Practica II: Series Temporales

Análisis Inteligente de Datos

Universidad: UNIVERSIDAD POLITÉCNICA DE MADRID

Campus: MONTEGANCEDO

Centro: ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INFORMÁTICOS

Identificación del Máster: MÁSTER UNIVERSITARIO EN INGENIERÍA INFORMÁTICA

Asignatura: ANÁLISIS INTELIGENTE DE DATOS

Nombres de los alumnos:

ABEL DE ANDRÉS GÓMEZ

GERMÁN SÁNCHEZ GRANADOS

Fecha de Entrega: 10/01/2016

Series Temporales

Explicación de los conjuntos de datos.

En cuanto a los datos con los que obtendremos nuestras series temporales van a ser la tasa de desempleo de jóvenes (entre 20 y 29 años) medido en tanto por ciento. Se trata de una serie cuya frecuencia es de cada tres meses (Marzo, Junio, Septiembre y Diciembre) desde Junio del año 1987 a Septiembre del año 2013.

Los datos han sido obtenidos del Banco de España.

Análisis sobre las características de la serie.

Análisis de la componente tendencial

La tendencia es una componente de la serie que refleja la evolución a largo plazo. Esta componente puede ser de carácter estacionaria (si se representa por una recta paralela al eje de abscisas), de naturaleza lineal (creciente o decreciente), de naturaleza exponencial, etc.

Una primera idea sobre la presencia de tendencia en la serie la obtendremos en su representación gráfica. Pero no siempre estará tan clara como en nuestra serie.

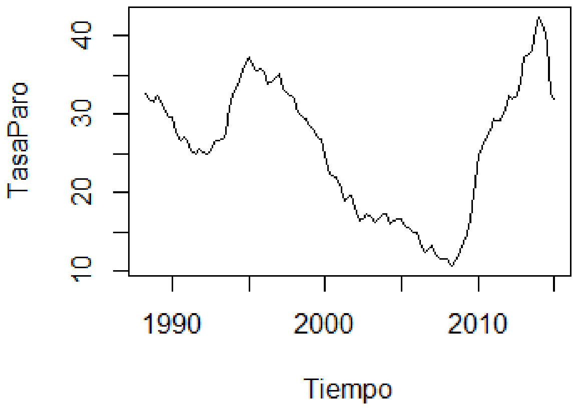

Ilustración 1. Gráfico de la Serie (TasaParo-Tiempo)

Como podemos observar en la gráfica, sí que existe una tendencia negativa, pero quizá no se pueda observar con tanta claridad.

Análisis de descomposición

A continuación vamos a descomponer nuestra serie temporal mediante una descomposición aditiva. El motivo por el cual hemos seleccionado la descomposición aditiva en vez de la multiplicativa ha sido que aunque existan fluctuaciones, estas no se ven afectadas por la tendencia. Otro motivo ha sido que existen oscilaciones constantes en el tiempo.

Consideramos que existe una tendencia lineal negativa, a pesar de que existan picos de aumento en la tasa del paro.

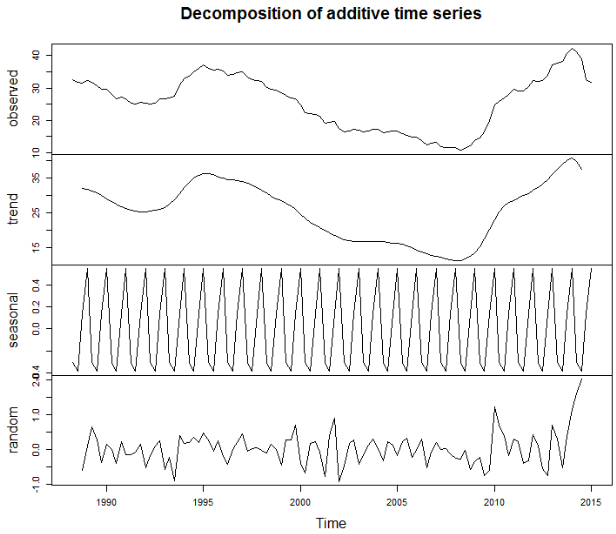

Ilustración 2 Descomposición aditiva

Como podemos observar, en el apartado de "trend", se puede comprobar de forma más clara la tendencia negativa, esto quiere decir que el paro va reduciéndose ligeramente desde el 1987 hasta el 2008 (aproximadamente). Por lo tanto, la tendencia es que la tasa de paro vaya reduciéndose a lo largo del tiempo. Aunque al final de la serie, existe una tendencia positiva, la tendencia negativa predomina a lo largo del tiempo, por lo tanto consideramos la tendencia como lineal negativa.

