Análisis de Series Temporales

T3

Serie 20 - Pernoctaciones Zamora

Pablo de Arriba Mendizábal
Pablo Martín De Benito

15 de Enero de 2024

OBJETIVO DE LA PRÁCTICA

Para la serie asignada se pide ajustar modelos de Box-Jenkins y analizarlos para después escoger el mejor candidato posible que ajuste a los datos.

En la primera página debe especificarse el nombre y apellidos del alumno/a (o alumnos), el nombre y número de la serie (según la numeración de la hoja en la que la profesora indicará la serie asignada) y el modelo final elegido para la serie.

También se incluirá en esa página la ACF de los residuales de dicho modelo y los contrastes de Portmanteau asociados, así como la expresión de Xt en función de los retardos correspondientes de Xt y at , donde Xt representa a la serie y at a los residuales del modelo elegido. Esta última expresión debe ser lo más explícita posible, con los valores numéricos resultado de la estimación y sin factores producto. En particular, la constante del modelo, si se necesita, debe incluirse con el valor numérico explícito resultado de la estimación.

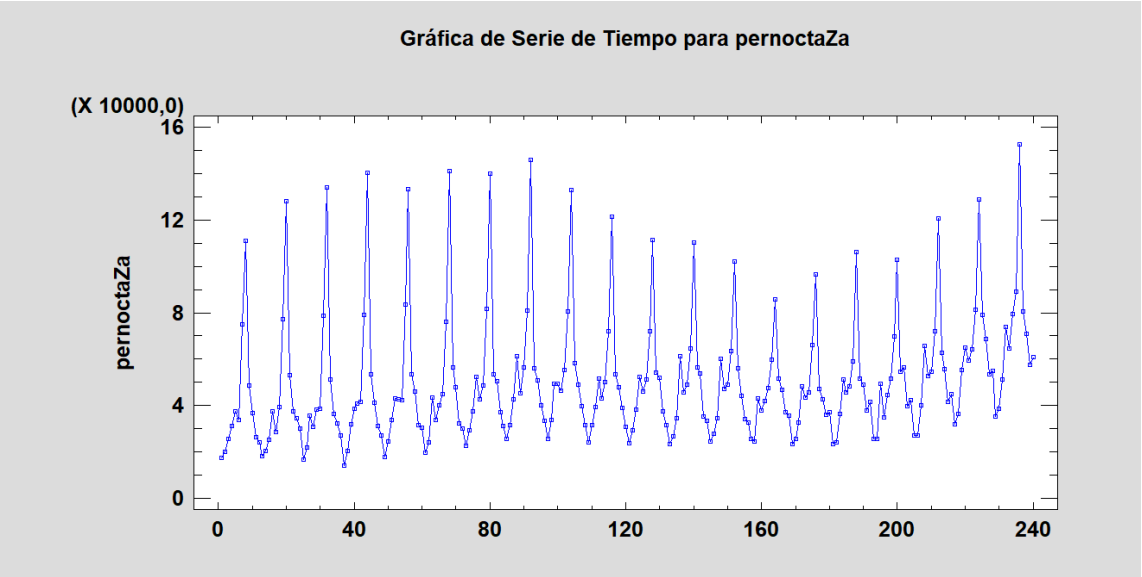
NOTAS:

1. Si hay un modelo que ajusta a la serie con un orden de diferenciación menor (d ó D) que el que se ha elegido, debe indicarse en la primera página y debe explicarse por qué no se ha elegido.
2. Si el ajuste se ha efectuado con datos transformados los epígrafes 3 y 4 (estudio de la capacidad de predicción y predicción) deben llevarse a cabo con la serie original, es decir, deshaciendo la transformación.

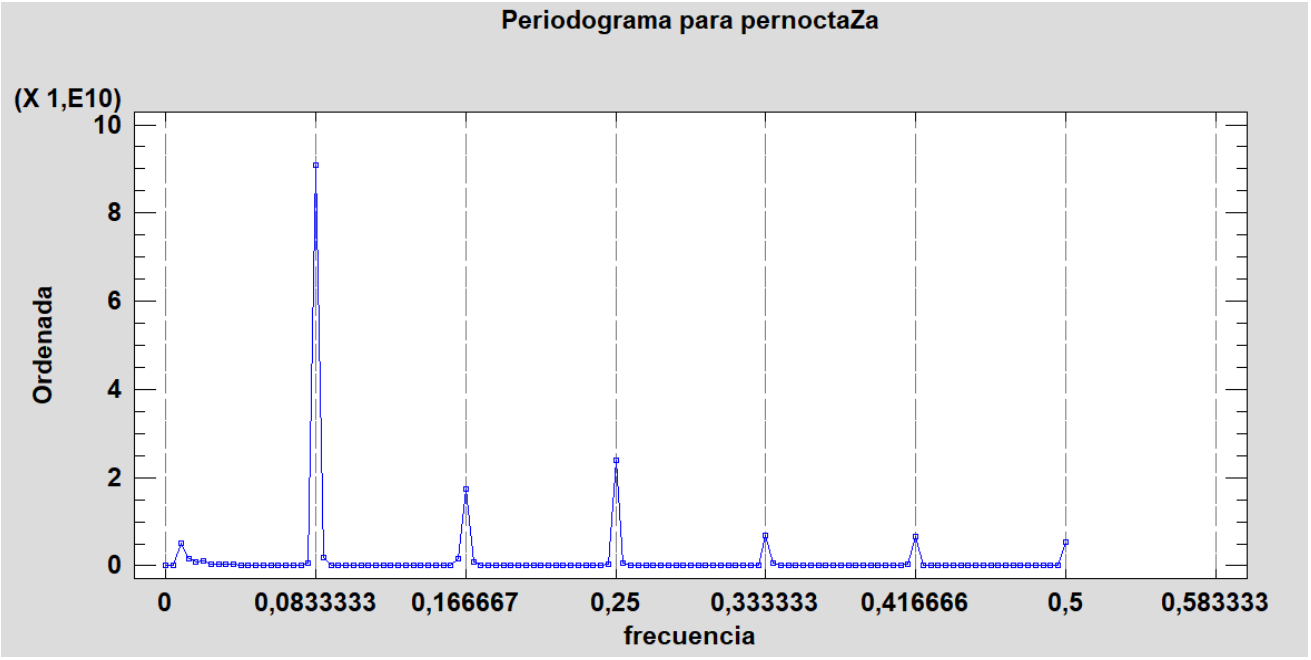
IDENTIFICACIÓN

La serie original que se nos proporciona recoge un total de 240 datos mensuales del número de pernoctaciones en alojamientos de la provincia de Zamora entre enero de 2000 hasta diciembre de 2019.

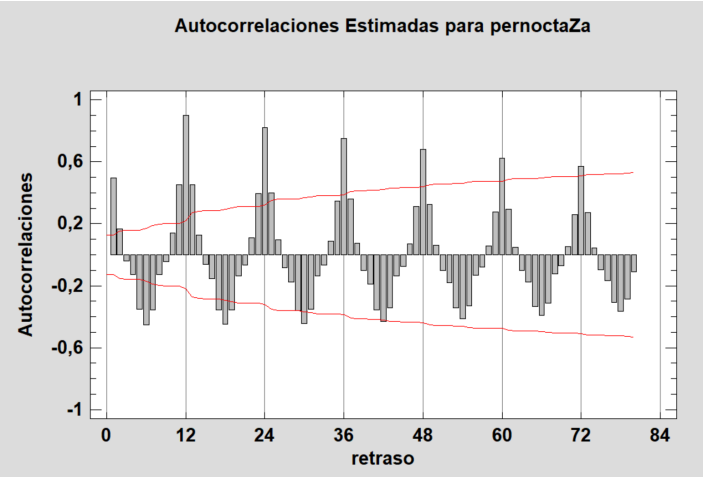
Lo primero que debemos hacer es describir la serie original dada, para ello utilizamos *Statgraphics* y la columna de datos proporcionada *pernoctaZa*.

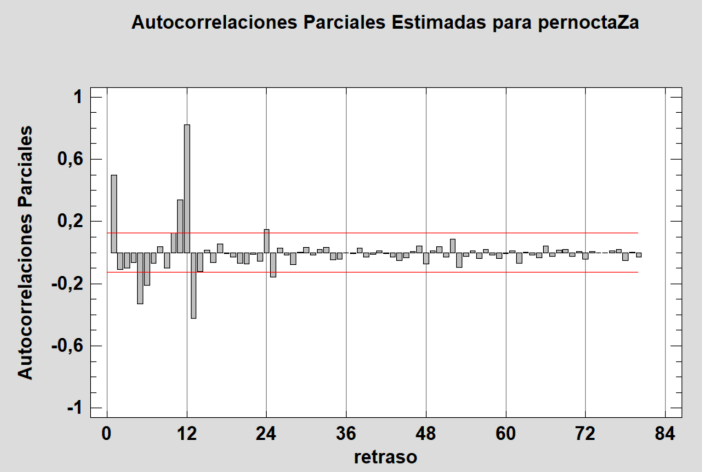


Observando el gráfico de la serie lo primero que nos damos cuenta es la aparición de ciclos que pueden estar adheridos a tendencia, estos ciclos podemos ver que tienen un periodo de doce meses, lo que concluye que la serie tiene una estacionalidad de periodo *s=12*.
Además, podemos destacar una varianza estable a lo largo de la serie, lo que hace que no necesitemos transformar los datos.



