Series Temporales

Daniel Vélez Serrano
Marzo 2022

Bibliografía

1. Peña, D. (2010), "Análisis de series temporales", Alianza Editorial

2. Aznar, A., Trívez, F.J. (1993), "Métodos de predicción en economía II, Análisis de Series Temporales", Ariel Economía

3. Matilla, M., Pérez, P., Sanz, B. (2013), "Econometría y predicción", Mc Graw Hill (UNED)

Índice

1 Introducción al concepto de serie temporal

Introducción

- En términos coloquiales, una serie temporal, es un conjunto de observaciones registradas en el tiempo.

- Se trata de construir un modelo que permita explicar dichas observaciones e identificar un patrón de comportamiento a partir del cual realizar predicciones.

- Cuando se plantea predecir el comportamiento futuro de una serie temporal, es preciso determinar.

  - La **unidad temporal** que se va a manejar: hora, día, semana, mes, año, etc.

  - El **horizonte de predicción** que se plantea predecir: día, semana, mes, año, década, etc.

  - ¿Cómo se va a **medir el error**?: En términos absolutos o relativos/porcentuales, en términos medios o medianos, etc.

Definición de proceso estocástico

- Un **proceso estocástico** es un modelo matemático que permite describir la evolución **aleatoria** de un sistema a lo largo del **tiempo**

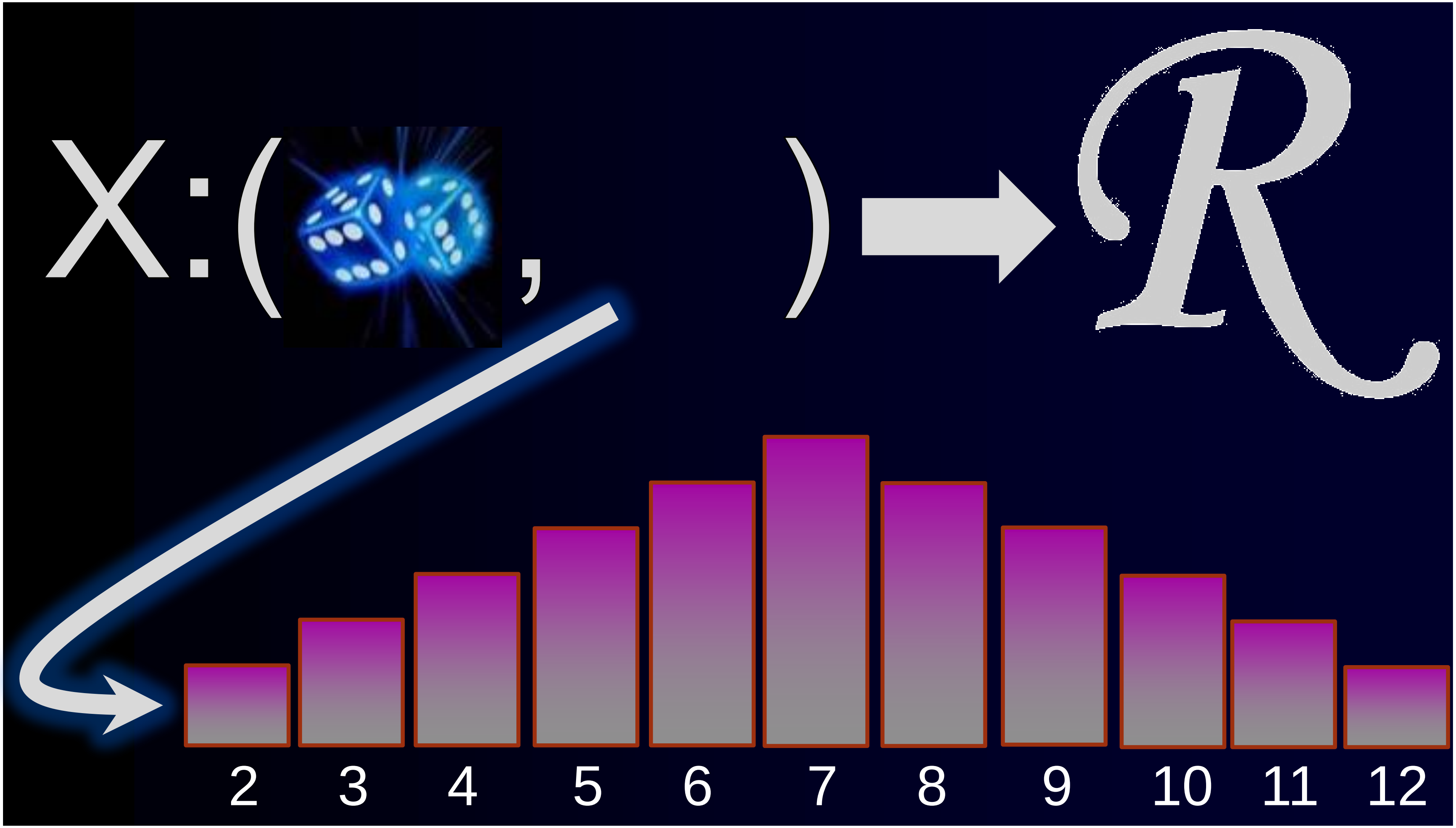
- Formalmente, es una aplicación:

$$X : \Omega x T \to S$$
$$(\omega, t) \to X(\omega, t)$$

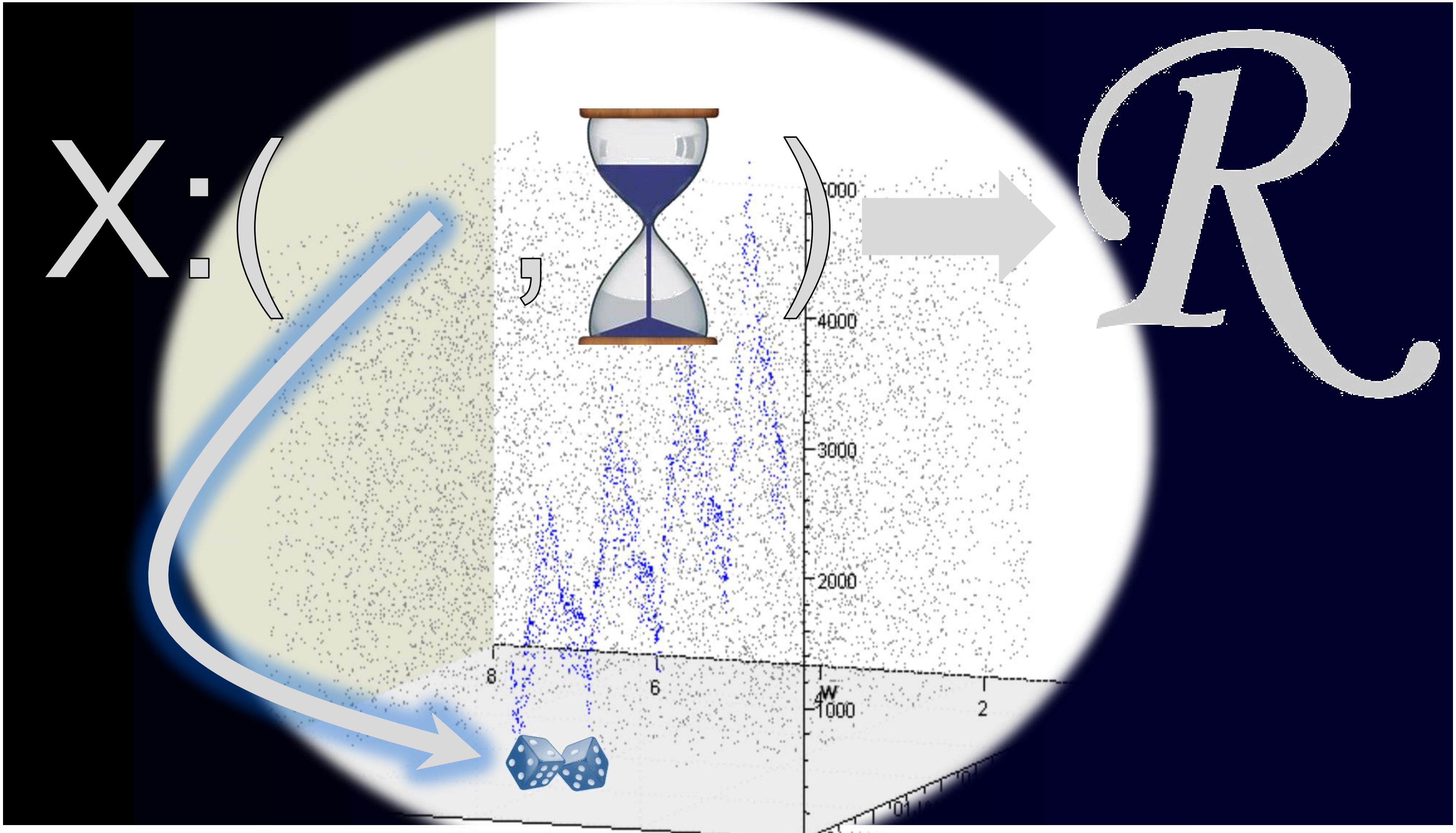
- T habitualmente representa el **tiempo**:

$T = \{0,1,\ldots\} \Rightarrow \{X_n, n \geq 0\}$ **Proceso en tiempo discreto**