Análisis mediante el uso de medias móviles

Otros medios para detectar la tendencia de una serie se basan en la aplicación de filtros a los datos. Entre esos filtros encontramos las medias móviles, que también las utilizaremos a continuación.

Se han realizado las medias móviles de orden 3,4 y 7 para nuestra serie de datos.

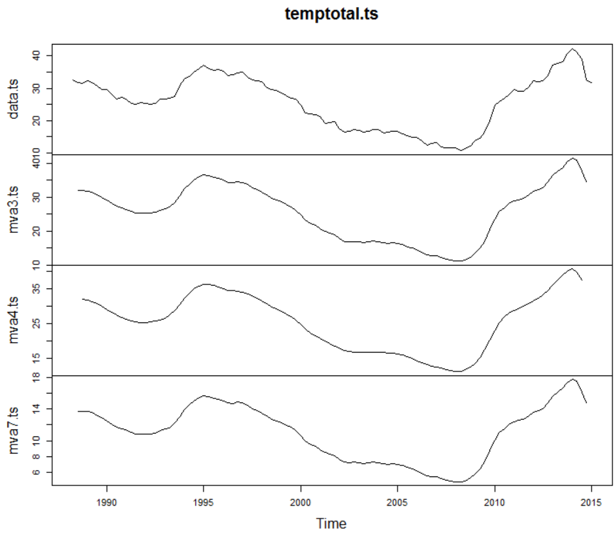

Ilustración 3 Medias Móviles

Mediante este análisis, obtenemos los mismos resultados que usando la descomposición, podemos ver que existe una tendencia lineal negativa desde 1987 hasta aproximadamente el 2008, y es a partir de este donde podremos encontrar una tendencia positiva (surgida por la crisis a la que comenzó ese año). Podemos suponer que si no existiera la crisis, el país podría haber seguido disminuyendo su tasa de empleo juvenil a lo largo del tiempo.

Análisis de la componente cíclica

La componente cíclica recoge las oscilaciones periódicas de amplitud superior a un año y se deben principalmente a la alternancia de etapas de aumento y disminución de la tasas de paro.

En cuanto esta componente, si podríamos decir que existe una componente cíclica, aunque no se puede establecer cada cuanto tiempo. Esto se debe a que cualquier decisión que se tome de un día a otro puede desencadenar un aumento o una disminución drástica en la tasa del paro.

Si nos damos cuenta, en el 1993 España estuvo en una crisis económica en el que la tasa de empleo, como se puede ver en la *Ilustración 1* aumento considerablemente. Posteriormente la tasa de paro ha ido disminuyendo mediante una tendencia lineal negativa hasta llegar a la crisis del 2008, donde vuelve a haber un aumento en la tasa de paro.

Análisis de la componente estacional

La componente estacional recoge las oscilaciones que se producen en periodos iguales o inferiores a un año y que se repiten de forma regular en los diferentes años. Esta componente aparece en las series en las que se toma más de un dato al cabo del año.

A plena vista observando la *Ilustración 1,* no se puede determinar la existencia de la componente estacional. Sin embargo, si realizamos un análisis más profundo sobre la serie mediante la descomposición de esta, podremos observar que sí existe una componente estacional.

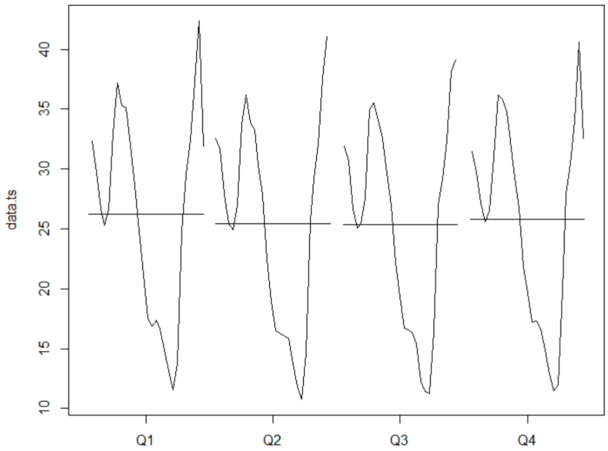

Ilustración 4 Gráfico de la estacionalidad trimestral

Como se puede observar en la *Ilustración 4*, podemos afirmar que si existe una componente estacional, del grafico se pueden sacar aquellos meses cuya tasa de paro juvenil es mayor y en cuales la tasa juvenil es menor. La grafica se divide en 4 trimestres.

Podemos destacar que en el segundo y tercer trimestre, la tasa de paro disminuye, mientras que en el primer y cuarto trimestre, la tasa de desempleo aumenta. Esto se debe a que en España se crea más trabajo los meses de verano y los del final de la primavera ya es un país en el que en esos meses hay una gran cantidad de turismo. Una vez que se termina la temporada de turismo, la tasa de desempleo aumenta.

