Estudio sobre la evolución del número de pasajeros en el metro de Barcelona

Un análisis basado en Series Temporales.

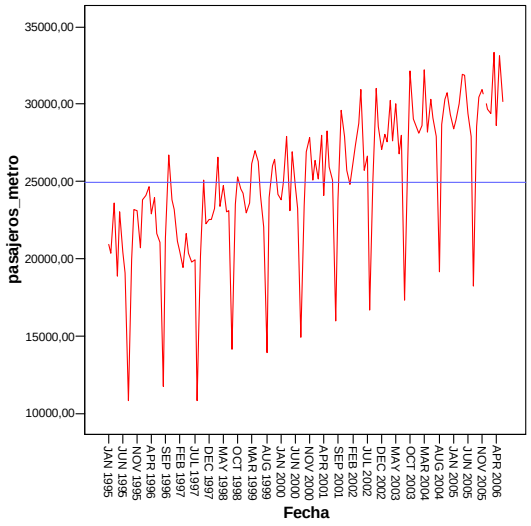
Autor: Origen

1.- Introducción

En el presente trabajo pretendemos modelar el número mensual de pasajeros del Metro de Barcelona, a partir de técnicas basadas en la metodología de Series Temporales. El modelo obtenido nos permitirá hacer predicciones sobre la variable de estudio para tiempos futuros. El poder realizar buenas predicciones sobre el número de pasajeros puede ser de gran ayuda a la hora de organizar los equipos necesarios para dar un buen servicio en el Metro.

La serie procede de la base de datos del Instituto Nacional de Estadística (http://www.ine.es/inebmenu/indice.htm) El periodo de observación de la serie está comprendido entre enero de 1995 y junio de 2006. La unidad de medida es "miles de viajeros".

El siguiente gráfico representa el comportamiento de la serie objeto de estudio:



2.- Metodología Estadística

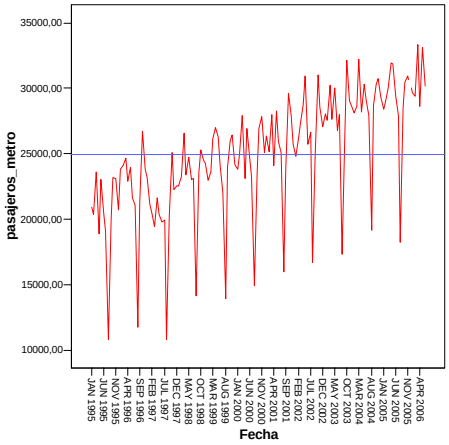
Para llevar a cabo el análisis de predicción de la serie "número mensual de pasajeros del metro de Barcelona", en primer lugar realizaremos, en la sección 3, un análisis descriptivo de la misma. Identificaremos las principales fuentes de variación haciendo uso de los métodos de descomposición. Además proporcionamos predicciones para la serie por medio del método clásico de suavizado exponencial

En la sección 4, haremos uso de la metodología de Box-Jenkins(véase Box, Jenkins and Reinsel (1994)) para la especificación de un modelo de la familia ARIMA que verosímilmente haya generado la serie de estudio. Asimismo, en la sección 5 analizaremos la presencia de posibles valores atípicos en la misma. Las predicciones de la serie haciendo uso de los modelos ajustados serán proporcionadas en la sección 6. Finalmente, estableceremos las conclusiones finales del trabajo en la sección 7.

Como herramientas de trabajo para el análisis estadístico y gráfico de los datos se utilizarán el software estadístico R (GNU S), versión 2.2.1 (véase R Development Core Team (2005)), el programa SPSS, versión 12, y el programa TSW (véase Gómez and Maraval (1996)).

3.- Análisis clásico de la serie

A la vista del gráfico de la serie:



Podemos decir que la serie presenta componente estacional (en periodos de un año se repite la misma conducta), tendencia-ciclo pues la serie no es estable en media (aumenta con el paso del tiempo) y componente irregular.

Además, se observa un corte en el mes de diciembre de 2005 y esto es debido a que el dato correspondiente a este mes no se ha recogido. También, en torno a febrero de 1997 se aprecia un comportamiento anómalo respecto a la tendencia y estacionalidad general de la serie.

Con la finalidad de describir adecuadamente la serie haremos uso tanto de los métodos de descomposición como de los métodos de suavizado exponencial.

3.1.- Métodos de descomposición

Aplicaremos los métodos de descomposición para identificar y aislar, de la forma más precisa posible, cada una de las componentes de variación presentes en la serie.

A simple vista tenemos una serie con componente tendencia-ciclo, componente estacional (de periodo 12) y componente irregular. Por tanto nuestra serie será de la siguiente forma: $X_t = f(T_t, S_t, I_t)$ donde X_t denota el dato en el instante t, T_t denota la componente tendencia, S_t la componente estacional e I_t la componente irregular, todo ello en el instante t.

Existen diferentes modelos de descomposición para series temporales: aditivo, multiplicativo y mixto, en nuestro caso nos decantaremos por el modelo aditivo pues la amplitud de las ondas no aumenta conforme transcurre el tiempo. En tal caso:

$$X_t = T_t + S_t + I_t$$

A continuación aplicaremos la descomposición clásica (basada en el concepto de medias móviles) y la descomposición STL, acrónimo de seasonal-trend descomposition based on Loess (véase Cleveland et al. (1990)). Para ello trabajaremos tanto con el SPSS como con el R restringiéndonos a modelos aditivos.

A partir de ahora eliminaremos los 6 últimos valores (a partir del dato perdido) para evitar problemas.