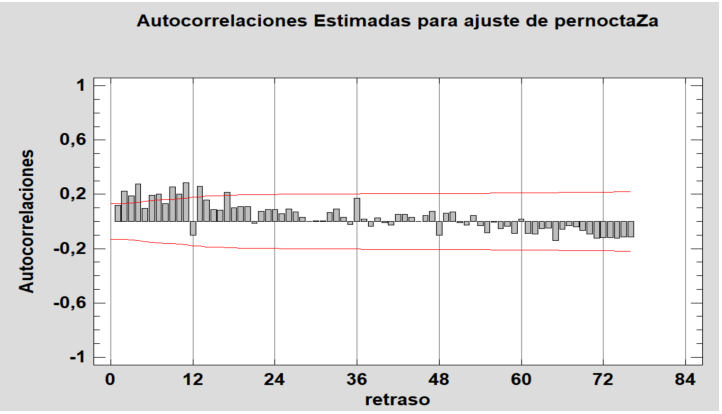
A la vista del periodograma, vemos que no existe una tendencia y vemos los máximos locales en las frecuencias 1/12, 2/12… que confirman la estacionalidad de periodo doce.

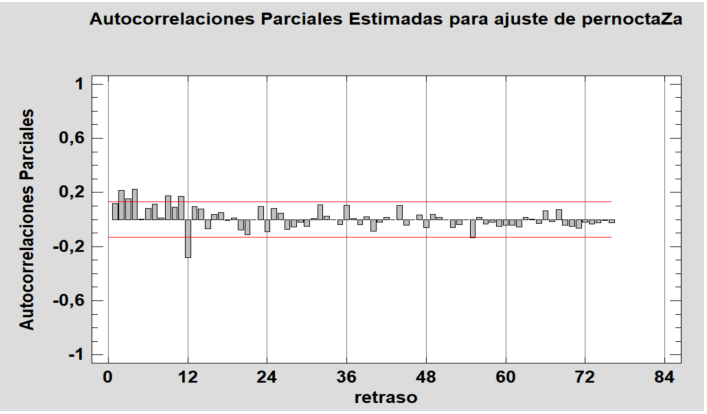




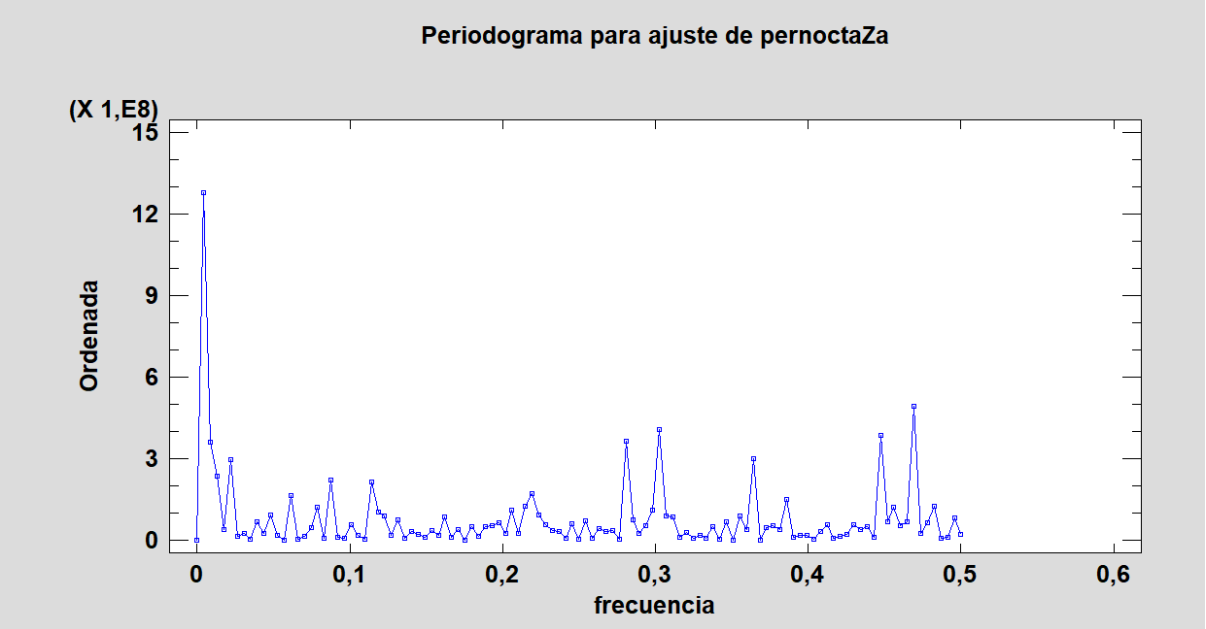
Analizando la ACF, vemos que la serie no es estacionaria en un primer momento, pues destacamos correlaciones muy notables y no se aprecia un decrecimiento exponencial.
Además, en la PACF aunque sí observamos un decrecimiento exponencial a partir del retraso 12, destacamos una gran cantidad de autocorrelaciones significativas.

Dicho esto, procederemos a probar varias diferenciaciones.
Como hemos notado una estacionalidad en la serie, diferenciamos una vez la parte estacional, arrojando los siguientes resultados:



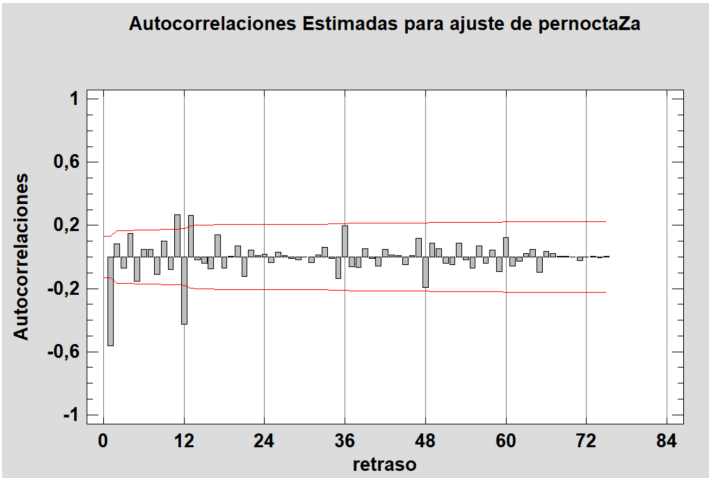


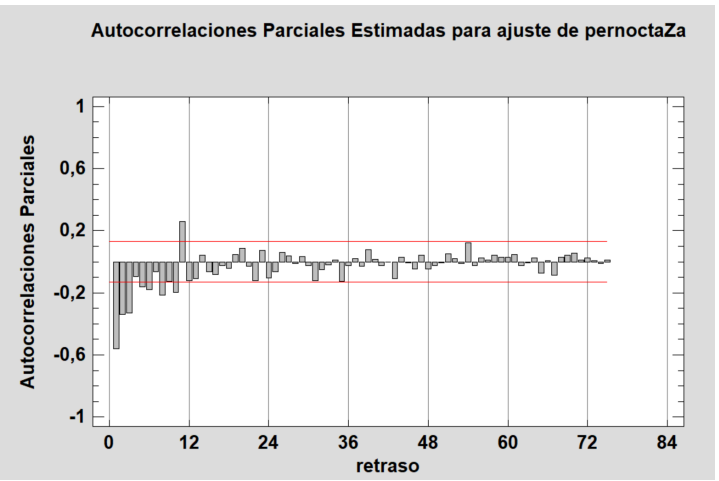
Podemos concluir que la ACF no nos da ninguna información destacable, no existe un decrecimiento exponencial, además de que salen bastantes correlaciones de la banda.
Tampoco podemos sacar nada en claro de la PACF pues obtenemos grandes primeras autocorrelaciones aunque luego sí parece decrecer, aunque no exponencialmente.



Además vemos como, diferenciando una vez la parte estacional, aparece una tendencia bastante grande, lo que rechazaría completamente esta diferenciación.

Así pues, comprobamos la serie diferenciando una vez en la parte regular y una vez en la parte estacional *(d=1 D=1)*.





Como podemos observar ahora vemos un decrecimiento exponencial por ciclos de la ACF, que aunque salen de la banda las correlaciones 1, 11, 12 y 13; podríamos decir que esta serie sí es estacionaria.

Para mayor claridad, observemos las varianzas de los datos ajustados de cada serie diferenciada.

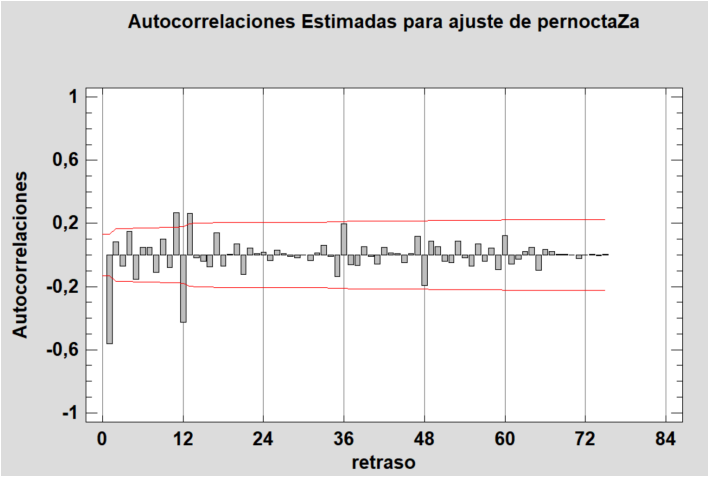
Resumen Estadístico

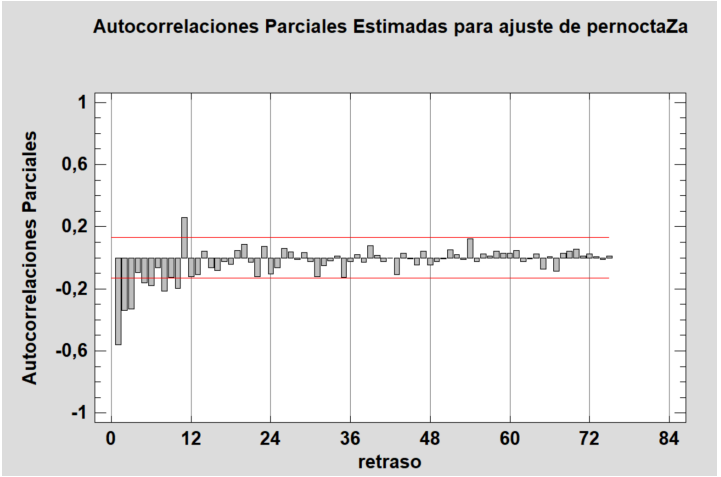
	SerieOriginal	Serie1ParteRegular	Serie1ParteEstacional	Serie1RegularYEstacional
Recuento	240	239	228	227
Promedio	50731,5	181,866	1606,81	22,326
Desviación Estándar	26879,9	26895,5	6318,94	8410,75
Coeficiente de Variación	52,9846%	14788,6%	393,26%	37672,5%
Mínimo	14390,0	-89817,0	-22180,0	-30677,0
Máximo	152287,	64966,0	30954,0	40364,0
Rango	137897,	154783,	53134,0	71041,0
Sesgo Estandarizado	11,0115	-5,02441	2,15997	0,688996
Curtosis Estandarizada	9,95769	7,77107	10,2844	10,4645