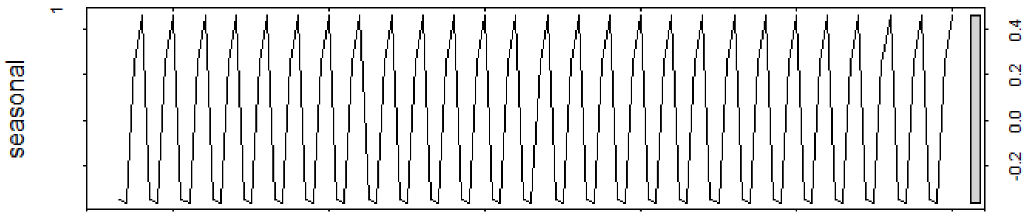

Ilustración 5 Gráfico de la estacionalidad de la serie

Análisis de la componente residual

La componente residual recoge las fluctuaciones erráticas que se dan por la ocurrencia de fenómenos imprevisibles (una huelga, una catástrofe, una crisis, etc.).

Clasificación de la serie

Una serie es **estacionaria** si la media y la variabilidad de mantienen constante a lo largo del tiempo.

Una serie es **no estacionaria** si la media y/o la variabilidad cambian a lo largo del tiempo y además:

- Pueden mostrar <u>cambios de varianza</u>.

- Pueden mostrar una tendencia, es decir que la media crece o baje a lo largo del tiempo.
- Pueden presentar efectos estacionales, es decir que el comportamiento de la serie es parecido en ciertos tiempos periódicos en el tiempo.

Por consiguiente, y teniendo en cuenta el análisis que hemos hecho hasta este punto, podemos decir que nuestra serie es no estacionaria.

Estudio de los residuos

Para realizar un estudio sobre los residuos, primero deberemos descomponer nuestra serie como ya hemos anteriormente y quedarnos con los residuos (remainder). Los residuos no dejan de ser una serie temporal.

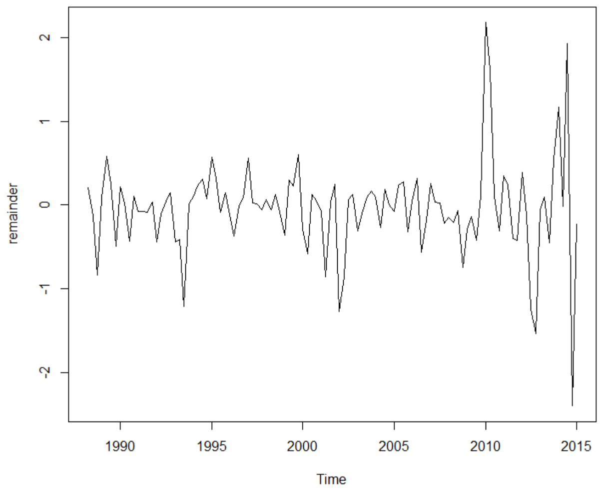

Ilustración 6 Gráfico de los residuos

Ahora debemos analizar si los residuos que se muestran en la Ilustración 6, corresponden a un Ruido Blanco. Un Ruido Blanco es una serie estacionaria en la que ninguna observación influye sobre las siguientes. Para ver si es ruido blanco, deberemos analizar la Función Autocorrelación Simple (FAS ó ACF) y la Función Autocorrelación Parcial (FAP ó PACF) de nuestro gráfico de residuos.

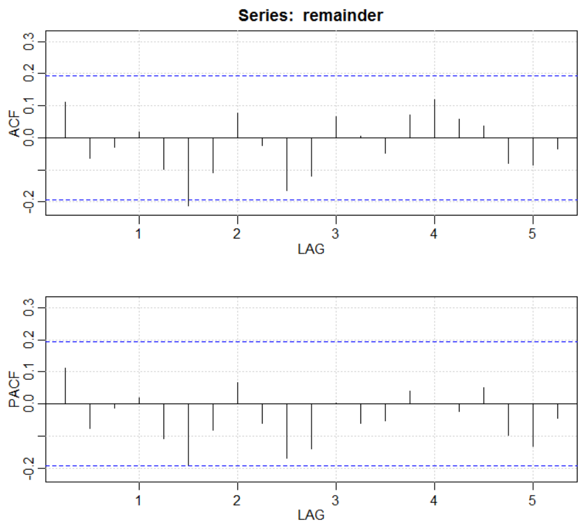

Ilustración 7 Gráfico de FAS y FAP de residuos

En la Ilustración 7, podemos ver una serie de barras verticales y unas bandas horizontales (en azul). Las bandas horizontales son los límites para considerar significativo un retardo (las barras verticales). Es decir, si una barra está dentro de las bandas la consideraremos no significativa en general.