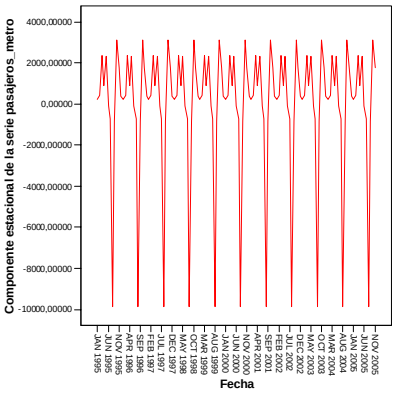
A. MÉTODOS DE DESCOMPOSICIÓN CLÁSICOS

Estos métodos están basados en el concepto de medias móviles. Haciendo uso del programa SPSS obtenemos los siguientes índices estacionales al aplicarle el modelo aditivo a nuestra serie pasajeros_metro:

Periodo	Indices estacionales	Periodo	Indices estacionales
1	230,444	7	-720,056
2	440,819	8	-9844,460
3	2348,703	9	-832,456
4	883,736	10	3109,594
5	2320,086	11	1783,728
6	-91,847	12	371,711

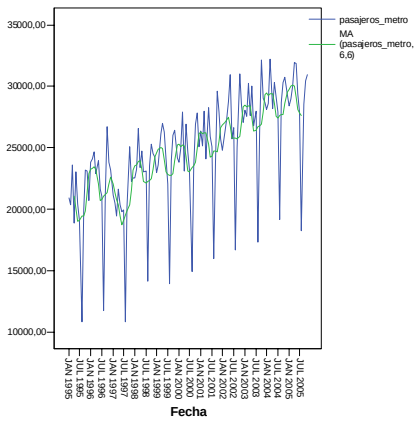
Como se observa el mes con menor índice estacional es agosto, le siguen septiembre, julio y junio. Los meses con mayor índice son octubre, marzo, mayo y noviembre. Una posible explicación a estas diferencias puede ser que las personas usan, en general, el metro para ir a trabajar por lo cual en las épocas de vacaciones el número de pasajeros disminuye. También es explicable el hecho de que el índice de agosto tenga una magnitud tan elevada (con signo negativo) pues gran parte de la población tiene vacaciones y abandona la ciudad.

En la gráfica de la componente estacional se observa esta diferencia de magnitud entre el índice de agosto y el resto:

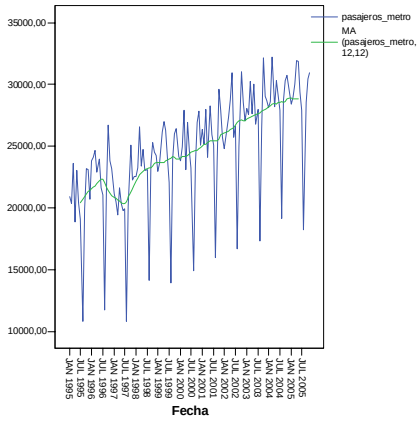


A continuación estimemos la tendencia realizando un suavizado mediante medias móviles centradas y comparémosla con la tendencia que nos proporciona el programa SPSS en la descomposición del modelo aditivo.

Como la serie tiene componente estacional, los órdenes que usaremos serán múltiplos del periodo pues si no, como se recoge en el primer gráfico, la tendencia se verá afectada por la componente estacional.



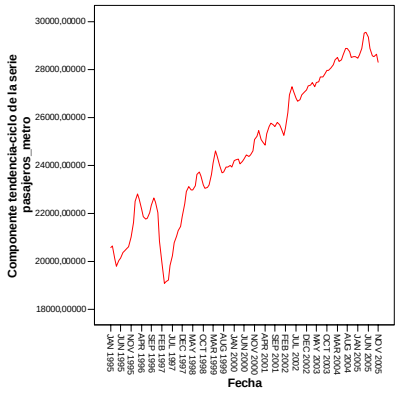
amplitud K=6
(perdemos 6 datos)



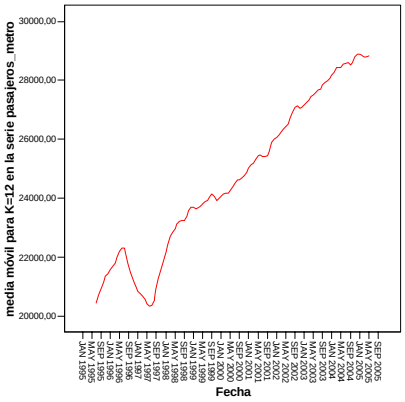
amplitud K=12
(perdemos 12 datos)

La mejor aproximación será la de K=12, pues se aproxima bien a la serie y no pierde demasiados datos.

A continuación vamos a representar la tendencia obtenida a partir de la descomposición para modelos aditivos proporcionada por el SPSS y al lado representamos la estimación obtenida a partir del método de media móvil para k=12.



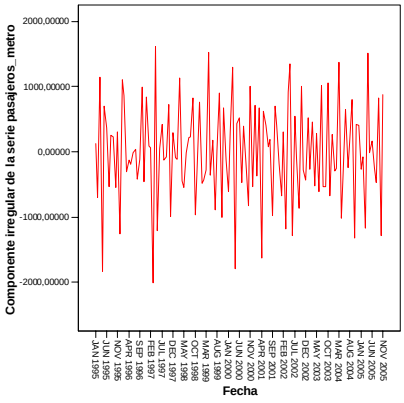
(tendencia descomposición SPSS)



(media móvil para k=12)

Según los gráficos obtenidos vemos que la tendencia que nos devuelve el SPSS de la descomposición estacional tiene todavía componente irregular. Podríamos considerar que la segunda es el suavizado de la primera. En ambas podemos sacar la misma conclusión: la serie aumenta y disminuye al principio en igual proporción para finalmente presentar tendencia creciente. Esta "irregularidad" puede ser debida a valores atípicos ó intervenciones en la serie. En la introducción ya comentábamos la existencia de un posible valor atípico en el primer tramo.

Finalmente, veamos que los residuos que hemos obtenido no tienen estructura estacional ni tendencia y además deben tener varianza constante y media 0, es decir, deben ser aleatorios. Si ocurre todo esto es que la descomposición es la adecuada.



Aparentemente se verifica todo lo exigido. Por lo que podemos dar por válida la descomposición realizada.