El resumen estadístico que vemos en la tabla se ha realizado mediante un análisis de las cuatro columnas de datos obtenidos de la serie original.
Cabe mencionar que además hemos incluido el análisis de la serie diferenciada una vez por la parte regular aunque no la hayamos analizado a fondo porque estaba claro que en un primer momento no aportaría información.

Como podemos observar en la fila de la *Desviación Estándar*, la menor varianza nos la da la serie diferenciada una vez en la parte estacional, pero como hemos visto anteriormente, esta dejaba ver una tendencia muy notable que rechazaría esa opción.
De esta forma, la mejor opción que obtendríamos es la serie diferenciada regular y estacionalmente una vez, tal y como habíamos concluido antes.

Una vez tenemos la serie estacionaria, habría que decidir los modelos candidatos que pensemos que puedan ajustar mejor, para ello analizamos la ACF y la PACF.





En un primer vistazo, podríamos sacar los siguientes posibles modelos:

➔ $SARIMA\ (1,1,1)(1,1,2)_{12}$

Si suponemos un decrecimiento exponencial en la parte regular y estacional en ambas ACF y PACF, entonces estaremos hablando de un AR(1), MA(1) AR(12), MA(12); además podemos suponer que en la ACF la parte estacional tiene dos autocorrelaciones sueltas, por lo que añadiremos también un coeficiente MA(24), quedando como modelo (1,1,1)(1,1,2).

➔ $SARIMA\ (3,1,1)(1,1,2)_{12}$

Si interpretamos que en la parte regular, la PACF muestra que hay tres correlaciones sueltas entonces tendríamos hasta un AR(3) en la parte regular quedando un (3,1,1)(1,1,2).

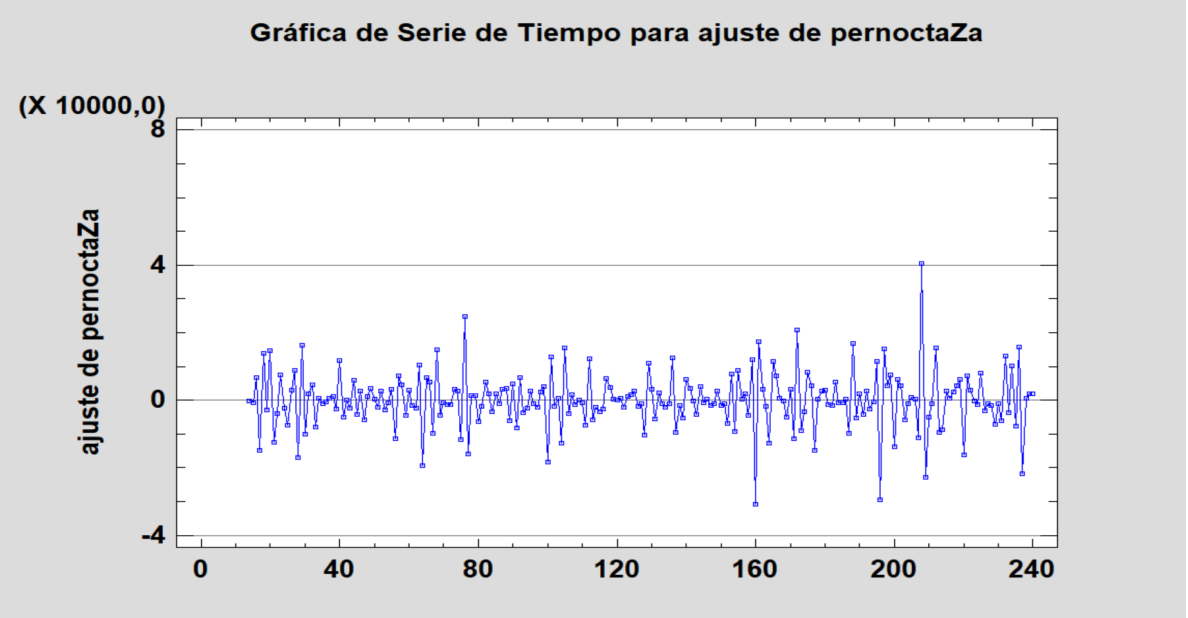
➔ $SARIMA\ (1,1,1)(0,1,2)_{12}$

Podemos suponer que no hay parte autoregresiva en la parte estacional ya que podemos pensar que no decrece exponencialmente la parte estacional.

➔ $SARIMA\ (0,1,1)(1,1,2)_{12}$

También podemos suponer que no existe parte autoregresiva en la parte regular pues no decrece exponencialmente a largo plazo.

Estos modelos, son suposiciones a la vista de los gráficos anteriores, no obstante, pueden surgir más modelos según los resultados de los coeficientes



Del gráfico de la serie, podemos darnos cuenta de que los datos están centrados por lo que podemos quizá considerar modelos sin la constante, probaremos los descritos previamente pero sin constante.

ESTIMACIÓN Y VALIDACIÓN

Para la etapa de estimación y validación de los modelos, vamos a construir una tabla con todos los parámetros y estimadores de todos los modelos que suponemos que pueden ajustar nuestra serie.

Para ello ejecutamos el siguiente código en SAS y recogemos los datos relevantes con respecto a tus coeficientes.

```
/* … Tratar la columna de datos */

proc arima data=pernoctaza plots(unpack)=all;
       identify var=pernoctaza(1,12);
        run;
       *estimate q=(1)(12,24) p=(1)(12) method=ml;          /* (1,1,1)(1,1,2) */
       *estimate q=(1)(12,24) p=(1) method=ml;              /* (1,1,1)(0,1,2) */
       *estimate q=(1)(12,24) p=(1,2,3)(12) method=ml;      /* (3,1,1)(1,1,2) */
       *estimate q=(1)(12,24) p=(1,2,3) method=ml;          /* (3,1,1)(0,1,2) */
       *estimate q=(1)(12,24) p=(12) method=ml;                    /*
(0,1,1)(1,1,2) */
       *estimate q=(1)(12,24) method=ml;                    /* (0,1,1)(0,1,2) */
       *estimate q=(1)(12) p=(1)(12) method=ml;             /* (1,1,1)(1,1,1) */
       *estimate q=(1)(12) p=(1) method=ml;                 /* (1,1,1)(0,1,1) */
       *estimate q=(1)(12) p=(12) method=ml;                /* (0,1,1)(1,1,1) */
       *estimate q=(1)(12)  method=ml;                      /* (0,1,1)(0,1,1) */
        forecast out=b lead=12 id=date interval=month;
run;

proc arima data=pernoctaza plots(unpack)=all;
       identify var=pernoctaza(1,12);
         run;
       *estimate q=(1)(12,24) p=(1)(12) method=ml noconstant; /* (1,1,1)(1,1,2) */
       *estimate q=(1)(12,24) p=(12) method=ml noconstatn;   /* (0,1,1)(1,1,2) */
       *estimate q=(1)(12,24) method=ml noconstant;          /* (0,1,1)(0,1,2) */
       *estimate q=(1)(12) p=(12) method=ml noconstant;      /* (0,1,1)(1,1,1) */
       *estimate q=(1)(12)  method=ml noconstant;            /* (0,1,1)(0,1,1) */
       forecast out=b lead=12 id=date interval=month;
       *estimate q=(12,24,36)  method=ml noconstant;         /* (0,1,0)(0,1,3) */
        *forecast out=m15 lead=24 id=date interval=month;
       *estimate q=(12,24)  method=ml noconstant;            /* (0,1,0)(0,1,2) */
       *forecast out=m15 lead=24 id=date interval=month;
       *estimate p=(1,2,3) q=(12,24)  method=ml noconstant;  /* (3,1,0)(0,1,2) */
        *forecast out=m13 lead=24 id=date interval=month;
run;
```