Por lo tanto, podemos ver que la mayoría de las barras en las gráficas son no significativas y por consiguiente podemos decir que no tienen una gran relación con el resto. Pudiendo afirmar que la mayoría de observaciones no van a estar relacionadas entre sí.

Como conclusión, los residuos si serán un Ruido Blanco, puesto que los retardos de sus FAS y su FAP se encuentran dentro del límite y significa que no están relacionados entre sí, cumpliéndose la característica para que una serie estacional sea Ruido Blanco.

Análisis de predicción

Método del suavizado Holt-Winters

Como ya conocemos, existen varios métodos para suavizar una serie y conseguir así que esta nos proporcione una predicción futura.

Nosotros hemos elegido este método puesto que hemos visto que nuestra serie tenía una tendencia de naturaleza lineal y negativa y además hemos observado que existe una variación en la componente estacional.

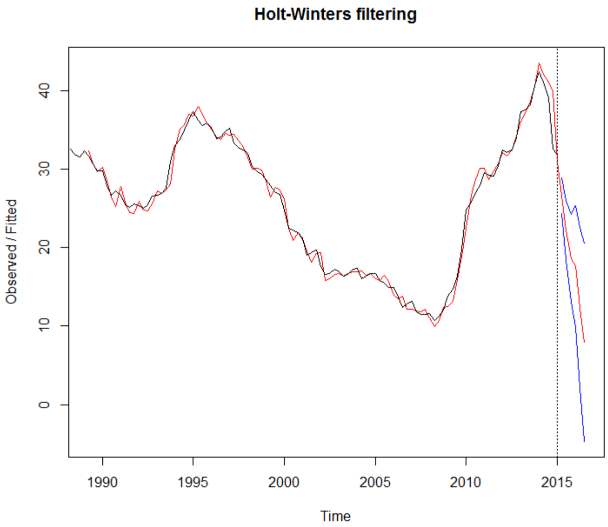

Ilustración 8 Gráfico de predicción de Holt-Winters

Como se puede comprobar en el gráfico, se observa que la tasa de paro disminuirá, es más, teniendo en cuenta la gráfica obtenida mediante los datos, se puede observar que ya en 2014, la tasa de paro está disminuyendo.

El grafico muestra tres predicciones en las que el paro disminuye a una velocidad diferente. Esto se debe que, aunque podamos predecir el futuro, no podemos conocerlo. Podemos decir que lo que diferencia entre las tres predicciones podría ser eventos que ocurran, por ejemplo métodos de gobierno, eventos económicos destacables, etc.

Ajuste de la serie a un modelo ARIMA

La metodología de Box y Jenkins se resume en cuatro fases:

- La **primera fase** consiste en identificar el posible modelo ARIMA que sigue la serie, lo que requiere:
  - Decidir que transformaciones aplicar para convertir la serie observada en una serie estacionaria.

    Como ya hemos visto, la serie no era estacionaria, por lo tanto hemos realizado unas transformaciones para hacer que si sea estacionaria. En un primer lugar hemos aplicado una transformación logarítmica, pero nos hemos dado cuenta de que aun así, la serie seguía siendo estacionaria.

