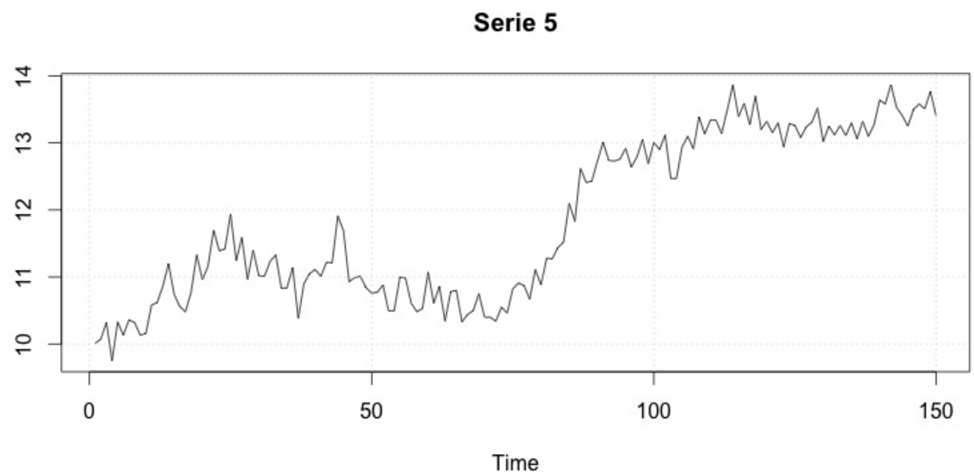
Universidade de São Paulo
Instituto de Matemáticas e Estatística
MAE 5870 – Análise de Séries temporais
Prova # 1 - Parte 2
Nome: Duván Humberto Cataño Salazar

Qual é o melhor modelo para sua série? Justifique. (Ajustar um modelo SARIMA, use criterio AIC, BIC e error quadrático medio para comparar os modelos e faca previsão para h=10).

Solución:

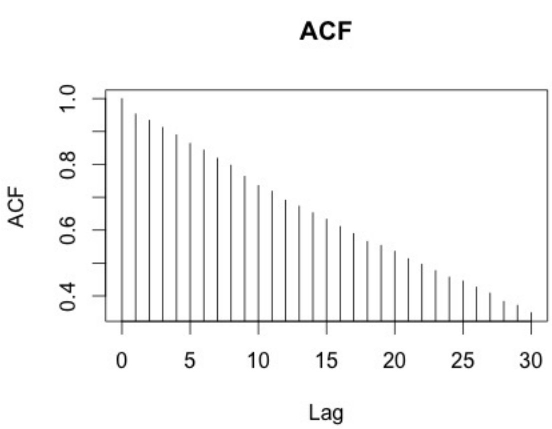
EXPLORACIÓN DE LOS DATOS:

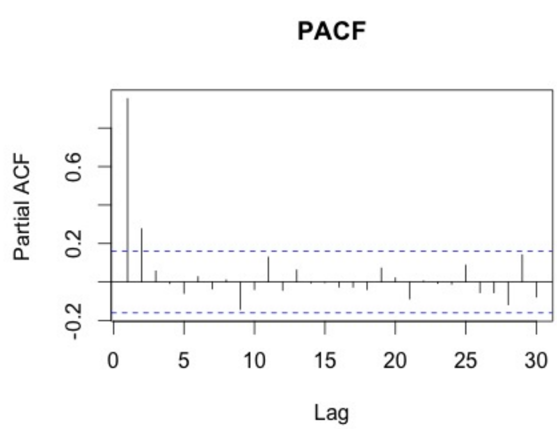
En primer lugar, graficamos la serie en consideración, la cual consta de 150 observaciones.



Como se puede notar, dicha serie muestra aproximadamente en la mitad del tiempo un cambio de nivel, una oscila alrededor de 11 y el otro alrededor de 13, por lo cual podemos pensar que la serie no es estacionaria. Notemos además que a primera vista, la serie no presenta cambios notables en la estructura, por lo cual suponemos que su varianza no cambia con el tiempo.