$T = [0,\infty) \Rightarrow \{X(t), t \geq 0\}$ **Proceso en tiempo continuo**



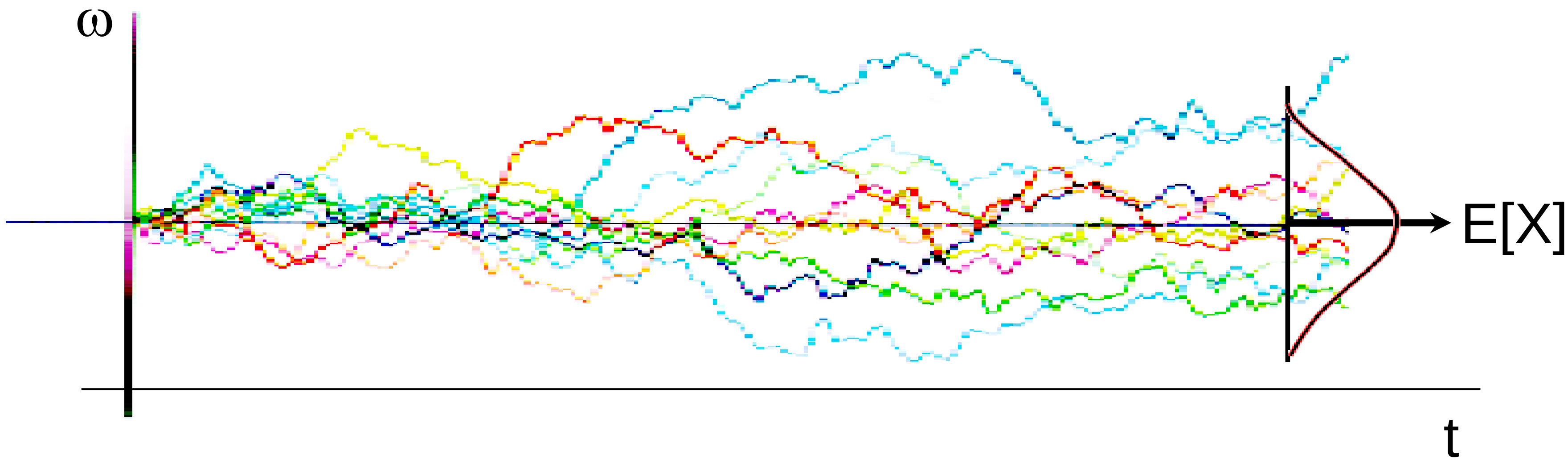
FIJADO UN INSTANTE TEMPORAL → VARIABLE ALEATORIA



FIJADA UNA REALIZACIÓN DEL PROCESO → SERIE TEMPORAL

Definición formal de serie temporal

- Así, dentro de un contexto matemático, una serie temporal se define como **una realización de un proceso estocástico.**



varias realizaciones de un proceso estocástico (varias series temporales)

- El objetivo que se plantea es **inferir el proceso estocástico que ha podido generar el conjunto de observaciones que definen la serie temporal** para poder conocer qué puede "esperarse" en el futuro (predicción).

Caracterización de una serie temporal

TEOREMA (KOLMOGOROV)

- "Un proceso estocástico queda caracterizado por sus F.D finito-dimensionales"

$$F(X(t_1), X(t_2), \ldots, X(t_N)), \quad t_1, t_2, \ldots, t_N \in Z, n \in N$$

- Para determinar el mecanismo generador del proceso (serie) se debe conocer:
  - la distribución conjunta
  - las marginales: $F(X_t)\ \forall t$
  - las distribuciones que permiten modelizar las relaciones entre variables contiguas:

    $F(X_t, X_{t+1})\ \forall t$
  - etc.

Problema: Solo se dispone de una observación por instante temporal, lo que hace inviable la estimación de dichas distribuciones.

Solución: Asumir que las distribuciones son estables (estacionarias) en el tiempo para que las distribuciones asociadas a diferentes instantes sean comparables.

Procesos estacionarios

- Un proceso es **estacionario en sentido estricto** si el comportamiento de una colección de v.a's solo depende de su posición relativa, no del instante "t".

$$F(X(t_1),\ldots,X(t_N)) = F(X(t_1+k),\ldots,X(t_N+k))\ \forall k \in Z, \forall n \in N, \forall t_1,\ldots,t_N \in Z$$

En particular, todas las marginales son iguales: $F(X(t)) = F(X(t+k))\ \forall t, \forall k$

Problema: La condición es demasiado restrictiva.