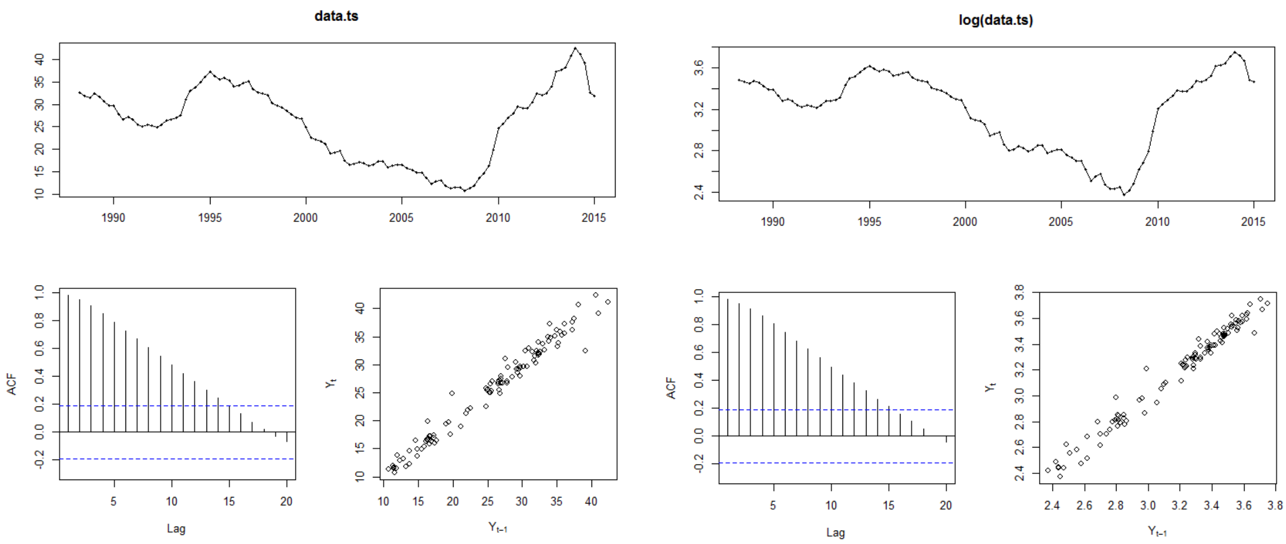

Ilustración 9 Serie sin transformación (izda.) y serie con transformación logarítmica (dcha.)

Por lo tanto, hemos decidido que en vez de aplicar una función logarítmica a nuestra serie, le vamos a aplicar una transformación diferencial.

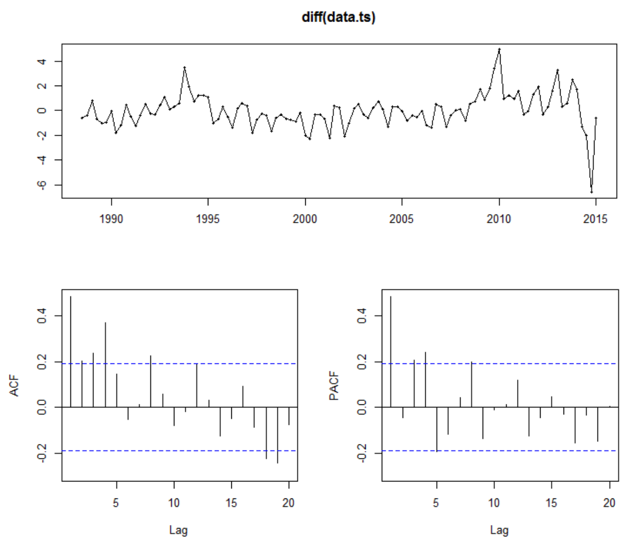

Ilustración 10 Serie con transformación diferencial

Como se puede ver en la Ilustración anterior, hemos conseguido que la serie sea estacionaria, no es tan perfecta, pero al menos hemos conseguido que la tendencia sea casi nula.

Ahora debemos pensar si debemos realizar una nueva diferencial para hacer nuestra serie más estacionaria o dejarla como esta. Para ello vamos a utilizar la desviación estándar:

Desviación estándar de la serie sin transformadas --> **8.37697**
Desviación estándar de la serie con una diferenciación --> **1.394642**
Desviación estándar de la serie con dos diferenciaciones --> **1.421934**

Como vemos, la desviación estándar aplicando dos transformaciones empeora y además, al realizar transformaciones, perdemos observaciones e información, por lo tanto sugerimos que solo se realice una diferenciación, y por lo tanto el grado de diferenciación será de 1.
A continuación se muestra la serie habiéndola aplicado dos transformaciones (en nuestro caso solo utilizaremos una).

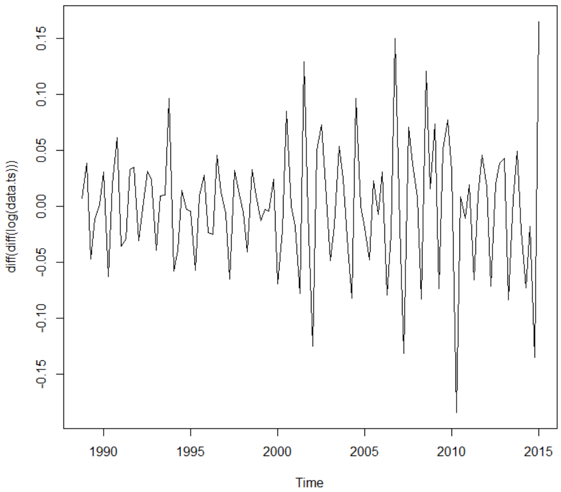

Ilustración 11 Serie con transformación logarítmica y diferencial II

Ahora sí, podemos comprobar que la tendencia es nula y la varianza es más constante que en la Ilustración 10, lo que hace que la serie sea estacionaria.