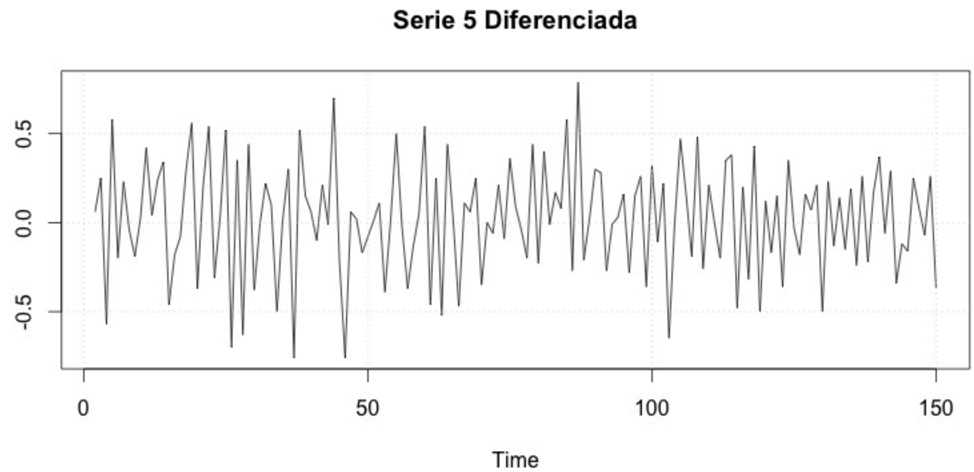
Con el fin de analizar las características de correlación, se observa la ACF y PACF de la serie.

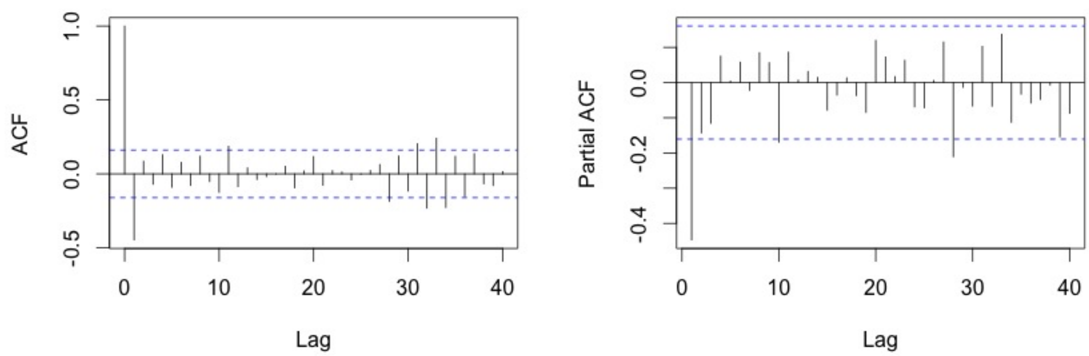




Como se puede observar en dichas gráficas, la ACF decae muy lentamente en forma aritmética, lo cual sugiere que la serie debe ser diferenciada y la PACF muestra valores significativos en sus dos primeros rezagos ($h = 1$ y $h = 2$.)

A continuación mostramos la gráfica de la primera diferencia de la serie, con sus respectivas ACF y PACF:





Observamos en primer lugar que la serie sólo oscila alrededor de cero, lo que sugiere que podemos suponer que primera diferencia de la serie es estacionaria; además nótese que tanto la ACF como la PACF, muestran ambas valores significativos en el primer rezago ($h = 1$).
De acuerdo con lo anterior expuesto, los modelos posibles para nuestra serie pueden ser los siguientes:

$$ARIMA(0,1,1)$$
$$ARIMA(1,1,1)$$
$$ARIMA(0,1,1)$$

A continuación empleamos los criterios de información AIC y BIC y el Error Cuadrático Medio (MLE) para determinar cuál de los modelos es el más adecuado para la modelación de nuestra serie.

	AIC	**BIC**	**MLE**
ARIMA(0,1,1)	50.40592	56.42719	0.07982
ARIMA(1,1,1)	50.62901	59.66091	0.07886
ARIMA(1,1,0)	51.50339	57.52466	0.08042

De la información de la tabla, se puede concluir que el mejor modelo para este caso es una ARIMA (0,1,1), el cual presenta en los criterios de información AIC y BIC los valores más pequeños, al igual que un MLE bajo, con muy poca diferencia con respecto al MLE del ARIMA (1,1,1).