	Modelo	**Parámetros**	**Estimadores**	**Error estándar**	**T-valor**	**P-valor**
1	$SARIMA\ (1,1,1)(1,1,2)_{12}$ con costante	Constante MA(1) MA(12) MA(24) AR(1) AR(12)	33.75960 0.86160 0.89250 -0.42610 -0.06827 0.45167	44.41258 0.03993 0.20239 0.08667 0.07555 0.21612	0.76 21.58 4.41 -4.92 -0.90 2.09	0.4472 <.0001 <.0001 <.0001 0.3661 0.0366
2	$SARIMA\ (1,1,1)(0,1,2)_{12}$ con costante	Constante MA(1) MA(12) MA(24) AR(1)	29.27118 0.85716 0.44573 -0.20719 -0.06253	37.97333 0.04049 0.06902 0.07085 0.07608	0.77 21.17 6.46 -2.92 -0.82	0.4408 <.0001 <.0001 0.0034 0.4112
3	$SARIMA\ (3,1,1)(1,1,2)_{12}$ con costante	Constante MA(1) MA(12) MA(24) AR(1) AR(2) AR(3) AR(12)	25.97227 0.85737 -0.39729 0.27002 -0.06889 0.0083803 -0.05485 -0.74225	32.39328 0.05340 2.48258 0.83686 0.08410 0.08218 0.07801 2.49087	0.80 16.06 -0.16 0.32 -0.82 0.10 -0.70 -0.30	0.4227 <.0001 0.8729 0.7470 0.4127 0.9188 0.4820 0.7657
4	$SARIMA\ (0,1,1)(1,1,2)_{12}$ con costante	Constante MA(1) MA(12) MA(24) AR(12)	33.88191 0.87493 0.89654 -0.42800 0.44459	42.33139 0.03372 0.20195 0.08756 0.21628	0.80 25.94 4.44 -4.89 2.06	0.4235 <.0001 <.0001 <.0001 0.0398
5	$SARIMA\ (0,1,1)(0,1,2)_{12}$ con costante	Constante MA(1) MA(12) MA(24)	29.26367 0.86964 0.45608 -0.20762	36.44081 0.03424 0.06888 0.07056	0.80 25.40 6.62 -2.94	0.4219 <.0001 <.0001 0.0033
6	$SARIMA\ (1,1,1)(0,1,1)_{12}$ con costante	Constante MA(1) MA(12) AR(1)	25.83557 0.86556 0.34660 -0.06345	32.07777 0.03915 0.06888 0.07547	0.81 22.11 5.03 -0.84	0.4206 <.0001 <.0001 0.4005
7	$SARIMA\ (0,1,1)(1,1,1)_{12}$ con costante	Constante MA(1) MA(12) AR(12)	27.05384 0.87274 0.17282 -0.23499	32.72726 0.03406 0.17941 0.17485	0.83 25.62 0.96 -1.34	0.4084 <.0001 0.3354 0.1790
8	$SARIMA\ (0,1,1)(0,1,1)_{12}$ con costante	Constante MA(1) MA(12)	25.74395 0.87635 0.35767	31.01507 0.03373 0.06849	0.83 25.98 5.22	0.4065 <.0001 <.0001

A la vista de estos resultados, nos podemos dar cuenta de que todos los modelos en los que incluimos la parte autoregresiva en la parte regular no es significativa en ninguno, por lo que debemos suprimirla, rechazando por lo tanto, todas las opciones de posibles AR(1), incluso AR(3) que habíamos supuesto analizando la ACF y PACF.

Además como hemos mencionado antes, vemos que, efectivamente, aunque existen buenos modelos, la constante en todos ellos no ajusta bien. Por ello, probamos los siguientes modelos sin constante.

Construimos los modelos sin constante:

	Modelo	**Parámetros**	**Estimadores**	**Error estándar**	**T-valor**	**P-valor**
9	$SARIMA\ (1,1,1)(1,1,2)_{12}$	MA(1) MA(12) MA(24) AR(1) AR(12)	0.85729 0.89038 -0.42605 -0.07074 0.45224	0.04051 0.20159 0.08622 0.07566 0.21525	21.17 4.42 -4.94 -0.93 2.10	<.0001 <.0001 <.0001 0.3498 0.0356
10	$SARIMA\ (0,1,1)(1,1,2)_{12}$	MA(1) MA(12) MA(24) AR(12)	0.87107 0.89370 -0.42731 0.44418	0.03411 0.20198 0.08739 0.21623	25.54 4.42 -4.89 2.05	<.0001 <.0001 <.0001 0.0400
11	$SARIMA\ (0,1,1)(0,1,2)_{12}$	MA(1) MA(12) MA(24)	0.86593 0.45445 -0.20866	0.03462 0.06869 0.07045	25.01 6.62 -2.96	<.0001 <.0001 0.0031
12	$SARIMA\ (0,1,1)(1,1,1)_{12}$	MA(1) MA(12) AR(12)	0.86903 0.17017 -0.23584	0.03445 0.18004 0.17524	25.22 0.95 -1.35	<.0001 0.3446 0.1784
13	$SARIMA\ (0,1,1)(0,1,1)_{12}$	MA(1) MA(12)	0.87270 0.35598	0.03411 0.06845	25.59 5.20	<.0001 <.0001
14	$SARIMA\ (0,1,0)(0,1,3)_{12}$	MA(12) MA(24) MA(36)	0.54011 -0.15596 -0.18216	0.06936 0.07830 0.07423	7.79 -1.99 -2.45	<.0001 0.0464 0.0141
15	$SARIMA\ (0,1,0)(0,1,2)_{12}$	MA(12) MA(24)	0.58327 -0.23105	0.06730 0.06914	8.67 -3.34	<.0001 0.0008
16	$SARIMA\ (3,1,0)(0,1,2)_{12}$	MA(12) MA(24) AR(1) AR(2) AR(3)	0.44292 -0.23645 -0.78649 -0.50702 -0.28510	0.06861 0.07097 0.06446 0.07581 0.06437	6.46 -3.33 -12.20 -6.69 -4.43	<.0001 0.0009 <.0001 <.0001 <.0001

Una vez obtenidos los resultados de los posibles modelos, debemos elegir los modelos que tengan los estimadores de los parámetros significativos, estos son los modelos que elegimos

- $SARIMA\ (0,1,1)(1,1,2)_{12}$ Modelo 10
- $SARIMA\ (0,1,1)(0,1,2)_{12}$ Modelo 11
- $SARIMA\ (0,1,1)(0,1,1)_{12}$ Modelo 13
- $SARIMA\ (0,1,0)(0,1,2)_{12}$ Modelo 15
- $SARIMA\ (3,1,0)(0,1,2)_{12}$ Modelo 16

No obstante, de dichos modelos elegido podemos observar como el modelo (10) su matriz de correlaciones existe algún parámetro con alta correlación con otro lo que haría no elegirlo.

Correlaciones de las estimaciones de parámetro				
Parámetro	MA1,1	MA2,1	MA2,2	AR1,1
MA1,1	1.000	0.074	-0.013	0.132
MA2,1	0.074	1.000	-0.746	0.948
MA2,2	-0.013	-0.746	1.000	-0.618
AR1,1	0.132	0.948	-0.618	1.000

Los demás modelos no habría problemas con la matriz de correlaciones

Modelo 11

Correlaciones de las estimaciones de parámetro			
Parámetro	MA1,1	MA2,1	MA2,2
MA1,1	1.000	-0.186	0.053
MA2,1	-0.186	1.000	-0.362
MA2,2	0.053	-0.362	1.000

Modelo 13

Correlaciones de las estimaciones de parámetro		
Parámetro	MA1,1	MA2,1
MA1,1	1.000	-0.216
MA2,1	-0.216	1.000

Modelo 14

Correlaciones de las estimaciones de parámetro			
Parámetro	MA1,1	MA1,2	MA1,3
MA1,1	1.000	-0.524	0.239
MA1,2	-0.524	1.000	-0.457
MA1,3	0.239	-0.457	1.000

Modelo 15

Correlaciones de las estimaciones de parámetro		
Parámetro	MA1,1	MA1,2
MA1,1	1.000	-0.452
MA1,2	-0.452	1.000

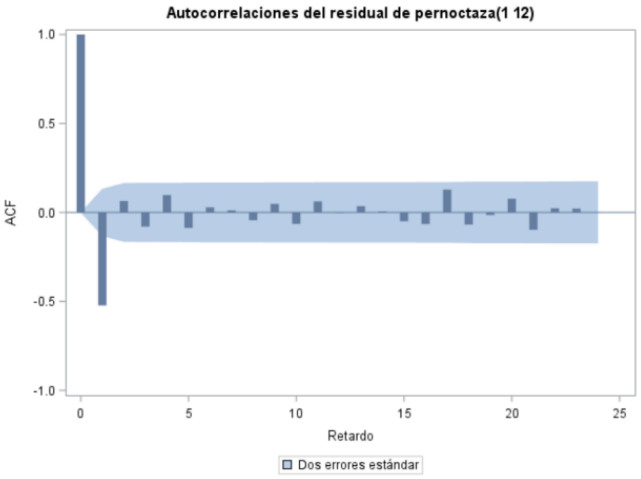
Modelo 16

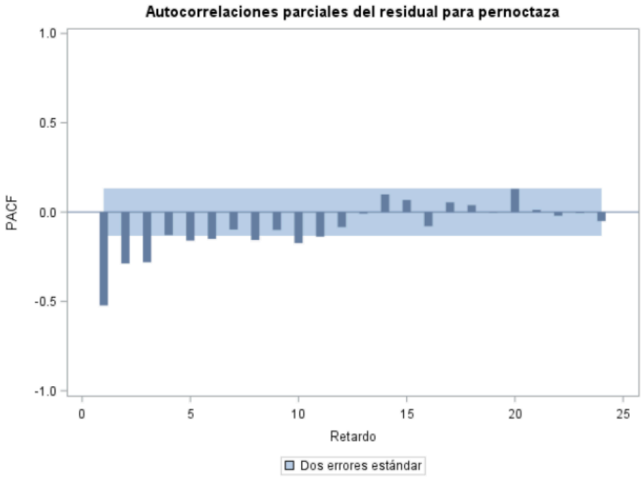
Correlaciones de las estimaciones de parámetro					
Parámetro	MA1,1	MA1,2	AR1,1	AR1,2	AR1,3
MA1,1	1.000	-0.358	0.142	0.104	0.099
MA1,2	-0.358	1.000	0.002	0.005	-0.069
AR1,1	0.142	0.002	1.000	0.599	0.315
AR1,2	0.104	0.005	0.599	1.000	0.595
AR1,3	0.099	-0.069	0.315	0.595	1.000

Además podemos ver como el modelo (14) (15) son similares solamente añadiendo un parámetro, por lo que vamos a quedarnos con el (15) al tener menor número de parámetros entre esos dos.