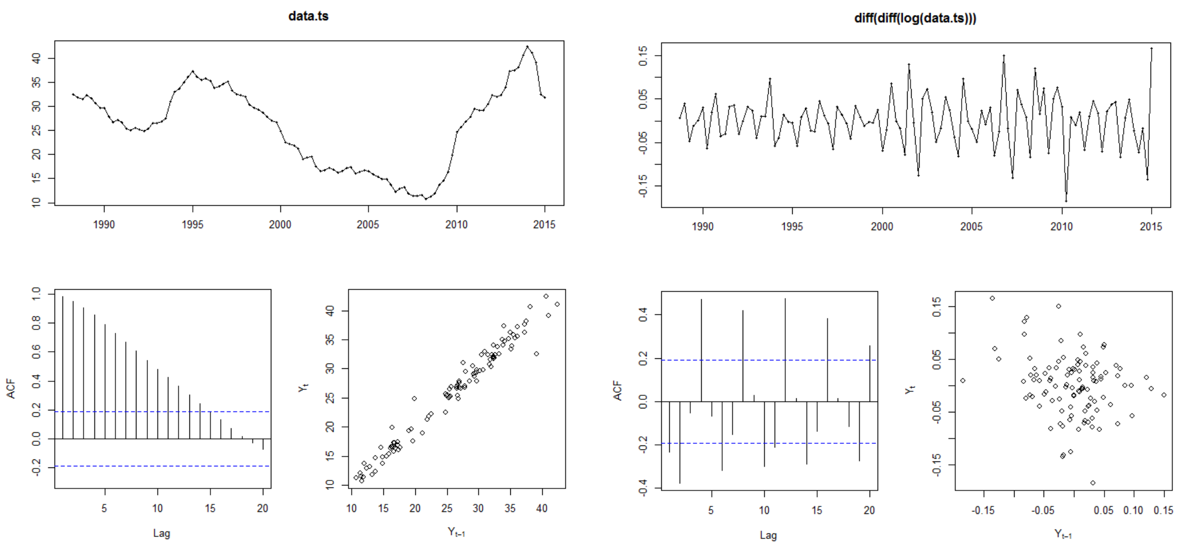

Ilustración 12 Serie antes y después de aplicar la transformación respectivamente

Otro motivo por el que sabemos que hemos realizado una transformación valida ha sido que el FAS en la serie sin transformación decrece muy lentamente, lo que significa que es una serie no estacionaria mientras que después de aplicarle una serie de transformaciones, podemos comprobar que ya no existe ese decrecimiento lento, lo que significa que ya es estacionaria.

Por último, si nos fijamos en los residuos, la serie no estacionaria posee unos residuos estructurados (poseen una estructura lineal positiva) mientras que en la serie ya estacionaria los residuos no poseen ninguna estructura, esto también nos muestra que la serie ya es estacionaria.

- Determinar un modelo ARMA para la serie estacionaria, es decir, los órdenes p y q de su estructura autorregresiva y de media móvil.

Como se ha comentado anteriormente, ha sido necesario una diferenciación para poder hacer que nuestra serie sea estacionaria, por lo tanto, nuestro parámetro d=1.

Una vez hecho esto, lo que haremos es estudiar la FAP y la FAS del modelo transformado, es decir, de la diferenciación de nuestra serie.

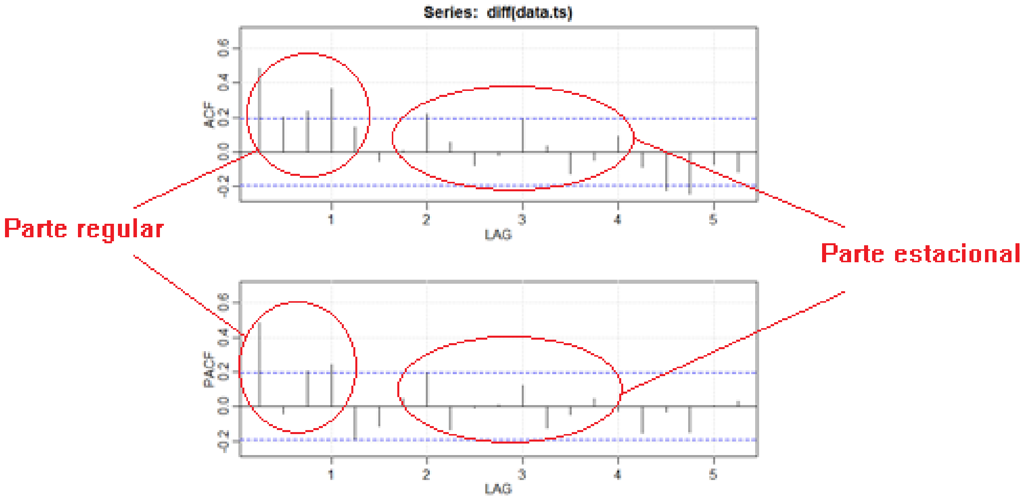

Ilustración 13 FAS y FAP de la serie diferenciada

Para realizar el estudio, tendremos en cuenta la parte regular y la parte estacional.