Vale la pena aclarar que además se aplicó un algoritmo a nuestros datos, con la función $arima(\cdot)$ de R, el cual realizó todas los posibles combinaciones de órdenes en los modelos para $p \leq 5$ $d = 1$ y $q \leq 5$, y el mejor modelo fue efectivamente el ARIMA(0,1,1). En EL anexo se adjunta el algoritmo empleado.

ESTIMACIÓN DE PARÁMETROS:

A continuación mostramos la estimación del parámetro θ en el modelo:

$$(1-B)x_t = (1+\theta)w_t \qquad \Longleftrightarrow \qquad x_t = x_{t-1} + w_t + \theta w_{t-1}$$

Empleando R por los métodos de ML y MSS se obtienen los siguientes resultados para el parámetro θ con su respectivo error estándar:

```
Series: s5
ARIMA(0,1,1)

Coefficients:
          ma1
      -0.4493
s.e.   0.0634

sigma^2 estimated as 0.07984:  log likelihood=-23.11


##########################################


Series: s5
ARIMA(0,1,1)

Coefficients:
          ma1
      -0.4475
s.e.   0.0635

sigma^2 estimated as 0.07982:  log likelihood=-23.2
```

Como se puede observar, los valores estimados por cada uno de los métodos no muestran diferencia significativa, por lo cual, nosotros en particular, trabajaremos en lo que sigue con

$$\theta = -0.4493,$$

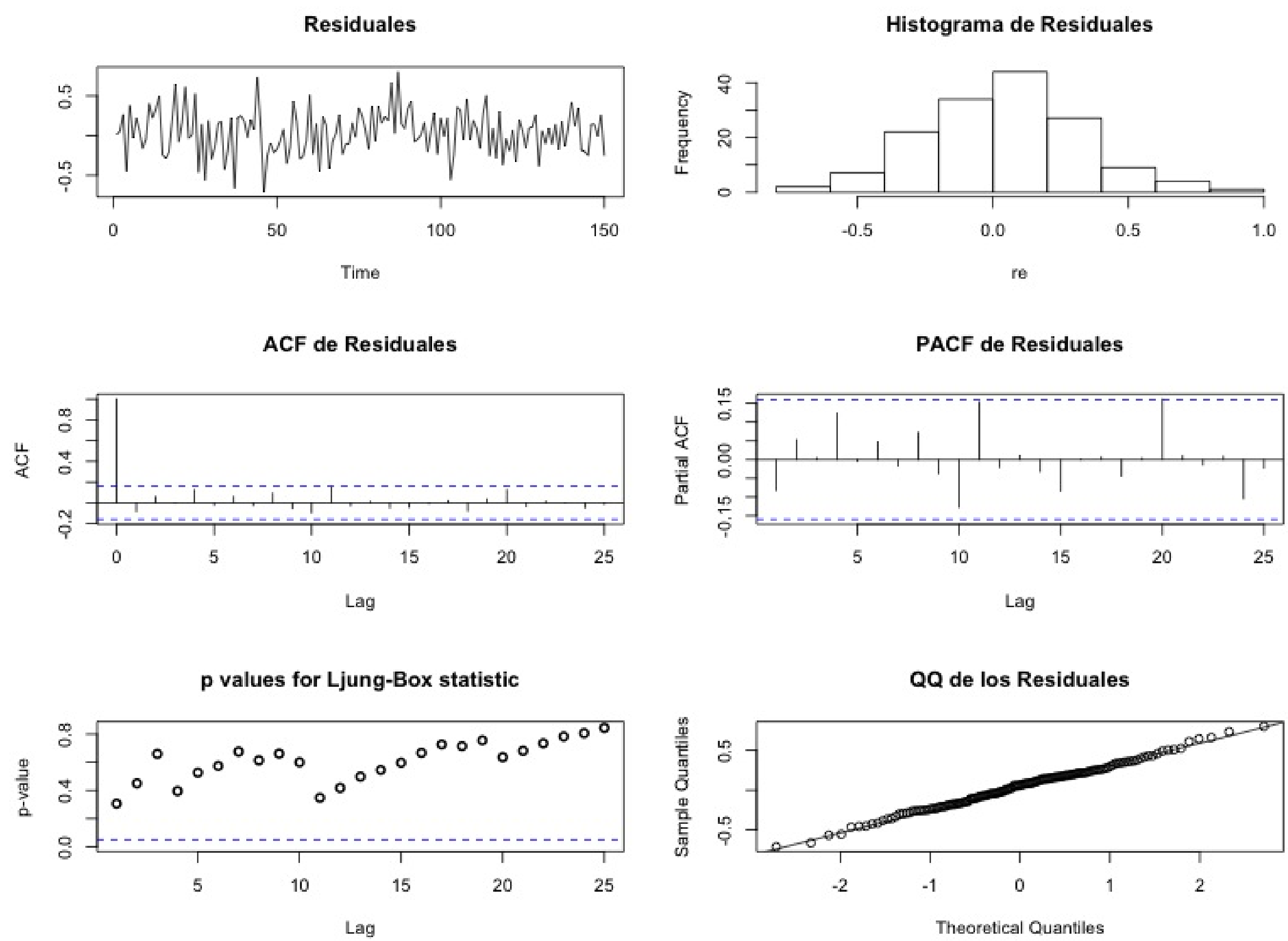
por lo tanto nuestro modelo está dado por:

$$x_t = x_{t-1} + w_t - 0.4493 w_{t-1},$$

donde w_t es una ruido blanco.

ANÁLISIS DE RESIDUALES:

En primer lugar, realizaremos gráficas y diferentes pruebas de hipótesis de normalidad para los residuales que deja dicho modelo al ser ajustado a un $ARIMA(0,1,1)$.



Gráficamente podemos ver que los residuales tienen todas las características de un ruido blanco. El histograma parece tener la estructura de una distribución normal con media 0 y varianza 1; tanto la ACF como la PACF no muestran indicios de no autocorrelación; el estadístico de Ljung-Box, muestra que los residuales no están correlacionados según los $p-$valores y finalmente los residuales están acorde a una normal si la miramos cuantil cuantil.

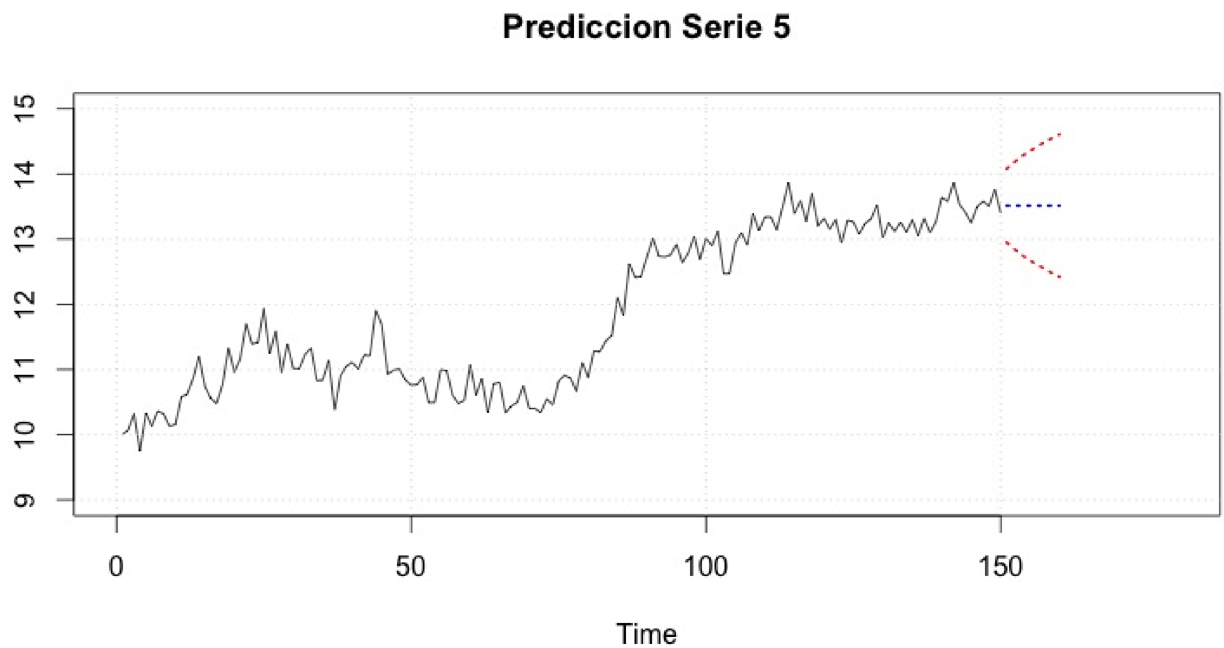
Para validar la afirmación de que los residuales siguen una distribución normal, además, se realizaron las pruebas de Shapiro-Wilks y de Kolmogorov-Smirnov, las cuales arrojaron los siguientes resultados:

```
Shapiro-Wilk normality test

data:  re
W = 0.9955, p-value = 0.9275

##########################################

One-sample Kolmogorov-Smirnov test

data:  re
D = 0.0513, p-value = 0.8251
alternative hypothesis: two-sided
```

