Series de tiempo univariadas - Presentación 23

Mauricio Alejandro Mazo Lopera

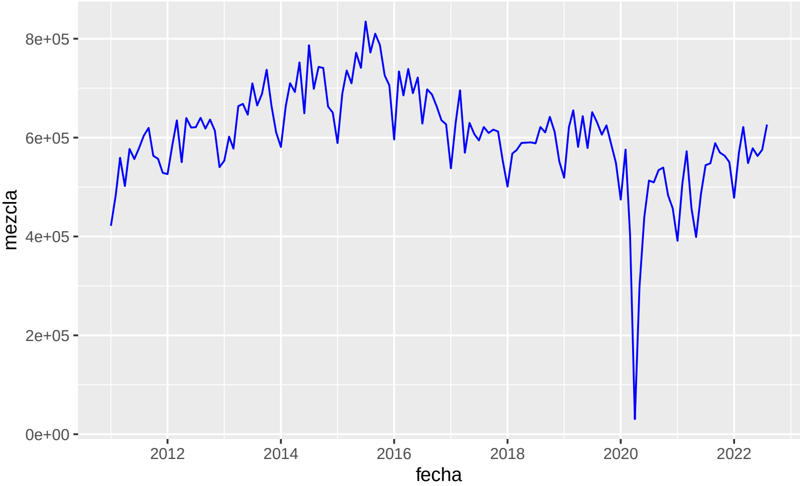
Universidad Nacional de Colombia
Facultad de Ciencias
Escuela de Estadística
Medellín

Considere la base de datos de Estadísticas de concreto premezclado (AQUÍ) disponible en el boletín técnico de Agosto de 2022:

```
require(tidyverse)
require(magrittr)
require(readxl)
bd_ini <- read_excel("../../DATOS/Anex_concreto_ago_22.xls",
                     sheet="Anexo 1 ", skip = 6)
bd_ini %<>% rename("mezcla"="Producción")
concreto <- data.frame(fecha=seq(as.Date("2011/1/1"),
                                 as.Date("2022/8/1"), "months"),
                     mezcla=bd_ini[1:140, 3])
```

Intervenciones - Predicciones: Ejemplo

```
concreto %>% ggplot(aes(x=fecha, y=mezcla))+
                geom_line(col="blue")
```



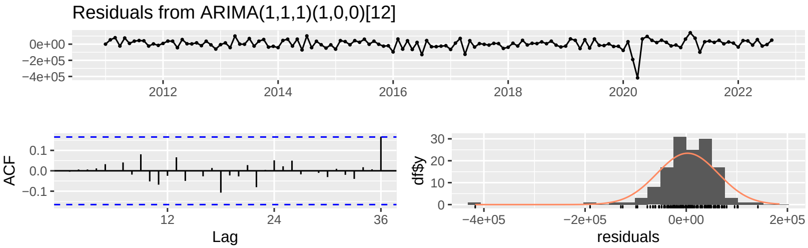
La caída del gráfico anterior se presenta en abril de 2020.

La caída del gráfico anterior se presenta en abril de 2020. Podemos aplicar un modelo $SARIMA(p, d, q) \times (P, D, Q)_s$ a la serie sin tener en cuenta el problema de la intervención:

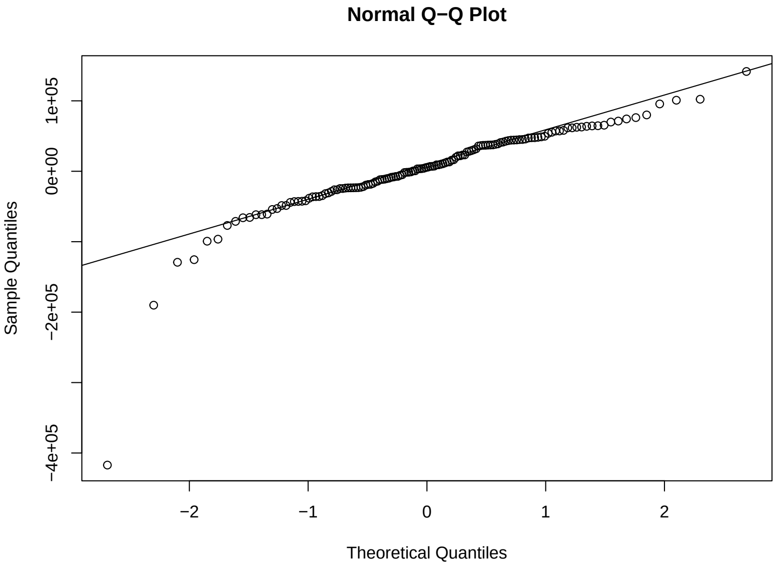
```
require(forecast)
ts_mezcla <- ts(concreto$mezcla, start=c(2011,1),
                frequency = 12)
modelo1_mez <- auto.arima(ts_mezcla,
                          stepwise = FALSE,
                          approximation = FALSE)
```

```r
require(lmtest)
modelo1_mez %>% coeftest()
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ar1   0.501120   0.146846  3.4126 0.0006436 ***
## ma1  -0.851733   0.098753 -8.6249 < 2.2e-16 ***
## sar1  0.355195   0.078233  4.5402  5.62e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,1)(1,0,0)[12]
## Q* = 7.7492, df = 22, p-value = 0.9978
##
## Model df: 3.   Total lags used: 25
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
1e+05
0e+00
−2e+05
−4e+05
−2
−1
0
1
2

```r
require(tseries)
modelo1_mez$residuals %>% shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.82939, p-value = 1.835e-11
```

```r
modelo1_mez$residuals %>% jarque.bera.test()
```

```
##
##  Jarque Bera Test
##
## data:  .
## X-squared = 1568.8, df = 2, p-value < 2.2e-16
```

Intervenciones - Predicciones: Ejemplo

Dividimos la base de datos pre-intervención para ver el orden del modelo a ser ajustado:

```r
ts_mezcla_pre <- window(ts_mezcla, start = c(2011,1),
                                   end=c(2020,3))
```

Usamos la función **auto.arima** del paquete **forecast**:

```r
require(forecast)
auto.arima(ts_mezcla_pre, stepwise = FALSE,
           approximation = FALSE)
```

```
## Series: ts_mezcla_pre
## ARIMA(2,1,0)(0,1,1)[12]
##
## Coefficients:
##           ar1      ar2     sma1
##       -0.8716  -0.2776  -0.7507
## s.e.   0.1115   0.1115   0.1303
##
```

Intervenciones - Predicciones: Ejemplo

Como vimos antes, el modelo a ajustar es un $SARIMA(2,1,0) \times (0,1,1)_{12}$. Usamos la función **arimax** del paquete **TSA** para ver el efecto de la intervención dada por el modelo:

$$(1-\phi_1 B-\phi_2 B^2)(1-B^{12})(1-B)X_t = \frac{\omega_1}{1-\delta_1 B} P_{1t}^{(112)} + w_t + \Theta_1 w_{t-12}$$

```
require(TSA)
modelo2_mez <- arimax(ts_mezcla, order=c(2, 1, 0),
                  seasonal = list(order = c(0, 1, 1)),
  xtransf=data.frame(
   abril2020a=1*(seq_along(ts_mezcla) == 112)),
             transfer=list(c(1, 0)))
```

Intervenciones - Predicciones: Ejemplo

Vemos el resumen del modelo:

```
require(lmtest)
modelo2_mez  %>% coeftest()
```

```
##
## z test of coefficients:
##
##                  Estimate  Std. Error z value  Pr(>|z|)
## ar1            -5.0129e-02  9.3555e-02 -0.5358    0.5921
## ar2            -1.5215e-02  1.0566e-01 -0.1440    0.8855
## sma1           -8.1444e-02  1.2471e-01 -0.6531    0.5137
## abril2020a-AR1  1.4105e-01         NaN     NaN       NaN
## abril2020a-MA0 -2.1152e+05  3.3029e+04 -6.4040 1.514e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