B. METODOS DE DESCOMPOSICIÓN STL

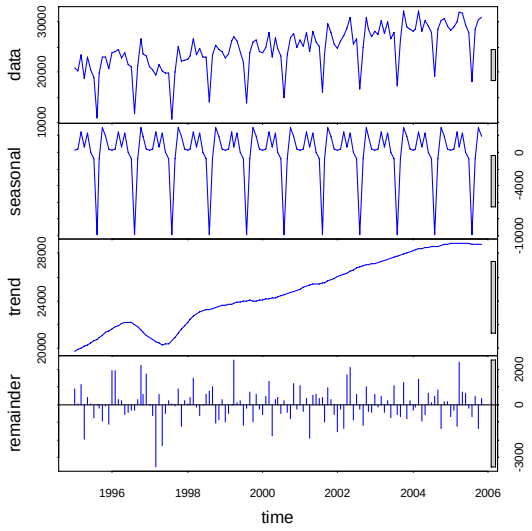
En el R realizaremos el método de descomposición STL, cuya principal herramienta es el suavizado de Loess. Mediante este procedimiento realizaremos una estimación de la tendencia de la misma.

Si c denota el tanto por cierto de los datos que utilizamos para hacer las regresiones locales en el método de suavizado de Loess entonces:

- c pequeña da importancia a los datos de forma individual, la tendencia se ajustará más a la serie original, quedando más patente la componente irregular.

- c es grande el suavizado será mayor, en este caso la tendencia quedará reducida prácticamente a una línea recta y obtenemos menos irregularidades.

Veamos los gráficos de las componentes de la serie tras realizar una descomposición con el método STL para c =1/5



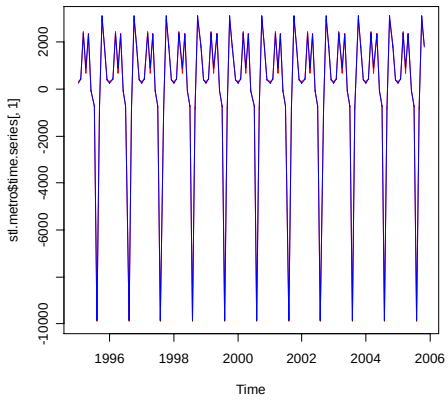
Con esta descomposición podemos hacer los mismos comentarios sobre las componentes tendencia y estacional que realizaremos para la descomposición clásica.

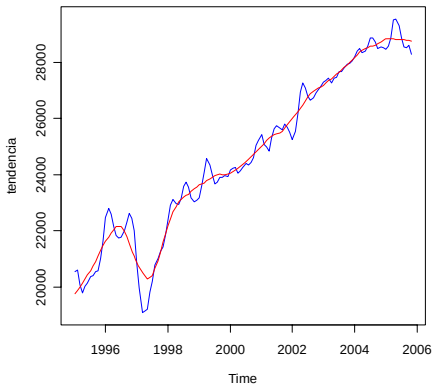
C. COMPARATIVA ENTRE AMBOS MÉTODOS

Los índices estacionales varían levemente entre ambos métodos, expresando la misma idea de cuáles son los meses con mayor número de pasajeros y cuáles los de menor.

periodo	d. clásica	d. stl
1	299.79599	230,444
2	430.03114	440,819
3	2439.35721	2348,703
4	677.95493	883,736
5	2355.37110	2320,086
6	-69.11038	-91,847
7	-719.86416	-720,056
8	-9889.47233	-9844,46
9	-750,99016	-832,456
10	2986.50689	3109,594
11	1857.63941	1783,728
12	382.7804	371,711

Comparemos ya los dos métodos, la línea roja representará a las componentes procedentes del método STL y la azul a las procedentes del método de descomposición clásica.





- La componente ciclo es aparentemente igual, lo vimos en la tabla (sufren variaciones que son muy pequeñas para las magnitudes de los datos)

- Obsérvese que la tendencia en el caso del SPSS es más irregular.

En realidad las dos descomposiciones serían válidas, pero parece que la descomposición STL consigue describir mejor la componente tendencia de la serie.

2.2.- Métodos de suavizado exponencial

Con este tipo de método lo que pretendemos, al igual que con los métodos anteriores, es eliminar las fluctuaciones aleatorias de la serie y aprovechar cualquier conducta "evidente" de la misma con la finalidad de predecir nuevos valores. Estos valores serán medias de valores pasados donde cada dato tendrá un peso diferente, que decrecerá de forma exponencial desde el dato más reciente hasta el más distante.

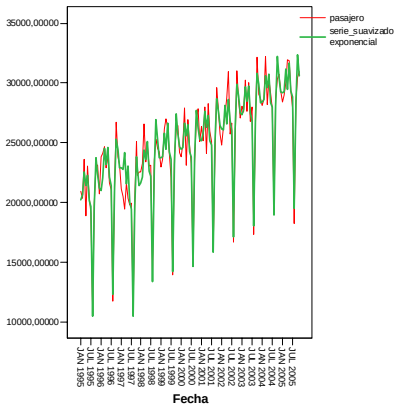
Nuestra serie presenta una tendencia aparentemente lineal, si obviamos el problema de los primeros meses, y además presenta componente estacional. Por ello utilizaremos el método de suavizado exponencial de Holt-Winter para el modelo aditivo.

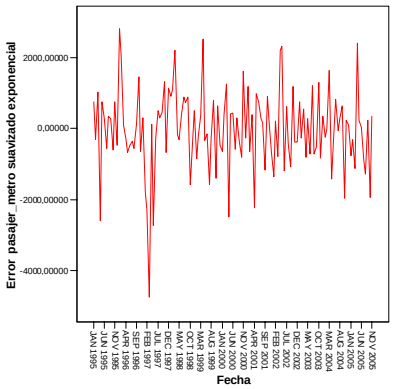
Para realizar este análisis utilizaremos el programa SPSS.