Como se puede observar en los resultados, los $p - valores$ en ambos casos, muestran suficiente evidencia para aceptar la hipótesis nula, de que los residuales siguen una distribución normal. Claramente dicha hipótesis se acepta con contundencia para diferentes valores de significancia, pues $p-$valores éstos son lo suficientemente grandes.

PREDICCIÓN:

A continuación mostramos la gráfica de la serie con el valor de la predicción con $h = 10$.



Según el resultado arrojado por R, tenemos que

$$x_{160}^{150} = 13.51409,$$

el cual tiene sentido según la dinámica del proceso.

Notemos además que dicho resultado está acorde con los modelos ARIMA teóricos, pues cuando h crece los intervalos de predicción también incrementan, esto es, los límites llegan a ser cada vez más amplios a medida que h crece. Esto significa que a medida que el horizonte de pronóstico crece, también crece indefinidamente a incertidumbre en el pronóstico. Tal y como se ilustra en la siguiente tabla.

h	lim_sup	lim_inf
151	14.07916	12.94903
152	14.15968	12.86851
153	14.23122	12.79697
154	14.29624	12.73195
155	14.35625	12.67193
156	14.41227	12.61592
157	14.46499	12.56320
158	14.51494	12.51325
159	14.56251	12.46568
160	14.60801	12.42018

ANEXO:

```
sarima.list = function(datos, p, d = 0, q, P = 0, D = 0, Q = 0, S = NA, include.mean = F, criterio = ""){
  M <- matrix(ncol = 10,nrow = (p+1)*(q+1)*(P+1)*(Q+1),dimnames=list(NULL,c("p","d","q","P","D","Q","S","co
  k <- 1
  n <- length(datos)
  for(i in 0:p){
    for(j in 0:q){
      for(l in 0:P){
        for(m in 0:Q){
          if ((i==0)&&(j==0)&&(l==0)&&(m==0)) next #Continua con la siguiente iteracion
          fit <- arima(datos, order = c(i, d, j),seasonal = list(order = c(l, D, m), period = S), include.m
          M[k,1]  <- i
          M[k,2]  <- d
          M[k,3]  <- j
          M[k,4]  <- l
          M[k,5]  <- D
          M[k,6]  <- m
          M[k,7]  <- S
          M[k,8]  <- fit$code  # 0: Convergencia, 1: No Convergencia
          M[k,9]  <- AIC(fit)  # AIC
          M[k,10] <- AIC(fit, k = log(length(datos)))  # BIC
          k <- k+1
        }
      }
    }
  }
  if(criterio == "AIC"){
    M <- M[order(M[,9]),]
  }
  if(criterio == "BIC"){
    M <- M[order(M[,10]),]
  }
  if(criterio == ""){
    M <- M
  }
  rownames(M) = rep("", (p+1)*(q+1)*(P+1)*(Q+1))
  return(M[1:((p+1)*(q+1)*(P+1)*(Q+1)-1),])
}

sarima.list(s5, p=5, d = 1, q=5, P = 0, D = 0, Q = 0, S = NA, include.mean = F, criterio = "")
```