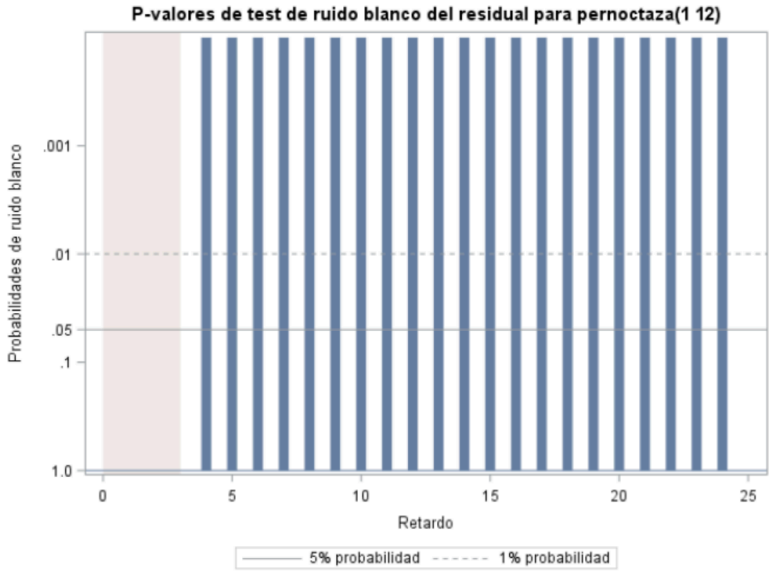
Una vez escogidos los mejores modelos estimados, vamos a analizar sus residuos correspondientes.
Para ello, el proc ARIMA de SAS, nos permite hacer un análisis a fondo que nos permitirá validar los modelos previamente seleccionados.

➔ Validación del Modelo 15 - $SARIMA\ (0,1,0)(0,1,2)_{12}$

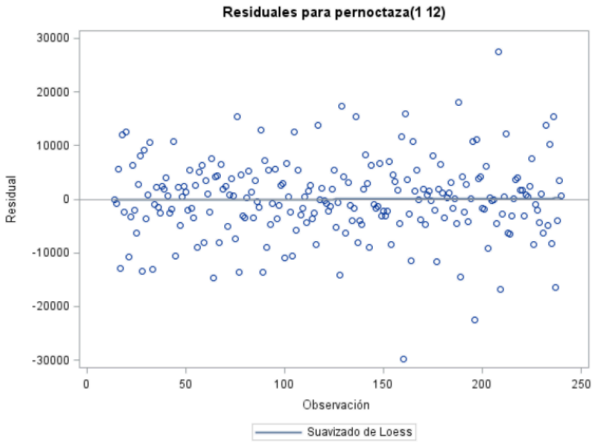




Analizamos la serie de los residuos y como podemos ver en la ACF y PACF, existen retardos que salen de la banda en ambas ACF y PACF y además de los primeros retardos los cuales están mejor estimados, lo que significaría que la serie de residuos no sería estacionaria y por tanto ruido blanco.



En cuanto a los tests de Ljung-Box, podemos observar en el gráfico cómo la probabilidad de ser ruido blanco en cada uno de los retardos es muy baja (barras altas) por lo que este modelo no podría ser válido.

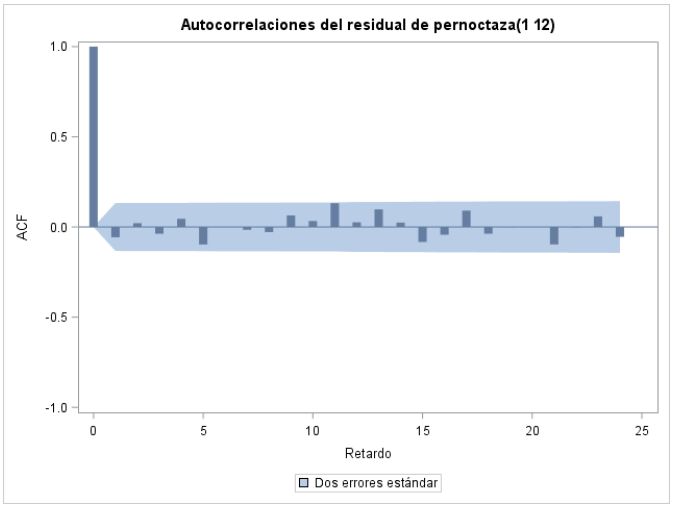


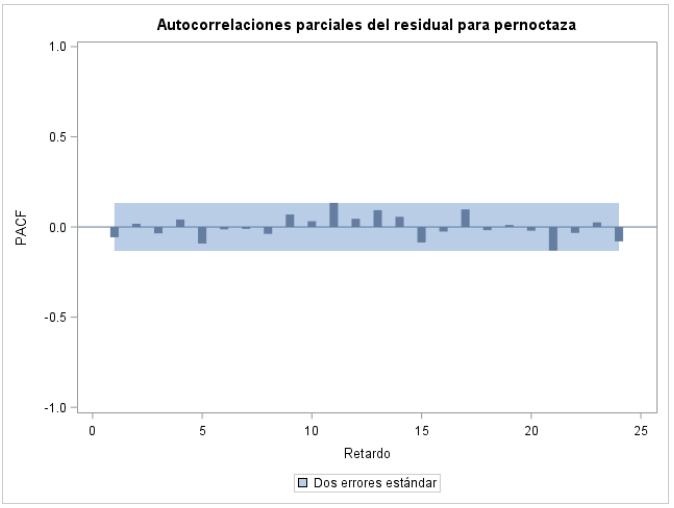
No obstante, observamos cómo la varianza de los residuos es homogénea a lo largo de las observaciones al no destacar ningún patrón visible en el gráfico. También podemos suponer media cero a la vista del gráfico.

SSE = 12114135084
AIC = 4688.474

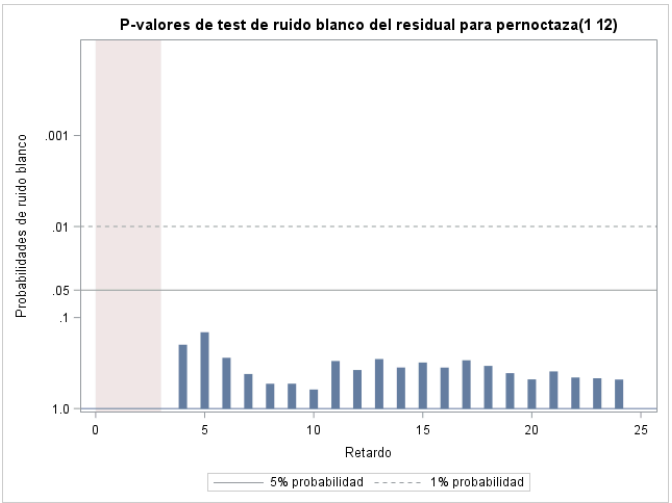
A la vista de estos resultados, podemos descartar este modelo debido a los problemas de heterocedasticidad que presentan los residuos.

➔ Validación Modelo 11 - $SARIMA\ (0,1,1)(0,1,2)_{12}$

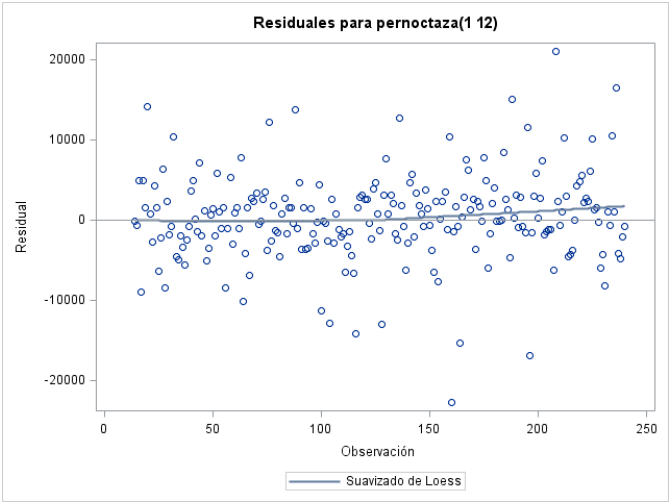




Obtenemos unos ACF y PACF casi idénticos al anterior modelo. En este caso las autocorrelaciones que antes se quedaban al límite de las bandas están incluso más cerca, aunque es casi inapreciable esta diferencia. Las primeras autocorrelaciones siguen siendo significativamente pequeñas, por lo que es aceptable considerarlo un ruido blanco.