La elección de los parámetros que ajustan al modelo la realizamos con una búsqueda de rejilla de forma que minimicemos el error cuadrático medio.

Aplicamos el modelo personalizado para tendencia lineal y modelo aditivo con búsqueda en rejilla de α, β y δ con pasos de 0.01, 0.01, 0.1 respectivamente y obtenemos los siguientes valores: $\alpha = 0.26$, $\beta = 0$, $\delta = 0$

La serie en color verde representa la serie obtenida tras aplicar el método de suavizado exponencial.





Como se aprecia se ajusta bastante bien a la original salvo en el periodo conflictivo, cosa que se ve claramente en el gráfico de los residuos (figura derecha), ya que en ese periodo se incrementa el error considerablemente.

En el apartado 5 utilizaremos este método para hacer predicciones y compararlas con las que obtendremos a partir de un modelo Arima que se ajuste a nuestros datos. (Veamos ya cuál es el mejor modelo Arima para ellos)

4.1.- Ajuste de un modelo ARIMA

La finalidad de esta sección es encontrar un modelo de la familia ARIMA que se ajuste a los datos para lo cual utilizaremos la metodología de Box-Jenkins.

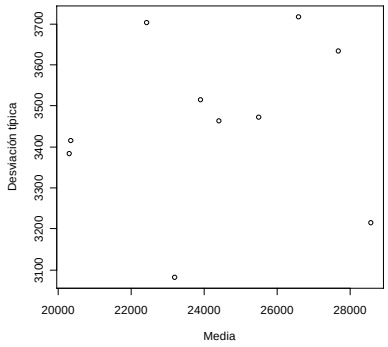
Esta metodología consta de las fases de identificación, estimación y diagnosis. El modelo ajustado nos permitirá hacer posteriormente, en la sección 6, predicciones para la serie estudiada.

4.1.- Identificación

En primer lugar, analizaremos la estacionariedad de la serie. Esta serie es claramente no estacionaria al presentar, al menos, una acusada tendencia y componente estacional.

A simple vista parece que la varianza es constante a lo largo de la serie pero la inestabilidad de la varianza a veces viene enmascarada por la tendencia o por la componente estacional.

Realizamos en el R el gráfico de dispersión-media:

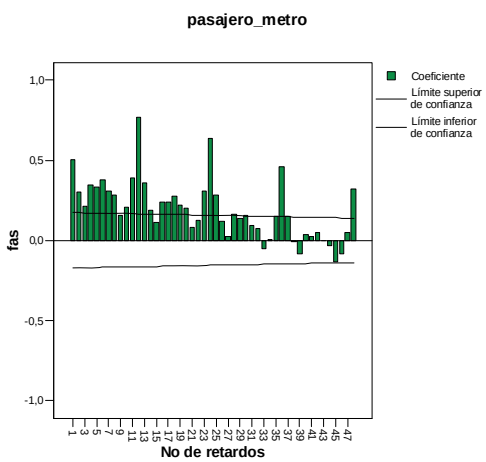


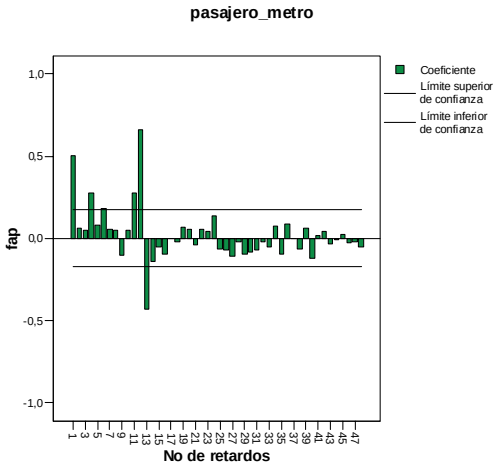
Se ve claramente que la varianza es estable ya que los puntos no siguen ninguna distribución concreta, además la transformación de Box-Cox que se propone es para:

$\lambda = 1 - \alpha = 1 - 0.05545 = 0.94455$

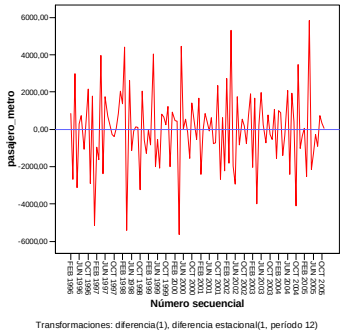
(1 es el λ asociado a la identidad)

Representamos la función de autocorrelación simple (fas) y la función de autocorrelación parcial (fap) de la serie:





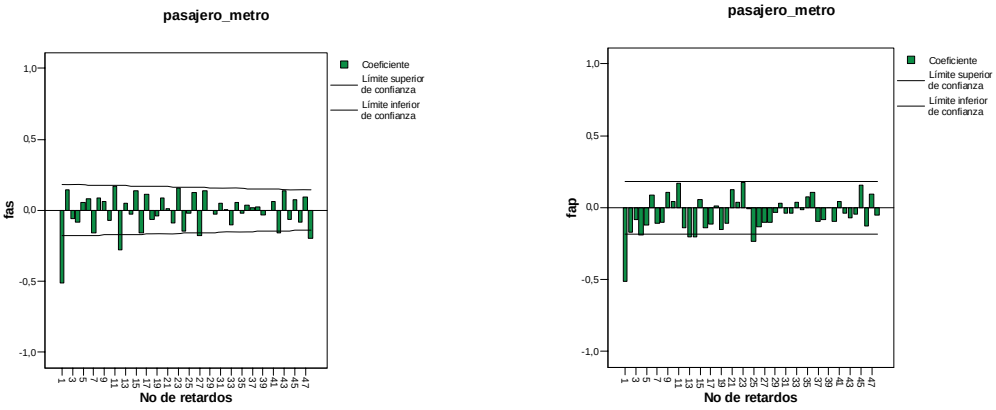
También en la fas se observa la tendencia (decrecimiento lento a 0) y la componente estacional (decrecimiento lento a 0 en los múltiplos de la estacionalidad) Por ello, para obtener una serie estacionaria aplicaremos una diferenciación estacional y una ordinaria a nuestros datos.