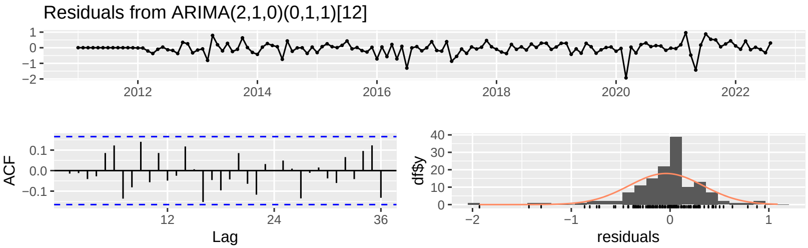
Como vimos antes, el δ_1 estimado no arrojó cálculos en el p-valor, ni en la desviación estándar ni tampoco en el estadístico Z. Este error puede ser resultado de las unidades tan grandes de la variable que estamos analizando. Dichas unidades están en cientos de miles. Dividimos entonces por 10000 y estimamos nuevamente el modelo

```
ts_mezcla2 <- ts_mezcla/100000
modelo3_mez <- arimax(ts_mezcla2, order=c(2, 1, 0),
                  seasonal = list(order = c(0, 1, 1)),
  xtransf=data.frame(
   abril2020a=1*(seq_along(ts_mezcla2) == 112)),
               transfer=list(c(1, 0)))
```

Vemos el resumen del modelo:

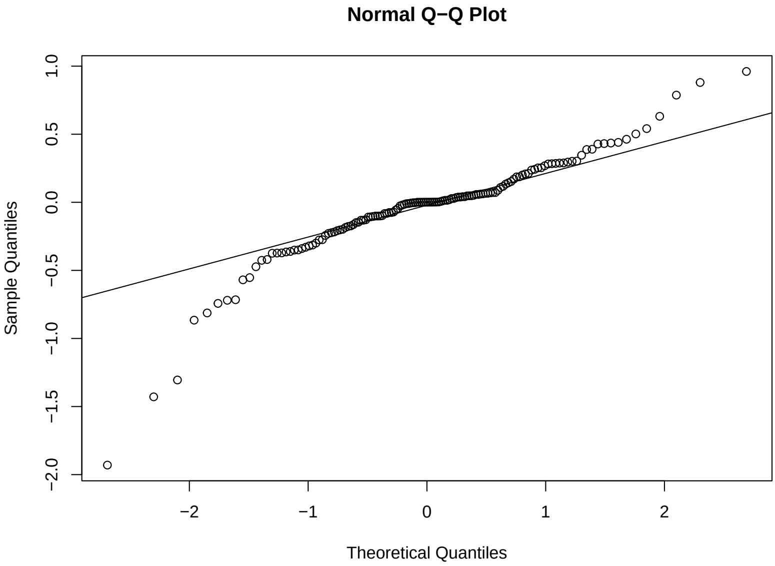
```
require(lmtest)
modelo3_mez  %>% coeftest()
```

```
##
## z test of coefficients:
##
##                  Estimate Std. Error  z value  Pr(>|z|)
## ar1             -0.763532   0.097475  -7.8331 4.759e-15 ***
## ar2             -0.246186   0.092339  -2.6661  0.007673 **
## sma1            -0.802139   0.096298  -8.3297 < 2.2e-16 ***
## abril2020a-AR1   0.328429   0.062813   5.2287 1.707e-07 ***
## abril2020a-MA0  -4.941668   0.358574 -13.7814 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,1,0)(0,1,1)[12]
## Q* = 25.493, df = 20, p-value = 0.1832
##
## Model df: 5.   Total lags used: 25
```

Ejemplo aplicado a datos reales: Desempleo



Intervenciones - Predicciones: Ejemplo

```
require(tseries)
modelo3_mez$residuals %>% shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.89219, p-value = 1.203e-08
```

```
modelo3_mez$residuals %>% jarque.bera.test()
```

```
##
##  Jarque Bera Test
##
## data:  .
## X-squared = 227.47, df = 2, p-value < 2.2e-16
```

```r
require(forecast)
predict(modelo1_mez, n.ahead = 5)
```

```
## $pred
##           Jan Feb Mar Apr May Jun Jul Aug      Sep      Oct      Nov      Dec
## 2022                                      623446.9 607965.1 601281.3 594565.0
## 2023 567765.7
##
## $se
##           Jan Feb Mar Apr May Jun Jul Aug      Sep      Oct      Nov      Dec
## 2022                                      61288.66 73077.66 78633.22 82108.57
## 2023 84734.50
```