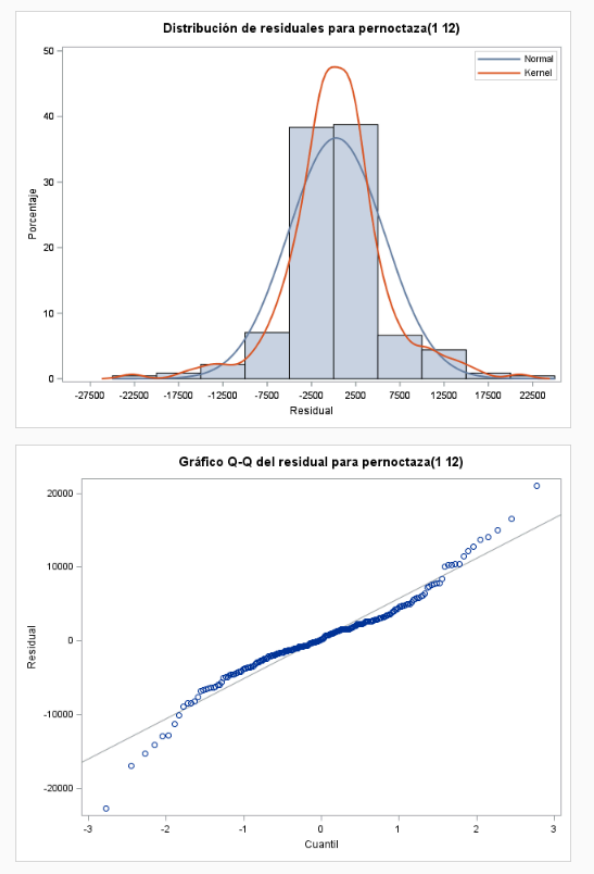
Al realizar los tests de Ljung-Box, se observa fácilmente que todos los p-valores son superiores a 0.05, lo cual indica que la probabilidad de ruido blanco es alta y no se puede rechazar esta hipótesis. Por este motivo consideramos los residuales un ruido blanco.



Respecto a la homogeneidad de varianzas, se puede confirmar al no observarse ningún patrón de megáfono ni similares. Todos los residuales están situados de forma aleatoria en torno al 0.

Tests para posición: Mu0=0				
Test	Estadístico		P valor	
T de Student	t	0.841996	Pr > |t|	0.4007
Signo	M	2.5	Pr >= |M|	0.7907
Rango con signo	S	981	Pr >= |S|	0.3231

El test de media de los residuos nos deja suponer media cero.



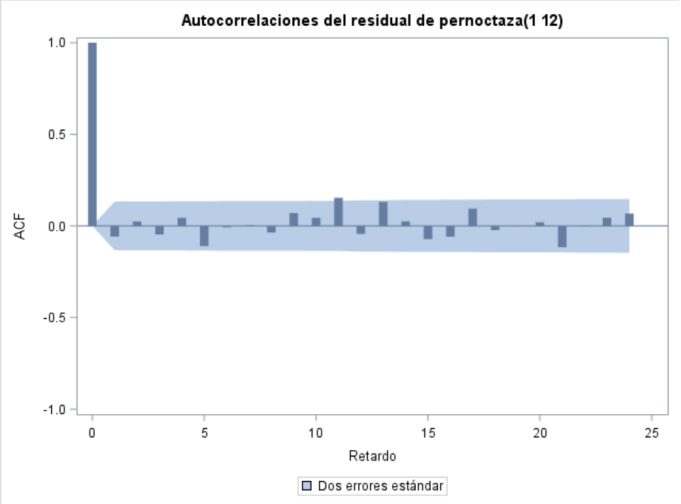
Para verificar la normalidad comprobamos el histograma y el Q-Q plot de los residuales. A la vista del histograma podría parecer que siguen una distribución normal, pero el Q-Q plot muestra que esto no es así ya que los residuales no aproximan bien a la recta.

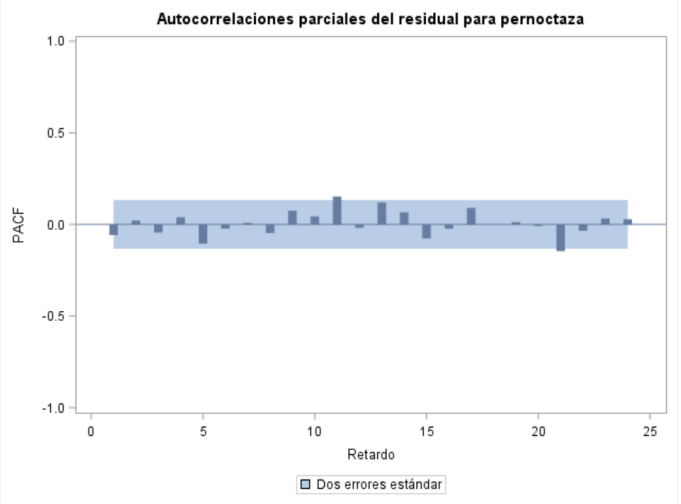
Test para normalidad				
Test	Estadístico		P valor	
Shapiro-Wilk	W	0.952962	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.095365	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.521871	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.118663	Pr > A-Sq	<0.0050

El test de Shapiro-Wilk es recomendable para muestras de tamaño < 50. En este caso tenemos 240 observaciones, por lo que es preferible fijarse en otros tests como el de Kolmogorov-Smirnov.
Como vemos en la tabla, los test de normalidad se rechazan por lo que los datos no siguen una distribución normal.
No obstante y aunque no es demasiado importante la normalidad, no podemos fiarnos de las bandas de confianza.

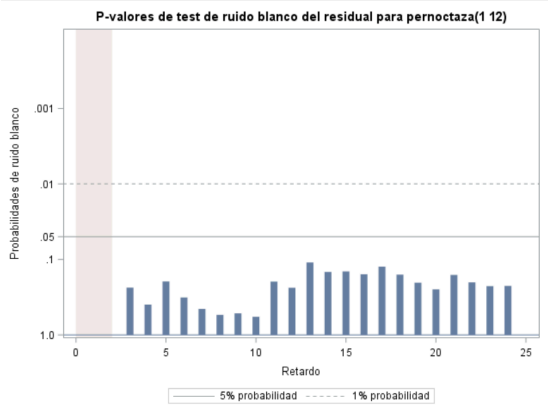
SSE = 6824594060.2
AIC = 4558.715

➔ Validación Modelo 13 - $SARIMA\ (0,1,1)(0,1,1)_{12}$

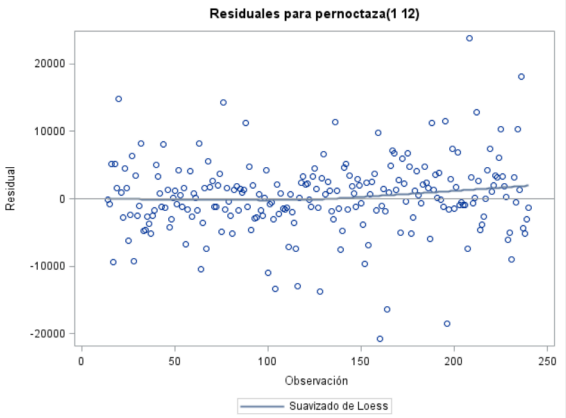




Al analizar los gráficos ACF y PACF podemos darnos cuenta como existe una autocorrelación significativa que sale de la banda, además de otra que se encuentra al borde de la significación.

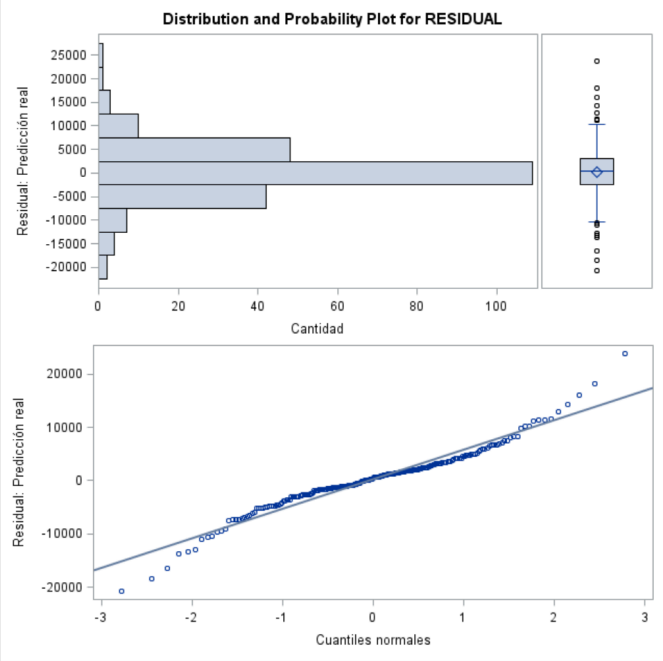


No obstante, los test de Ljung-Box nos arrojan unos muy buenos resultados en los primeros retardos que en realidad son los más importantes.
Que aunque en los retardos más altos, nos indica una probabilidad de ruido blanco más baja, no se acerca demasiado a la significación puedo aceptarlo.



En cuanto a la varianza se refiere, el gráfico de residuales nos indica que podemos suponer homogeneidad a lo largo de las observaciones al no destacar ningún patrón visible.
También podemos suponer una media cero a la vista del gráfico y los tests.