Como se aprecia en el gráfico parece que la serie resultante es estacionaria, aunque faltaría comprobar que la varianza es estable. El nuevo gráfico de dispersión-media lo corrobora y en este caso se obtiene un $\alpha = 0.01558$ y por tanto $\lambda = 0.98442 \approx 1$, lo que nos indica que no es necesario ninguna transformación más.

Una vez transformada la serie, de manera que, podamos considerar que procede de un proceso estacionario, identifiquemos a continuación la estructura estacionaria de la serie.

Las principales herramientas para tratar de identificar el modelo que se ajusta a nuestros datos son las funciones de autocorrelación simple (fas) y parcial (fap).

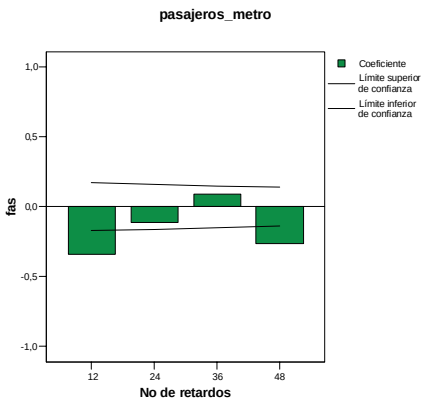


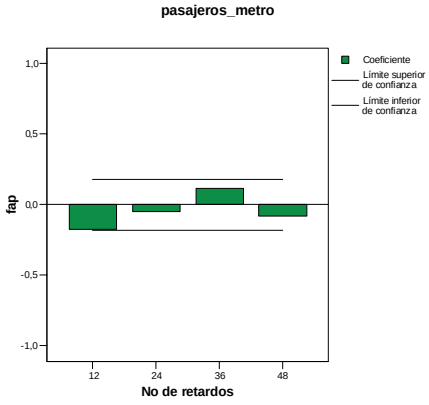
A la hora de identificar la parte ordinaria nos fijamos en los primeros retardos. Los modelos más factibles son:

- MA(1) pues en la fas se aprecia un retardo no significativo y en la fap se observa un decrecimiento lento a 0. (También podíamos considerar MA(2) porque el segundo retardo en la fas esta próximo a las bandas pero de momento no lo haremos pues si fuese necesario el MA(2) en la fase de sobreajuste lo obtendríamos)

- AR(2) pues en la fap se observan dos retardos significativos y la fas decrece de forma sinusoidal a 0.

Para estudiar la parte estacional nos fijamos en los retardos múltiplos de la estacionalidad, es decir en los múltiplos de 12:





En este caso consideraremos:

- MA(1) pues en la fas hay un retardo significativo y en la fap se aprecia un decrecimiento lento a 0 de forma sinusoidal. (Hemos rechazado que el cuarto retardo sea significativo en la fas pues el segundo y el tercero no lo son, así que este hecho lo asociaremos a la aleatoriedad)

- También podríamos pensar en un AR(1) pero parece menos verosímil, ya que el primer retardo de la fas no es claramente significativo.

Por tanto los modelos que consideraremos serán:

ARIMA(0,1,1)xARIMA(0,1,1)12
ARIMA(2,1,0)xARIMA(0,1,1)12
ARIMA(0,1,1)xARIMA(1,1,0)12
ARIMA(2,1,0)xARIMA(1,1,0)12

4.2.- Estimación

En la fase de estimación queremos estimar los parámetros de cada modelo propuesto en la fase de identificación y comprobar si estos modelos son válidos, en el sentido de que los correspondientes coeficientes sean significativamente distintos de cero.

Las estimaciones de los coeficientes de cada modelo son:

ARIMA(0,1,1)xARIMA(0,1,1)12		
	Estimación	p-valor
MA1	0,72203920	0,00000000
SMA1	0,71210510	0,00000005

ARIMA(0.1,1)XARIMA(1,1,0)12		
	Estimación	p-valor
MA1	0,73613990	0,00000000
SAR1	-0.46822242	0,00000012

ARIMA(2,1,0)xARIMA(0,1,1)12		
	Estimación	p-valor
AR1	-0,67098649	0,00000000
AR2	-0,28013101	0,00203223
SMA1	0,68722743	0,00000011

ARIMA(2,1,0)XARIMA(1,1,0)12		
	Estimación	p-valor
AR1	-0,69878296	0,00000000
AR2	-0,27744572	0,00258466
SAR1	-0,45897580	0,00000034

Todos los coeficientes son significativos, con lo que en principio todos los modelos son factibles. Para elegir de estos modelos cuál es el más adecuado utilizaremos el coeficiente de información de Akaike (AIC) y el criterio bayesiano de Schwarz (SBC) El modelo que minimice estas cantidades será considerado el más idóneo.

	AIC	SBC
ARIMA(0,1,1)xARIMA(0,1,1)12	2046,1228	2051,6642
ARIMA(2,1,0)xARIMA(0,1,1)12	2052,1039	2060,416
ARIMA(0,1,1)xARIMA(1,1,0)12	2072,789	2078,3312
ARIMA(2,1,0)xARIMA(1,1,0)12	2078,0163	2086,3283

A la vista del AIC y SBC el modelo con el que nos quedaremos será el ARIMA(0,1,1)xARIMA(0,1,1)12 pues tiene menor AIC y SBC. También se observa que los modelos con parte estacional de medias móviles presentan menor AIC y SBC que los de parte autorregresiva.