En cuanto a la parte regular, hemos considerado que se trata de un autorregresivo de orden 1 ya que en la FAS existe una estructura decreciente (positiva) y en la FAP existe una única barra significativa (y también positiva).

Respecto a la parte estacional, hemos considerado que se trata de una media móvil de orden 1 ya que la FAS existe una barra positiva y significativa y en la FAP existen barras alternadas.

Por consiguiente nuestro modelo para la serie será:

ARIMA(1,1,1)x(1,0,1)$_4$

- La **segunda fase**: Seleccionando provisionalmente un modelo para la serie estacionaria, se pasa a la segunda etapa de estimación, donde los parámetros AR y MA del modelo se estiman por máxima verosimilitud y se obtienen los residuos del modelo.
  Es necesario destacar que se han probado otra serie de modelos, puesto que para estimar un modelo a simple vista es necesario tener bastante experiencia. Entre ellos los siguientes:

MODELO	AIC's
ARIMA(1,1,0)x(0,0,0)	349.5193
ARIMA(1,1,0)x(1,0,0)	331.2742
ARIMA(1,1,1)x(1,0,0)	331.2742
ARIMA(1,1,1)x(1,0,1)	326.8445

Si usamos la orden de **auto.arima** de nuestra serie temporal, podremos ver que nuestra decisión no ha sido del todo perfecta, pero ha estado muy cerca, puesto que el modelo estimado automáticamente es **ARIMA(1,1,0)x(1,0,1)$_4$** cuya AIC es de 326.16**.** Es un poquito mejor que nuestro modelo.

```
 > auto.arima(data.ts)
Series: data.ts
ARIMA(1,1,0)(1,0,1)[4]

Coefficients:
         ar1     sar1     sma1
      0.5770   0.9043  -0.6372
s.e.  0.0832   0.0795   0.1565

sigma^2 estimated as 1.119:  log likelihood=-159.08
AIC=326.16   AICc=326.55   BIC=336.85
```

Puesto que no existe una gran diferencia entre los modelos, continuaremos el estudio con el modelo elegido por nosotros.

Continuando con nuestro modelo, concluimos que se ha estimado mediante un AR(1) y un MA(1), por lo tanto p=1 y q=1, también para la parte estacional. No se ha realizado ninguna transformación para la parte estacional, por lo tanto D=0.

Después se muestran los coeficientes para la creación de la ecuación de nuestro modelo y otros datos como el AIC(Akaike information criterion) para la toma de decisiones sobre qué modelo elegir.

```
> data.ts.5
Series: data.ts
```

```
ARIMA(1,1,1)(1,0,1)[4]

Coefficients:
         ar1      ma1    sar1     sma1
      0.7325  -0.2162  0.9162  -0.6534
s.e.  0.1268   0.1651  0.0684   0.1418

sigma^2 estimated as 1.101:  log likelihood=-158.42
AIC=326.84   AICc=327.44   BIC=340.21
```

- La **tercera fase** es el diagnostico, donde se comprueba que los residuos no tienen estructura y dependencia y siguen un proceso de ruido blanco. Si los residuos muestran estructura se modifica el modelo para incorporarla y se repiten las etapas anteriores hasta obtener un modelo adecuado.

  Como se muestra en la *Ilustración 15*, los residuos tienen una estructura característica de un ruido blanco, ya que ninguna de las observaciones depende de las otras, o dicho de otra forma, ninguna barra vertical supera los límites horizontales (sucesión de barras no significativas). Además como se puede apreciar, tampoco tienen ninguna estructura.

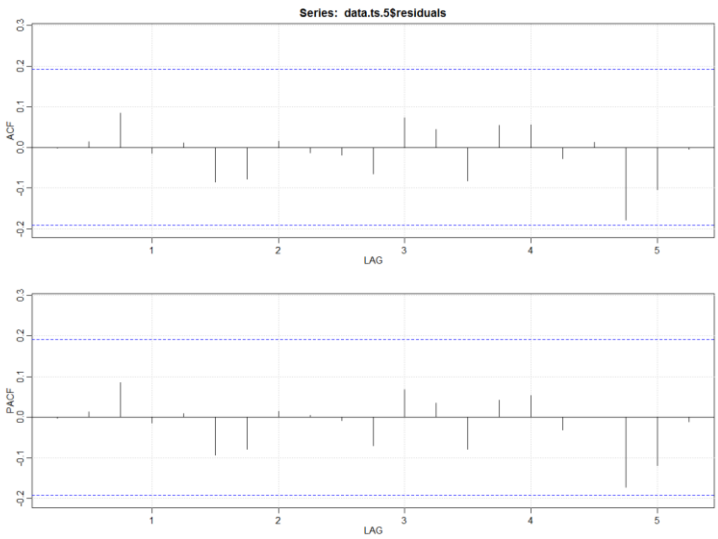

Ilustración 14 Gráfico de la FAP y FAS de los residuos del modelo

Por otro lado, si realizamos test a los residuos, podremos comprobar que todos los test realizados de Box-Ljung, tienen un p-value mayor que 0.05 por lo tanto, los residuos son independientes y es justamente lo que se busca para que el modelo sea aceptable.