Tests para posición: Mu0=0				
Test	Estadístico		P valor	
T de Student	t	0.825002	Pr > |t|	0.4102
Signo	M	8.5	Pr >= |M|	0.2882
Rango con signo	S	1121	Pr >= |S|	0.2587

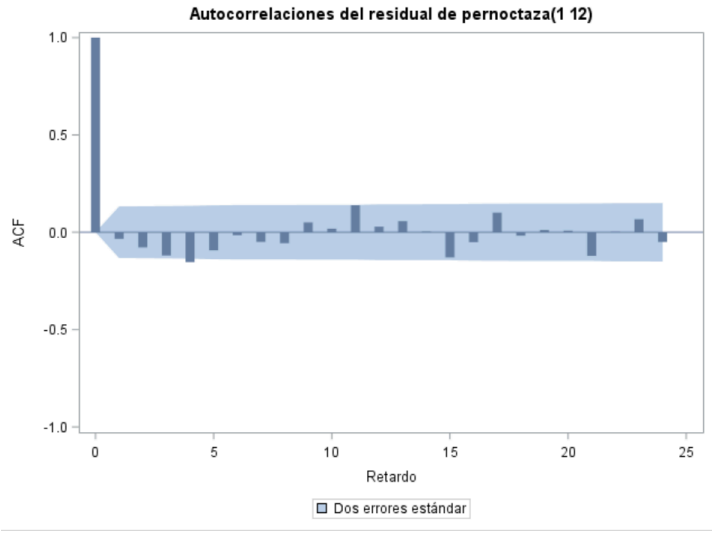


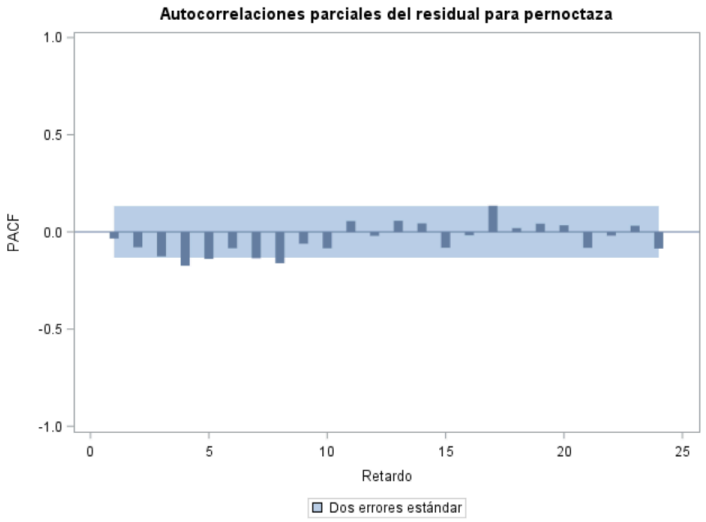
Test para normalidad				
Test	Estadístico		P valor	
Shapiro-Wilk	W	0.95201	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.091217	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.521928	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.063276	Pr > A-Sq	<0.0050

Podemos rechazar que los residuos son normales a través de los gráficos histograma y Q-Q Plot viendo la forma de estos, no asemejándose a un comportamiento normal y corroborándolo con los resultados de los tests de Shapiro-Wilk y K.S.

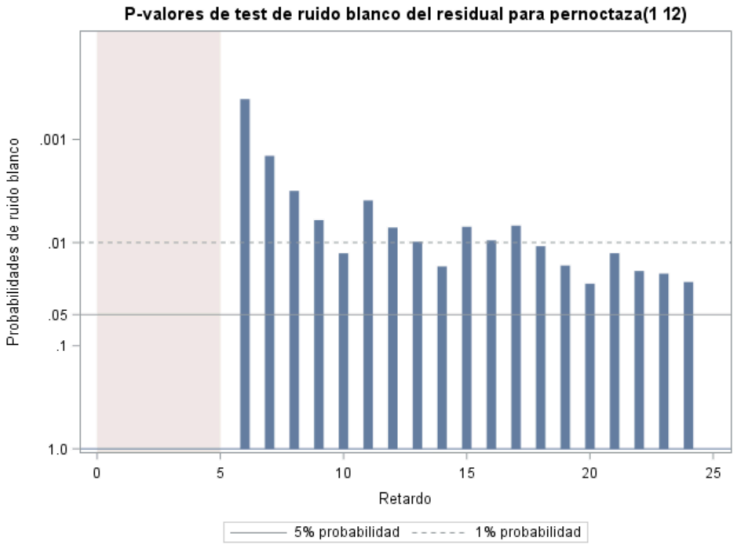
SSE = 6996739944.4
AIC = 4562.616

➔ Validación Modelo 16 - $SARIMA\ (3,1,0)(0,1,2)_{12}$

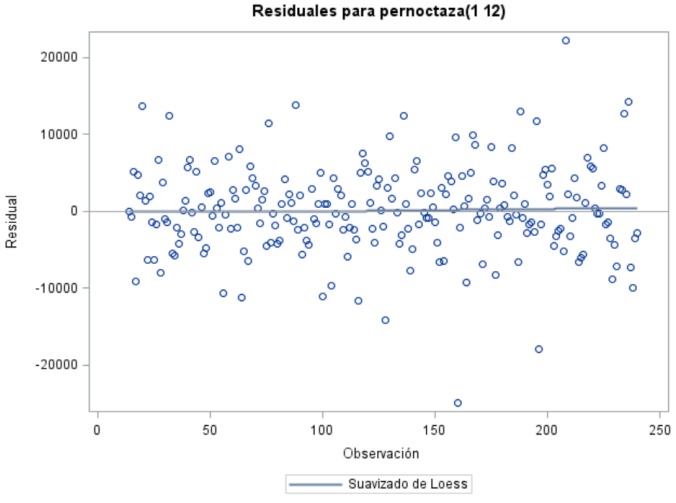




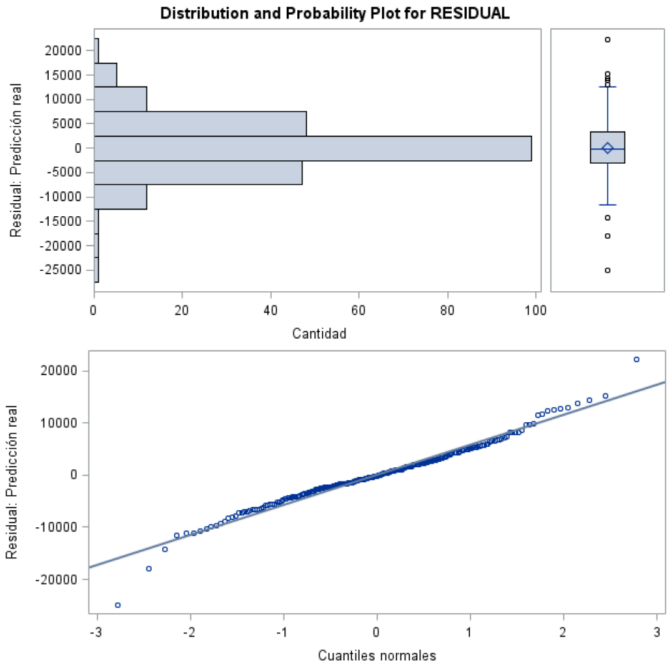
En primer lugar observamos la serie de residuales y aunque la mayoría de autocorrelaciones se encuentran dentro de la banda, sí es verdad que los primeros retardos, que son los mejor estimados, salen fuera de la banda como por ejemplo el 4. Pasa lo mismo en la PACF, aunque no es un problema mayor y podría validarse.



Al observar los test de Ljung-Box obtenemos malos resultados, y es que como nos muestra el gráfico, la probabilidad de que la serie de residuales sea ruido blanco es muy baja, sobre todo en los primeros retardos, lo que puede ser un motivo para descartar el modelo.



No destacamos ningún patrón en los residuos por lo que decimos que son homogéneos.



Podemos suponer normalidad a la vista de los gráficos, que nos pueden indicar residuos normales y para ratificarlo, vemos como el test Kolmogorov-Smirnos, que es mucho más convincente con nuestros datos, lo aceptan.

Test para normalidad				
Test	Estadístico		P valor	
Shapiro-Wilk	W	0.97377	Pr < W	0.0003
Kolmogorov-Smirnov	D	0.056752	Pr > D	0.0745
Cramer-von Mises	W-Sq	0.205546	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.274046	Pr > A-Sq	<0.0050

SSE = 7445558128.8
AIC = 4583.021

Aunque los tests de Ljung-Box no nos han dado unos resultado buenos, podemos validar los residuos, dejando pasar ese defecto, para comparar con los demás modelos válidos.

	$SARIMA\ (0,1,1)(0,1,2)_{12}$	$SARIMA\ (0,1,1)(0,1,1)_{12}$	$SARIMA\ (0,1,0)(0,1,2)_{12}$	$SARIMA\ (3,1,0)(0,1,2)_{12}$
SSE	6824594060.2	6996739944.4	12114135084	7445558128.8
AIC	4558.715	4562.616	4688.474	4583.021

Trás realizar los análisis de los residuos de los modelos, podemos concluir con tres modelos válidos para la siguiente fase de comparación de modelos.
Estos son

- **(11)** $SARIMA\ (0,1,1)(0,1,2)_{12}$
- **(13)** $SARIMA\ (0,1,1)(0,1,1)_{12}$
- **(16)** $SARIMA\ (3,1,0)(0,1,2)_{12}$

COMPARACIÓN DE MODELOS

Para esta etapa, vamos a volver a realizar la estimación sin utilizar los últimos 24 datos y construir una predicción 24 lugares hacia adelante y calcular con ellos los errores de predicción y suma de cuadrados (SSEp) para comparar los tres modelos de nuevo.

Escogemos un *k=24* porque debemos escoger ciclos enteros para hacer backcasting, así que decidimos coger dos ciclos de periodo 12.