4.3.- Diagnosis

Realicemos la fase de diagnosis para el modelo elegido: ARIMA(0,1,1)xARIMA(0,1,1)12

Análisis de los coeficientes estimados

Matriz de correlación

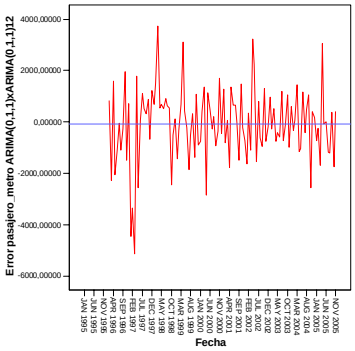
	MA1	SMA 1
MA1	1,0000000	-0,0186764
SMA1	-0,0186764	1,0000000

Los estimadores de los coeficientes del modelo son incorrelados, pues el coeficiente de correlación entre ambos es muy pequeño.

Análisis de los residuos

Queremos comprobar si los residuos representan un proceso de ruido blanco, es decir, tienen varianza constante, media cero y son incorrelados. Además comprobaremos si son normales pues esto nos permitirá obtener bandas de confianza para las predicciones basadas en normalidad.

- Homogeneidad en la varianza de los residuos:

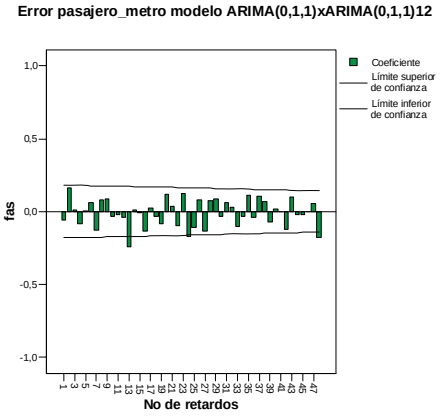


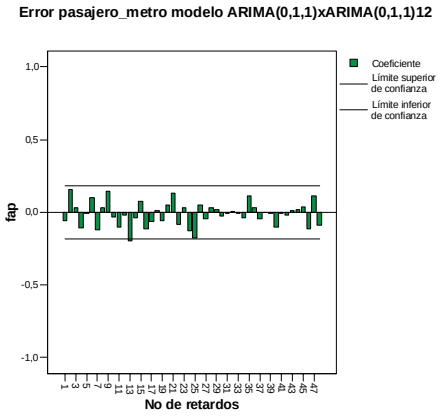
Asumiremos la estabilidad en la varianza aunque parezca que es mayor al principio, pero eso es debido a la irregularidad que se produce a comienzos del año 1997.

- Media nula

La media es -81,9115062 que teniendo en cuenta las magnitudes que utilizamos aunque no sea próxima a 0 como el intervalo de confianza para ella al 95% es (339,7552171 , 175,9322047) que contiene al 0, no rechazaremos que la media sea nula.

- Residuos incorrelados

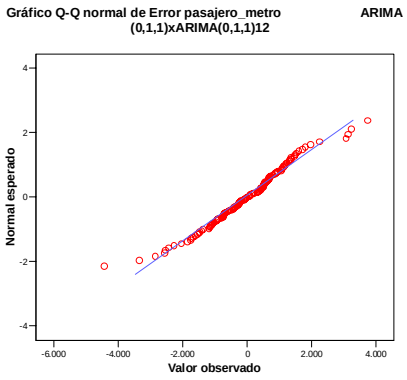




Se observa que los retardos 2, 13, 24 y 48 pueden ser significativos. El test de Box-Ljung resuelve que ningún retardo es significativo, para los retardos problemáticos obtenemos los siguientes p-valores: 0,174 - 0,203 - 0,145 - 0,123. Todos son no significativos, luego podemos considerar que los residuos son incorrelados.

- Normalidad en los residuos

Análisis gráfico:



A la vista del gráfico no podemos rechazar la normalidad, pero tampoco podemos aceptarla pues los datos iniciales y finales se comportan de manera no deseada ya que distan bastante de la recta, y esto es un signo de no normalidad. Para salir de dudas veamos dos test de normalidad.

Pruebas de normalidad:

Obtenemos que el test de Kolmogorov-Smirnov (Lilliefors) es altamente no significativo (p-valor > 0.200), es decir, no podemos rechazar la hipótesis nula, es decir la hipótesis de normalidad. En cambio, el test de Shapiro-Wilk es significativo (p-valor= 0.023) Por tanto, no tenemos asegurada la normalidad de los residuos.

Realicemos ahora el sobreajuste del modelo ARIMA(0,1,1)xARIMA(0,1,1)12. Veamos que cualquier nuevo coeficiente que añadamos al mismo va a ser no significativo, es decir no va a aportar información significativa al modelo debido a problemas de colinealidad. En algunos casos convierte en no significativo coeficientes que antes lo eran .

	coeficiente introducido	significación	coeficiente antiguo	significación
ARIMA(0,1,2)xARIMA(0,1,1)12	MA2	0,49225927	SMA1	0,33337345
ARIMA(1,1,1)xARIMA(0,1,1)12	AR1	0,26503460	SMA1	0,35256640
ARIMA(0,1,1)xARIMA(0,1,2)12	SMA2	0,17543445		
ARIMA(0,1,1)xARIMA(1,1,1)12	SAR1	0,62655001		

Hemos visto que al introducir nuevos parámetros al modelo estos no lo mejoran por lo que nos quedamos con el modelo inicial: ARIMA(0,1,1)XARIMA(0,1,1)12

Luego el modelo Arima del que podemos considerar que nuestra serie es una trayectoria es:

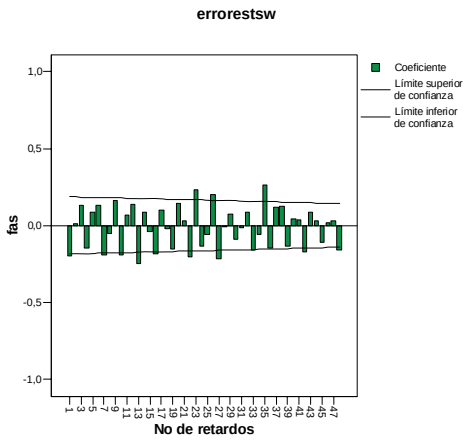
$$(1-B)(1-B_{12})X_t = (1-0{,}72B)(1-0{,}71B_{12})Z_t$$

donde X_t es el valor de la serie en el tiempo t $\{Z_t\}_t$ es un proceso de ruido blanco y B es el operador de retardo ($B_kY_t = Y_{t-k}$ para cualquier serie $\{Y_t\}_t$)

5.- Búsqueda de valores atípicos

La distribución de la serie original, así como la de los residuos (sobre todo con la irregularidad presente en torno al año 1997) nos hace plantearnos la posibilidad de que existan valores atípicos en la serie Para determinar su existencia vamos a utilizar el programa TSW. El análisis de la serie con este software partiendo del modelo ajustado de la sección anterior nos proporciona 5 valores atípicos.