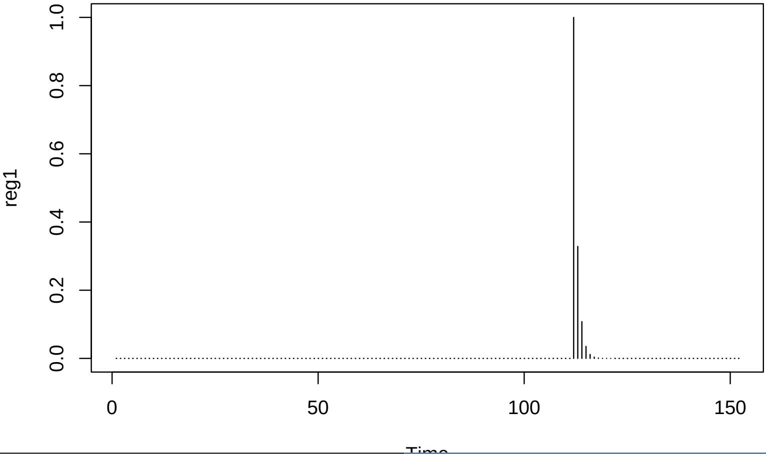
Para hacer predicciones con el modelo intervenido cuando existe una reducción gradual del efecto de la intervención es necesario definir los regresores correspondientes a la intervención:

```r
delta1 <- 0.328429
reg1 <- stats::filter(1 * (seq.int(length(ts_mezcla2) + 12)== 112),
              filter = delta1, method = "rec",
              sides = 1)
```

Intervenciones - Predicciones: Ejemplo

Vemos cuál es la magnitud del efecto que se reduce lentamente

```
plot(reg1, type="h")
```



Intervenciones - Predicciones: Ejemplo

Para realizar predicciones podemos ajustar un modelo con la función **arima**:

```
xreg <- cbind(I1=stats::filter(1*(seq_along(ts_mezcla2) == 112),
                                  filter=0.328429,method = "rec",
                                  sides = 1))

modelo3_mez_a<-arima(ts_mezcla2, order = c(2, 1, 0),
                     seasonal = list(order = c(0,1,1),
                                     period = 12),
                     xreg = xreg)
```

Intervenciones - Predicciones: Ejemplo

```
modelo3_mez_a %>% coeftest()
```

```
##
## z test of coefficients:
##
##        Estimate Std. Error  z value  Pr(>|z|)
## ar1   -0.763531   0.095419  -8.0019 1.225e-15 ***
## ar2   -0.246186   0.092242  -2.6689  0.007609 **
## sma1  -0.802139   0.090856  -8.8287 < 2.2e-16 ***
## xreg  -4.941669   0.357075 -13.8393 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Las predicciones serían:

```r
reg1 <- ts(reg1, start = start(ts_mezcla2),
           frequency = frequency(ts_mezcla2))
reg1_new <- window(reg1, start = c(2022, 9),
                   end=c(2023,1))
predict(modelo3_mez_a, newxreg=reg1_new, n.ahead=5)
```

