction to time series analysis and forecasting (Montgomery, Douglas C. | Jennings, Cheryl L. | Kulahci, Murat)

INE: https://www.ine.es/jaxiT3/Datos.htm?t=10822#!tabs-grafico

TOURSPAIN:
http://estadisticas.tourspain.es/esES/estadisticas/frontur/informesdinamicos/paginas/mensual.aspx