Al analizar los residuos obtenemos que estos son correlados pues tienen p-valores significativos:



1	-->	**0,031**
2	-->	0,097
3	-->	0,086
4	-->	0,056
5	-->	0,071
6	-->	0,058
7	-->	**0,019**
8	-->	**0,030**
9	-->	**0,015**
10	-->	**0,005**
11	-->	**0,007**
12	-->	**0,006**

A la vista de estos resultados debemos volver a la fase de identificación. Habíamos identificado cuatro modelos como plausibles y habíamos decidido que los que tenían parte estacional de medias móviles eran mejores y tomamos el de menor AIC, ARIMA(0,1,1)xARIMA(0,1,1)12. Veamos si con el otro modelo de los posibles, ARIMA(2,1,0)xARIMA(0,1,1)12, no tenemos problemas al identificar los valores atípicos, es decir si podemos considerar que tiene los residuos incorrelados e inclusos normales.

En el nuevo modelo aparecen cinco valores atípicos dos de ellos aditivos, dos cambios transitorios y un cambio de nivel. Tal como habíamos observado al principio hay dos atípicos en los meses de enero y marzo de 1997 (cambio de nivel y atípico aditivo, respectivamente) Además aparecen otro atípico aditivo en abril de 1999 y cambios transitorios en los meses de abril de 2002 y abril del 2005.

Las estimaciones que obtenemos en el TSW para nuestro modelo tras localizar los atípicos son:

Variables en el modelo:

	Estimación	Error Estándar	¿significativa?
AR1	0.89118	0.08058	0.89118 > 2 x 0.08058 = 0.16118
AR2	.51285	0.080082	0.51285 > 2 x 0.080082 = 0.160164
SMA1	-0.83258	0.12009	0.83258 > 2 x 0.12009 = 0.24018

Atípicos:

		Estimación	Error Estándar	¿significativa?
LS	1-1997	-3157.0	690.67378	3157.0 > 2 x 690.67378 =131.1344756
AO	3-1997	-3105.2	824.87702	3105.2 > 2 x 824.87702 =1649.75404
AO	4-1999	4271.3	782.56649	271.3 > 2 x 782.56649 =1565.13298
TC	4-2002	3739.7	696.22749	3739.7 > 2 x 696.22749 =1392.45498
TC	4-2005	3623.1	719.31348	3623.1 > 2 x 719.31348 =1438.62696

Por tanto, las variables y los atípicos son significativos, no deberíamos eliminar ninguno de ellos. Además en este modelo disminuye el AIC hasta 1976.48 (sin atípicos era 2052,1039 y para el modelo ARIMA(0,1,1)xARIMA(0,1,1)$_{12}$ era 2046,12).

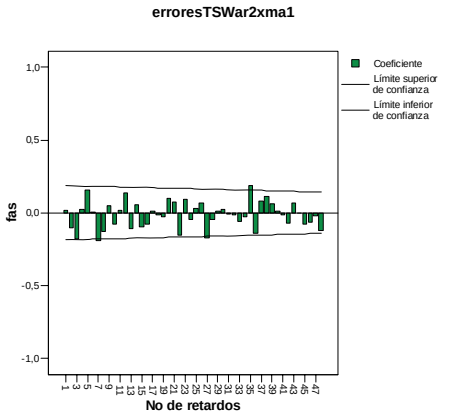
En cuanto a los atípicos obtenidos cabe destacar los correspondientes a 1997 que se pueden justificar pues se realizó durante esos meses una huelga de personal, lo que disminuyó considerablemente el número de pasajeros pues se cerraron algunos servicios. (La huelga era para pedir un incremento en el sueldo a percibir tras la jubilación).

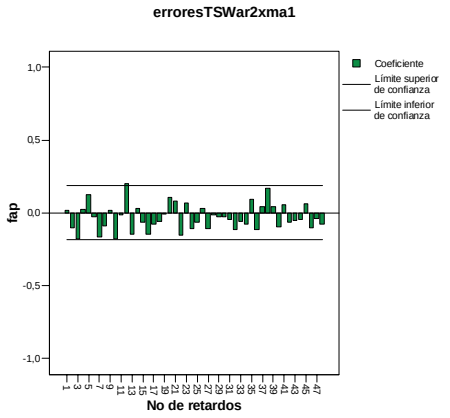
Diagnosis ARIMA(2,1,0)xARIMA(0,1,1)12

Análisis de los residuos obtenidos tras búsqueda de atípicos

Queremos comprobar si los residuos que se obtienen tras la búsqueda de atípicos son un proceso de ruido blanco. Veamos que en este caso no se produce lo mismo que en el modelo anterior que al introducir los atípicos los residuos se volvían correlados.