Para realizar esto en nuestro programa SAS tenemos que omitir estos *k=24* datos del data set y con ellos realizar la estimación de los modelos, una vez estimados los modelos, realizamos las predicción *k=24* datos hacia adelante.
Para ello, ejecutamos la siguiente porción de código.

```
* Comprobar la capacidad de predicción;
data b; set pernoctaza; if _N_ < 217;
run;

proc arima data=b plots(unpack)=all ;
       identify var=pernoctaza(1,12);
        run;
       estimate q=(1)(12,24) method=ml noconstant;          /* 11-(0,1,1)(0,1,2) */
       forecast out=m11b lead=24 id=date interval=month;
       estimate q=(1)(12)  method=ml noconstant;            /* 13-(0,1,1)(0,1,1) */
       forecast out=m13b lead=24 id=date interval=month;
       estimate p=(1,2,3) q=(12,24)  method=ml noconstant;  /* 16-(3,1,0)(0,1,2) */
       forecast out=m16b lead=24 id=date interval=month;
run;
quit;

data w; set pernoctaza; if _N_ > 216;
run;
data m11w; set m11b; keep FORECAST; if _N_>216; rename forecast=forecnm;
run;
data m13w; set m13b; keep FORECAST; if _N_>216; rename forecast=forecnm;
run;
data m16w; set m16b; keep FORECAST; if _N_>216; rename forecast=forecnm;
run;

* para modelo 11;
data r11; merge w m11w; residuals = pernoctaza - forecnm;
run;
data x;
set r11;
sum2+RESIDUALs*RESIDUALs;
var = sum2/_N_;
RUN;
quit;
```

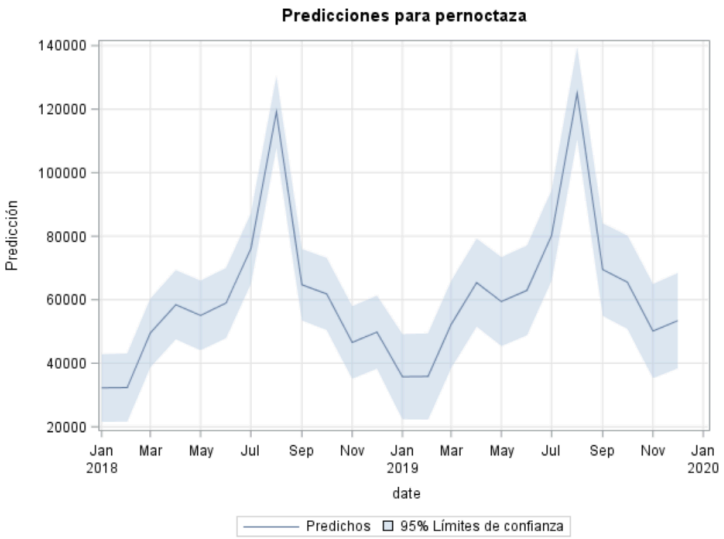
```
* para modelo 13;
data r13; merge w m13w; residuals = pernoctaza - forecnm;
run;
data x;
set r13;
sum2+RESIDUALs*RESIDUALs;
var = sum2/_N_;
RUN;
quit;

* para modelo 16;
data r16; merge w m16w; residuals = pernoctaza - forecnm;
run;
data x;
set r16;
sum2+RESIDUALs*RESIDUALs;
var = sum2/_N_;
RUN;
quit;
```

Obtenemos los siguientes resultados agrupados en una tabla:

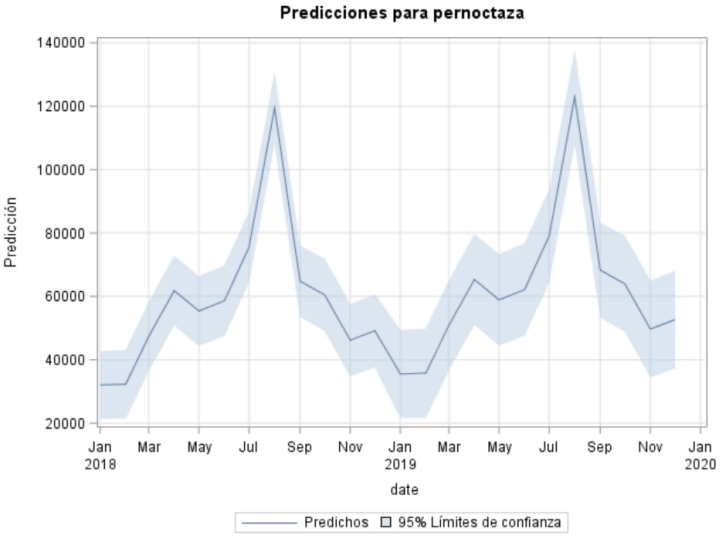
	11 - $SARIMA\ (0,1,1)(0,1,2)_{12}$	**13** - $SARIMA\ (0,1,1)(0,1,1)_{12}$	**16** - $SARIMA\ (3,1,0)(0,1,2)_{12}$
SSE	6824594060.2	6996739944.4	7445558128.8
SSEp	2049966553.9	2319605659.4	3047822549.6
AIC	4558.715	4562.616	4583.021

Parece que el ***modelo 11*** muestra tener los mejores valores en cuanto SSE, capacidad de predicción y AIC con respecto a los otros dos.
Observemos las bandas de predicción de cada uno de los modelos.

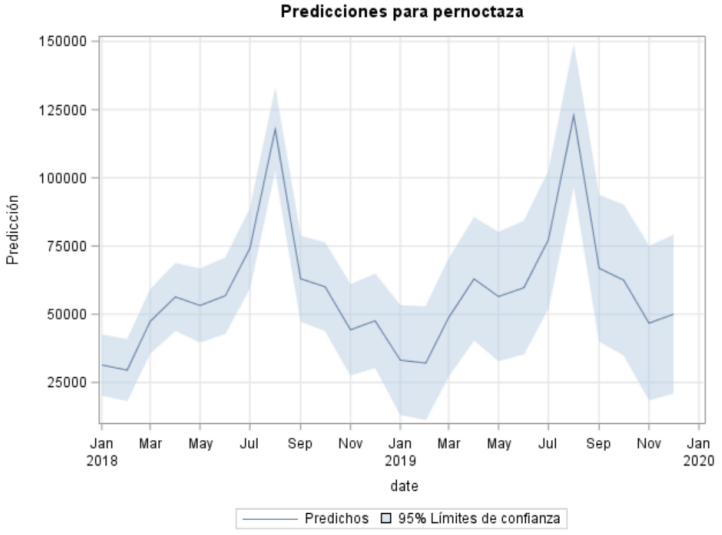
Modelo 11



Modelo 13



Modelo 16



En cuanto a las bandas de confianza, parece que no existen grandes diferencias entre los dos modelos de arriba, pero el tercer modelo parece que las bandas de predicción son más anchas que las otras dos, aunque los tres parecen predecir con bastante exactitud, lo cual no nos indica nada para destacar uno de ellos.

A la vista de los resultados, Análisis de Residuos (aunque los residuos no sean normales), SSE, SSEp, los AIC y las bandas de predicción, concluimos con el que mejor modelo parece ser el modelo

(11) $SARIMA\ (0,1,1)(0,1,2)_{12}$ siendo el que mejor estimación de parámetros tiene y correlaciones, siendo los residuos homogéneos de media cero y además el que tiene los mejores valores de comparación entre modelos.

- SSE = 6824594060.2
- AIC = 4558.715
- SSEp = 2049966553.9

La ecuación del modelo sería la siguiente:

$$(1 - B)(1 - B^{12})\, X_t = (1 - \theta B)(1 - \Theta_1 B^{12} - \Theta_2 B^{24}) a_t$$

$$(1 - B - B^{12} + B^{13}) X_t = (1 - \theta B - \Theta_1 B^{12} + \theta\Theta_1 B^{13} - \Theta_2 B^{24} + \theta\Theta_2 B^{25}) a_t$$

$$X_t - X_{t-1} - X_{t-12} + X_{t-13} = a_t - \theta a_{t-1} - \Theta_1 a_{t-12} + \theta\Theta_1 a_{t-13} - \Theta_2 a_{t-24} + \theta\Theta_2 a_{t-25}$$

$$X_t = X_{t-1} + X_{t-12} - X_{t-13} + a_t - \theta a_{t-1} - \Theta_1 a_{t-12} + \theta\Theta_1 a_{t-13} - \Theta_2 a_{t-24} + \theta\Theta_2 a_{t-25}$$

PREDICCIÓN

Una vez elegido el modelo, vamos a utilizarlo para calcular predicciones para valores futuros más allá de los 240 datos que disponemos, con un intervalo de predicción para las mismas.

Para ello utilizamos el proc ARIMA en SAS con la opción *lead = k*, siendo k el número de predicciones que queremos calcular. Esta vez, realizamos la estimación del modelo con los 240 datos, sin reservar *k*, pues ahora no estamos probando la capacidad de predicción del modelo.

Para un *k = 24*, es decir, dos años de predicción, obtenemos los siguientes resultados.

Predicciones para la variable pernoctaza				
Obs	Predicción	Error Std	95% Confidence Limits	
241	45008.7304	5464.0593	34299.3709	55718.0899
242	48798.8569	5512.9454	37993.6825	59604.0314
243	63806.0748	5561.4018	52905.9276	74706.2219
244	83090.0303	5609.4396	72095.7308	94084.3298
245	73807.6636	5657.0694	62720.0112	84895.3160
246	83923.6526	5704.3016	72743.4268	95103.8784
247	97059.8477	5751.1460	85787.8087	108331.8866
248	154093.9839	5797.6118	142730.8736	165457.0942
249	91654.1374	5843.7081	80200.6799	103107.5949
250	81202.0115	5889.4437	69658.9139	92745.1091
251	66785.9974	5934.8269	55153.9505	78418.0443
252	69437.5389	5979.8656	57717.2177	81157.8601
253	52039.4914	7039.0730	38243.1618	65835.8210
254	56012.4922	7129.5435	42038.8437	69986.1407
255	70089.7823	7218.8803	55941.0370	84238.5276
256	91078.9879	7307.1249	76757.2864	105400.6895
257	81455.0089	7394.3164	66962.4151	95947.6028
258	93877.3853	7480.4918	79215.8909	108538.8797
259	105339.5881	7565.6856	90511.1169	120168.0594
260	165635.5461	7649.9307	150641.9574	180629.1348
261	99342.6969	7733.2581	84185.7895	114499.6044
262	88623.3947	7815.6972	73304.9096	103941.8797
263	74643.7065	7897.2758	59165.3304	90122.0826
264	77527.0064	7978.0202	61890.3741	93163.6386