```
## $pred
##          Jan Feb Mar Apr May Jun Jul Aug      Sep      Oct      Nov      Dec
## 2022                                     6.022790 6.165838 5.812795 5.471153
## 2023 4.888438
##
## $se
##           Jan Feb Mar Apr May Jun Jul Aug       Sep       Oct       Nov
## 2022                                      0.3968852 0.4078307 0.4670014
## 2023 0.5417144
##           Dec
## 2022 0.5080391
## 2023
```

NOTA: Recuerde que estas predicciones están es cientos de miles y se deben multiplicar por 100000 para obtener los valores reales.

Análisis de observaciones atípicas (outliers)

Como vimos antes, las series de tiempo están afectadas por eventos tales como: días de fiesta, huelgas, promociones, cambios políticos, errores de medición o registros erróneos, entre otros.

Cuando el período de ocurrencia de estos eventos externos (exógenos) es desconocido, a las observaciones generadas por ellos se les llama observaciones atípicas u **outliers**.

Estas observaciones atípicas pueden afectar la inferencia del modelo haciéndola poco confiable o aún inválida. Por esta razón, se deben aplicar procedimientos que permitan su detección y remoción de sus efectos.

Análisis de observaciones atípicas (outliers)

En la literatura se presentan varios procedimientos como el de Chang, Tiao y Chen (1988)[1]:

1. Identificar un modelo ARIMA y estimarlo asumiendo que no hay observaciones atípicas (por ejemplo, usar la función **auto.arima** con la serie completa).
2. El procedimiento de detección de outliers se aplica a la serie de residuales para verificar si están presentes (**checkresiduals**).
3. Si es así, se estima un modelo ajustado, el cual incluye los efectos de los outliers como intervenciones.
4. La detección y el ajuste continúa en la medida que sea necesario, después que el modelo intervenido es estimado.

---

[1] Chang, I., Tiao, G.C. and Chen, C. (1988) Estimation of Time Series Parameters in the Presence of Outliers. Technometrics, 30, 193-204.

El procedimiento anterior es útil en cierta medida, pero puede presentar algunas deficiencias:

- Si el modelo inicial está mal identificado, pueden aparecer observaciones atípicas.
- La eficiencia del procedimiento de detección de outliers puede estar afectada por los sesgos en los parámetros debido a la presencia de outliers.
- Algunos outliers pueden estar enmascarados y no ser identificados.
- Se pueden detectar algunos outliers espúreos (falsos).

Análisis de observaciones atípicas (outliers)

Para evitar los problemas anteriores, Chen, Liu y Lon-Mu (1993)[2], proponen un procedimiento iterativo para la estimación conjunta de los parámetros del modelo y de los efectos de los outliers:

- Se inicia como antes, con un modelo identificado y con estimadores potencialmente sesgados, debido a la presencia de outliers.
- A continuación, a los residuales del modelo estimado se aplica un procedimiento iterativo de detección de outliers.
- A continuación la serie original es ajustada (para remover los efectos de los outliers) de acuerdo a los tipos de outliers detectados.

---

[2] Chen, C. and Liu, Lon-Mu (1993). "Joint Estimation of Model Parameters and Outlier Effects in Time Series". Journal of the American Statistical Association, 88(421), pp. 284-297.

Análisis de observaciones atípicas (outliers)

- Después se estima el modelo para la serie ajustada y se examinan los residuales de nuevo. Los tres pasos de
  1. detección de outliers,
  2. ajuste de la serie por los efectos de los outliers, y
  3. estimación de los parámetros de la serie ajustada, son iterados hasta que no se encuentren más outliers.