- Residuos incorrelados





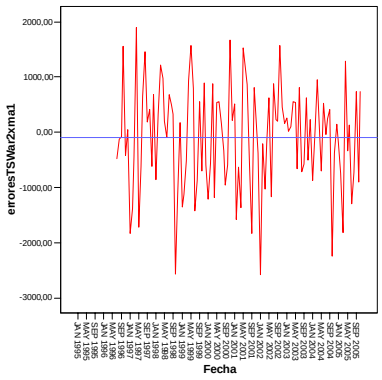
Se observan algunos retardos en el límite, podrían ser significativos, pero el test de Box-Ljung nos permite concluir que son incorrelados (pues todos tienen p-valores mayores que 0.05)

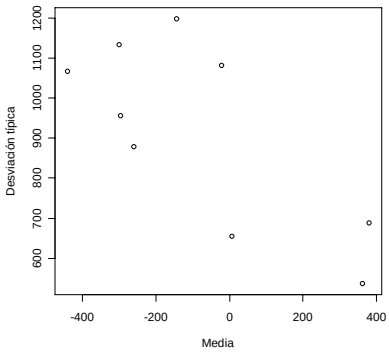
Comprobemos que los residuos cumplen el resto de condiciones para que el modelo sea válido.

- Media nula

La media es -88,6252 pero aunque no es próxima a 0 el intervalo de confianza para ella al 95% es (-263,8923 , 86,6418) que contiene al 0, por lo que no podremos rechazar que la media es nula.

- Homogeneidad en la varianza de los residuos:

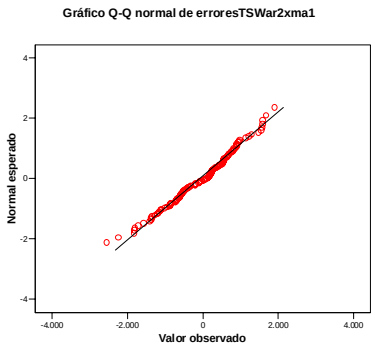




Asumiremos la estabilidad en la varianza. Pues aunque en el gráfico de dispersión media parece que la varianza disminuye conforme aumenta la media, obtenemos un $\lambda = 1 + 0.0176$ que es próximo a 1.

- Normalidad en los residuos

<u>Análisis gráfico:</u>



A la vista del gráfico no podemos rechazar la normalidad, pero tampoco podemos aceptarla al cien por cien, para salir de dudas veamos los test de normalidad.

<u>Pruebas de normalidad:</u>

Obtenemos que tanto el test de Kolmogorov-Smirnov (Lilliefors) como el de Shapiro-Wilk son no significativos con probabilidades de significación 0,083 y 0,261 respectivamente, luego no podemos rechazar la hipótesis nula de normalidad.

6.- Predicción

Poder hacer predicciones es la finalidad que tiene el estudio de series temporales, sacar conclusiones de los valores pasados para poder predecir, en base de los modelos obtenidos, el comportamiento futuro.

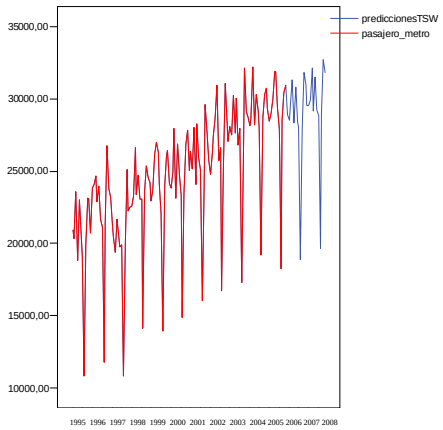
Podemos hacer predicciones bien con el método de suavizado exponencial o a partir del modelo ARIMA que hemos ajustado tras aplicar metodología de Box-Jenkins. Veamos cual es mejor y haremos predicciones para el próximo año con el que sea más adecuado.

Veremos cual de los dos tiene menor error cuadrático medio (ECM) en el periodo enero 2004 - noviembre 2005 y este será el elegido a la hora de hacer predicciones

ECM(arima) 770.829,7
ECM(suav.exp) 1.048.621

Por tanto es más aconsejable utilizar el modelo ARIMA que hemos ajustado para obtener predicciones futuras de la serie.

Veamos cuales son las predicciones para el año 2006 a partir de julio y hasta noviembre del año 2007.



mes-año	predicción
7-2006	27985,24
8-2006	18851,71
9-2006	28109,15
10-2006	31847,59
11-2006	30969,08
12-2006	29509,63
1-2007	29521,08
2-2007	29897,51
3-2007	32157,46
4-2007	29188,45
5-2007	31527,26
6-2007	29248,52
7-2007	28816,27
8-2007	19656,13
9-2007	28970,5
10-2007	32680,12
11-2007	31803,9

7.- Conclusión final

En conclusión, aplicando distintas metodologías de series temporales, hemos encontrado un modelo que explica el comportamiento de nuestra serie y con el que podemos hacer predicciones. Debemos notar que aunque el modelo inicial ARIMA(0,1,1)xARIMA(0,1,1)$_{12}$ parecía mejor, la existencia de valores atípicos nos ha obligado a cambiar a otro modelo que se ajusta más a los requisitos teóricos y a la propia serie. Por tanto, consideramos que el modelo que mejor se adapta a nuestros datos será: ARIMA(2,1,0)xARIMA(0,1,1)$_{12}$, junto con cinco atípicos.

El modelo final será:

$$X_t = \frac{(1+0{,}83B)}{(1-0{,}89B-0{,}51B^2)(1-B)(1-B^{12})} Z_t +$$

$$-3157{,}0S_t^{(25)} - 3105{,}2I_t^{(27)} + 4271{,}3I_t^{(29)} + \frac{3739{,}7}{(1-0{,}7B)} I_t^{(88)} + \frac{3623{,}1}{(1-0{,}7B)} I_t^{(124)}$$

Bibliografía

- Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994) Times Series Analysis: Forecasting and Control. Prentice Hall.

- Cleveland,R. B. and Cleveland, W. S. and McRae,J. E. and Terpenning, (1990) STL: A seaonal--trend descomposition procedure based on Loess. Journal Official Statistics, volumen ="6",páginas="3-73"}

- Gómez, V. y Maravall, A. (1996) Programs TRAMO (Time Series Regression with Arima noise, Missing observations and Outliers) and SEATS (Signal Extraction in Arima Time Series). Instructions for the user. Documento de trabajo 9628, Servicio de Estudios, Banco de España.

- R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.