Por último, podemos decir que el test de Jarque Bera es un test de normalidad en los datos, y como podemos comprobar, el p-value resultante es casi nulo, por lo que los datos provienen de una distribución normal.

```
LB.test(data.ts.5)

  Box-Ljung test
```

```
data:  residuals from  data.ts.5
X-squared = 3.6917, df = 8, p-value = 0.8838

> #or
> LB.test(data.ts.5, lag=20)

  Box-Ljung test

data:  residuals from  data.ts.5
X-squared = 11.406, df = 16, p-value = 0.7838

> #normality of the residuals
> jarque.bera.test(residuals(data.ts.5))

  Jarque Bera Test

data:  residuals(data.ts.5)
X-squared = 845.34, df = 2, p-value < 2.2e-16
```

- La **cuarta fase** es la predicción, una vez que se ha obtenido un modelo adecuado se realizan predicciones con el mismo.

  Como se puede ver en la Ilustración 15, mediante el modelo elegido, hemos conseguido predecir cuáles serán las posibles observaciones futuras. Como podremos observar (mediante los puntos) la tasa de paro juvenil disminuirá a partir del 2013.

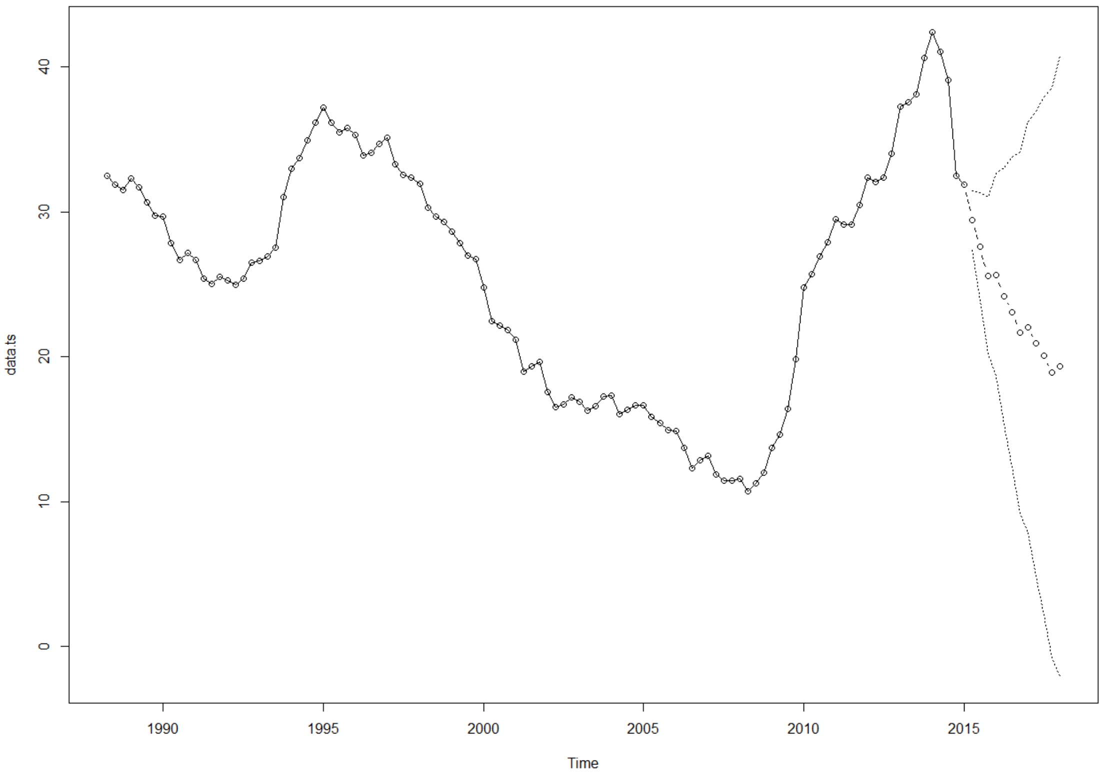

Ilustración 15 Predicción del modelo seleccionado