- En este momento, la información acumulada de los outliers es empleada para estimar conjuntamente los efectos de los outliers y producir una serie final de observaciones ajustadas.
- Después de este paso, se estima el modelo con la serie ajustada para obtener las estimaciones finales de los parámetros.
- Finalmente, el procedimiento de detección de outliers es aplicado a la serie de residuales de la serie original usando las estimaciones finales de los parámetros del modelo.

Análisis de observaciones atípicas (outliers)

Este procedimiento difiere del anterior en varios aspectos.

1. La detección de outliers se hace iterativamente basada tanto en los residuales ajustados como en las observaciones ajustadas. Es decir, una vez una observación atípica es detectada, su efecto puede ser removido de la serie observada de la misma forma que puede ser removido de los residuales del modelo estimado. Al ajustar la serie observada, este procedimiento evita la necesidad de formular y estimar un modelo de intervención.

Análisis de observaciones atípicas (outliers)

Este procedimiento difiere del anterior en varios aspectos.

1. La detección de outliers se hace iterativamente basada tanto en los residuales ajustados como en las observaciones ajustadas. Es decir, una vez una observación atípica es detectada, su efecto puede ser removido de la serie observada de la misma forma que puede ser removido de los residuales del modelo estimado. Al ajustar la serie observada, este procedimiento evita la necesidad de formular y estimar un modelo de intervención.
2. Los outliers son detectados basados en estimadores robustos de los parámetros del modelo. Finalmente, en este procedimiento, los efectos de los outliers son estimados conjuntamente usando regresión múltiple. Como resultado, este procedimiento produce estimaciones más robustas de los parámetros del modelo, y reduce los outliers espúreos y el efecto de enmascaramiento en el proceso de detección.

Análisis de observaciones atípicas (outliers)

En el RStudio existe una función llamada **tso** dentro del paquete **tsoutliers**. Entre los outliers que detecta se destacan:

- **Observación atípica Aditiva (AO)**: Es un evento que afecta a una serie de tiempo solamente durante un período. Si suponemos que el outlier ocurre en el momento $t = T$, la serie observada se puede representar como:

  $$X_t = w_t + \omega_a P_t^{(T)}$$

  donde ω_a es un parámetro a ser estimado.

Análisis de observaciones atípicas (outliers)

En el RStudio existe una función llamada **tso** dentro del paquete **tsoutliers**. Entre los outliers que detecta se destacan:

- **Observación atípica Aditiva (AO)**: Es un evento que afecta a una serie de tiempo solamente durante un período. Si suponemos que el outlier ocurre en el momento $t = T$, la serie observada se puede representar como:

$$X_t = w_t + \omega_a P_t^{(T)}$$

  donde ω_a es un parámetro a ser estimado.

- **Observación atípica Innovativa (IO)**: Es un evento cuyo efecto es propagado de acuerdo al modelo ARIMA del proceso. Si suponemos que el outlier ocurre en el momento $t = T$, la serie observada se puede representar como:

$$X_t = \frac{\theta(B)}{\phi(B)} w_t + \frac{\theta(B)}{\phi(B)} \omega_i P_t^{(T)}$$

  donde ω_i es un parámetro a ser estimado.

- **Observación atípica de cambio de nivel (LS)**: Es un evento que afecta permanentemente la serie a partir de un período dado. Si suponemos que el outlier ocurre en el momento $t = T$, la serie observada se puede representar como:

$$X_t = w_t + \frac{1}{1-B}\omega_I P_t^{(T)} \quad \text{ó} \quad X_t = w_t + \omega_I S_t^{(T)}$$

Análisis de observaciones atípicas (outliers)

- **Observación atípica de cambio de nivel (LS)**: Es un evento que afecta permanentemente la serie a partir de un período dado. Si suponemos que el outlier ocurre en el momento $t = T$, la serie observada se puede representar como:

$$X_t = w_t + \frac{1}{1-B}\omega_I P_t^{(T)} \quad \text{ó} \quad X_t = w_t + \omega_I S_t^{(T)}$$

- **Observación atípica de cambio temporal (TC)**: Es un evento que impacta inicialmente la serie y luego desaparece gradualmente. Si suponemos que el outlier ocurre en el momento $t = T$, la serie observada se puede representar como:

$$X_t = w_t + \frac{1}{(1-\delta B)}\omega_c P_t^{(T)}, \quad \text{para} \quad 0 < \delta < 1$$

Análisis de observaciones atípicas (outliers)

El Efecto de las observaciones atípicas sobre la función de autocorrelación muestral Es bien conocido que las observaciones atípicas pueden influenciar fuertemente la ACF muestral y por tanto afectar la identificación de los modelos de series de tiempo.

Los distintos tipos de observaciones atípicas pueden tener efectos cualitativamente diferentes.

En muestras grandes la ACF muestral puede resultar seriamente afectada ante la existencia de observaciones atípicas AO, LS o TC:

Análisis de observaciones atípicas (outliers)

El Efecto de las observaciones atípicas sobre la función de autocorrelación muestral Es bien conocido que las observaciones atípicas pueden influenciar fuertemente la ACF muestral y por tanto afectar la identificación de los modelos de series de tiempo.

Los distintos tipos de observaciones atípicas pueden tener efectos cualitativamente diferentes.

En muestras grandes la ACF muestral puede resultar seriamente afectada ante la existencia de observaciones atípicas AO, LS o TC:

1. Una observación atípica aditiva (AO) grande puede anular completamente la información de la ACF muestral.

II. Cuando $T >> k$, siendo T el período de ocurrencia de la observación atipica y k el orden del coeficiente de autocorrelación muestral, una observación atípica de cambio de nivel (LS) empuja la ACF muestral hacia 1, la cual es la cota no estacionaria de la función de autocorrelación. La ACF muestral decae lentamente a medida que k crece.

II. Cuando $T >> k$, siendo T el período de ocurrencia de la observación atipica y k el orden del coeficiente de autocorrelación muestral, una observación atípica de cambio de nivel (LS) empuja la ACF muestral hacia 1, la cual es la cota no estacionaria de la función de autocorrelación. La ACF muestral decae lentamente a medida que k crece.

III. Para el caso de una observación atípica de cambio temporal (TC), los valores de la ACF muestral son dominados por el factor de decaimiento δ.

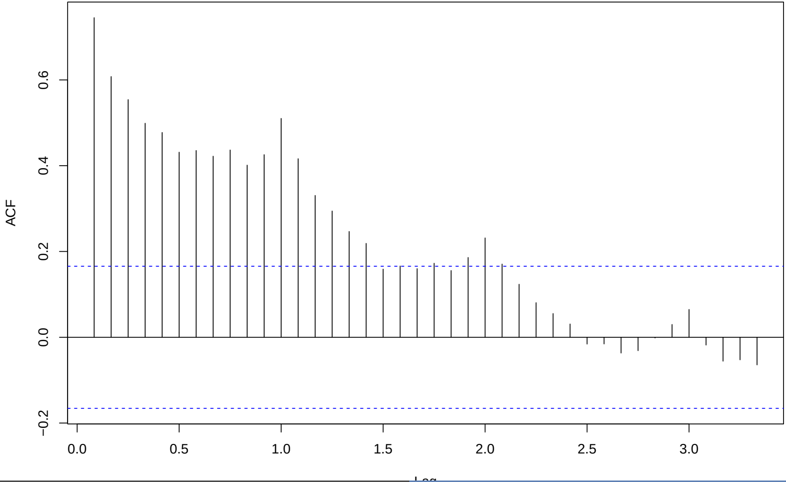
II. Cuando $T >> k$, siendo T el período de ocurrencia de la observación atipica y k el orden del coeficiente de autocorrelación muestral, una observación atípica de cambio de nivel (LS) empuja la ACF muestral hacia 1, la cual es la cota no estacionaria de la función de autocorrelación. La ACF muestral decae lentamente a medida que k crece.

III. Para el caso de una observación atípica de cambio temporal (TC), los valores de la ACF muestral son dominados por el factor de decaimiento δ.

IV. Finalmente, la existencia de una observación atípica innovativa (IO), no altera el cálculo de la ACF muestral puesto ella tiende a ρ_k cuando n y ω_i son grandes.

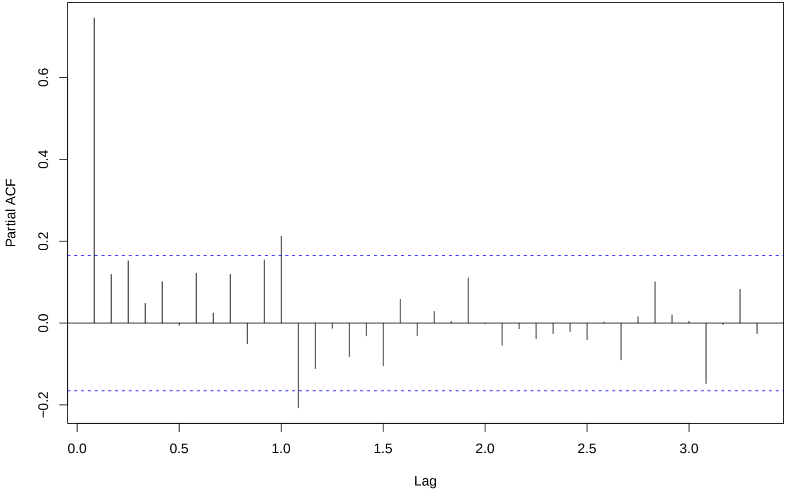
Retomemos el ejemplo relacionado con la mezcla de cemento:

```
ts_mezcla %>% acf(lag.max = 40)
```



Análisis de observaciones atípicas (outliers)

```
ts_mezcla %>% pacf(lag.max = 40)
```



Análisis de observaciones atípicas (outliers)

Realicemos un análisis de los valores outlier:

```r
require(tsoutliers)
mod_outliers <- tso(ts_mezcla2, delta=0.7)
mod_outliers
```

```
## Series: ts_mezcla2
## Regression with ARIMA(0,1,1)(2,0,0)[12] errors
##
## Coefficients:
##           ma1    sar1    sar2    TC112    TC125
##       -0.5982  0.4033  0.2272  -4.0295  -1.4239
## s.e.   0.0818  0.0809  0.0859   0.3675   0.3769
##
## sigma^2 = 0.2114:  log likelihood = -89.46
## AIC=190.91   AICc=191.55   BIC=208.52
##
## Outliers:
##   type ind    time coefhat   tstat
## 1   TC 112 2020:04  -4.030 -10.966
## 2   TC 125 2021:05  -1.424  -3.778
```

Análisis de observaciones atípicas (outliers)

Los dos valores atípicos que fueron detectados son tipo cambio temporal en las posiciones 112 y 125, en las fechas abril de 2020 y mayo de 2021. Sin embargo, note que el valor de δ fue fijado en 0.7 (así viene por defecto en la función **tso**).

Análisis de observaciones atípicas (outliers)

Los dos valores atípicos que fueron detectados son tipo cambio temporal en las posiciones 112 y 125, en las fechas abril de 2020 y mayo de 2021. Sin embargo, note que el valor de δ fue fijado en 0.7 (así viene por defecto en la función **tso**). Si intentamos con otro valor:

```
mod_outliers <- tso(ts_mezcla2, delta=0.3)
mod_outliers
```

```
## Series: ts_mezcla2
## Regression with ARIMA(2,1,0)(1,1,0)[12] errors
##
## Coefficients:
##           ar1      ar2     sar1    LS111   TC112    TC125
##       -0.8797  -0.3069  -0.4597  -1.5314  -4.204  -1.3247
## s.e.   0.0921   0.0912   0.0812   0.2754   0.318   0.2609
##
## sigma^2 = 0.1391:  log likelihood = -53.71
## AIC=121.43   AICc=122.37   BIC=141.34
##
## Outliers:
##   type ind    time coefhat   tstat
## 1   LS 111 2020:03  -1.531  -5.562
## 2   TC 112 2020:04  -4.204 -13.221
## 3   TC 125 2021:05  -1.325  -5.078
```

Análisis de observaciones atípicas (outliers)

Nuevamente, intentamos con otro valor:

```
mod_outliers <- tso(ts_mezcla2, delta=0.5)
mod_outliers
```

```
## Series: ts_mezcla2
## Regression with ARIMA(2,1,0)(2,1,0)[12] errors
##
## Coefficients:
##           ar1      ar2     sar1     sar2     AO67    LS111    TC112    TC125
##       -0.7194  -0.2072  -0.5087  -0.2459  -0.8817  -1.7321  -3.6644  -1.4636
## s.e.   0.0919   0.0916   0.0954   0.0984   0.2695   0.3162   0.3380   0.2780
##
## sigma^2 = 0.1418:  log likelihood = -54.16
## AIC=126.32   AICc=127.85   BIC=151.91
##
## Outliers:
##   type ind    time coefhat   tstat
## 1   AO  67 2016:07 -0.8817  -3.271
## 2   LS 111 2020:03 -1.7321  -5.478
## 3   TC 112 2020:04 -3.6644 -10.841
## 4   TC 125 2021:05 -1.4636  -5.265
```

Conclusión: De los tres modelos, el que tiene el menor AIC es el relacionado con $\delta = 0.3$. **TAREA:** Ajuste los tres modelos con la función **arimax** y realice un análisis de los residuales de cada uno.