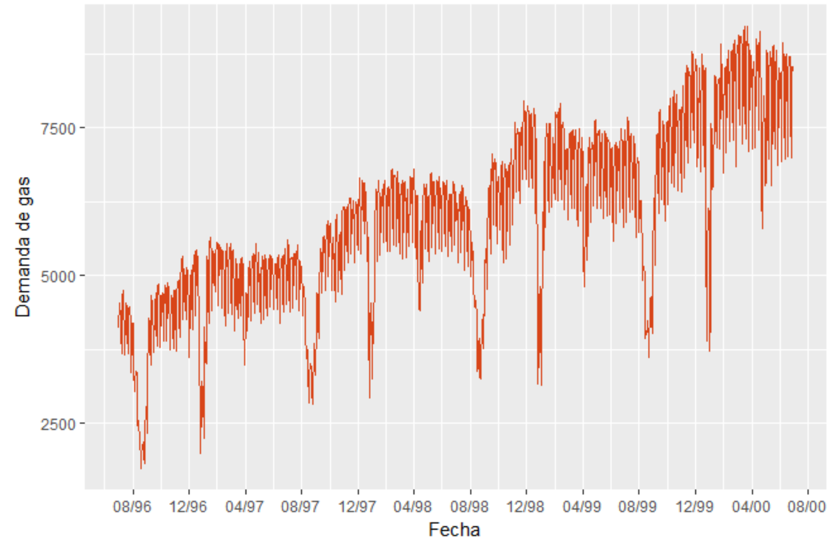
Se suele asumir una relajación: **estacionariedad débil.**

- Un proceso es **estacionario en sentido débil** si: $\begin{cases} 1a.\ E[X_t] = \mu < +\infty, \forall t \\ 1b.\ V[X_t] = \gamma_0 < +\infty, \forall t \\ 1c.\ \text{cov}[X_t, X_{t+k}] = \gamma_k, \forall t, \forall k \end{cases}$

- Un caso particular es un proceso de **ruido blanco** si : $\begin{cases} 2a.\ E[a_t] = 0 < +\infty, \forall t \\ 2b.\ V[a_t] = \sigma_a^2 < +\infty, \forall t \\ 2c.\ \text{cov}[a_t, a_{t+k}] = 0, \forall t, \forall k \neq 0 \end{cases}$

Ejemplo motivador:
Estimación a corto plazo de la demanda industrial de gas

- Se desea estimar la demanda de gas de una zona en la que el consumo es de carácter fundamentalmente industrial (archivo *demandaIndustrialGas.csv*).



- Las características fundamentales de esta serie son:
  - Alto consumo industrial en días laborables.
  - Bajo consumo industrial en fines de semana, días festivos, puentes y periodos vacacionales.

- Utilizar el periodo 01jul1996-30jun1999 para ajustar el modelo.

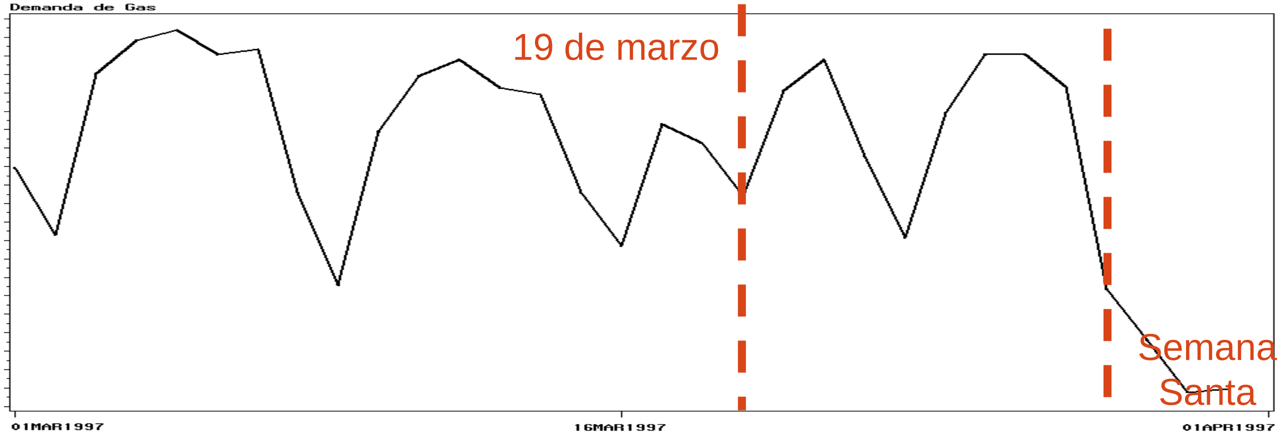
- Validar los resultados sobre el periodo 01jul1999-30jun2000: ofrecer una medida de error medio y mediano mensual a distintos horizontes (d+1, d+2,…, d+10).

Análisis descriptivo de la serie

- Un análisis gráfico de la serie permite identificar sus principales componentes.

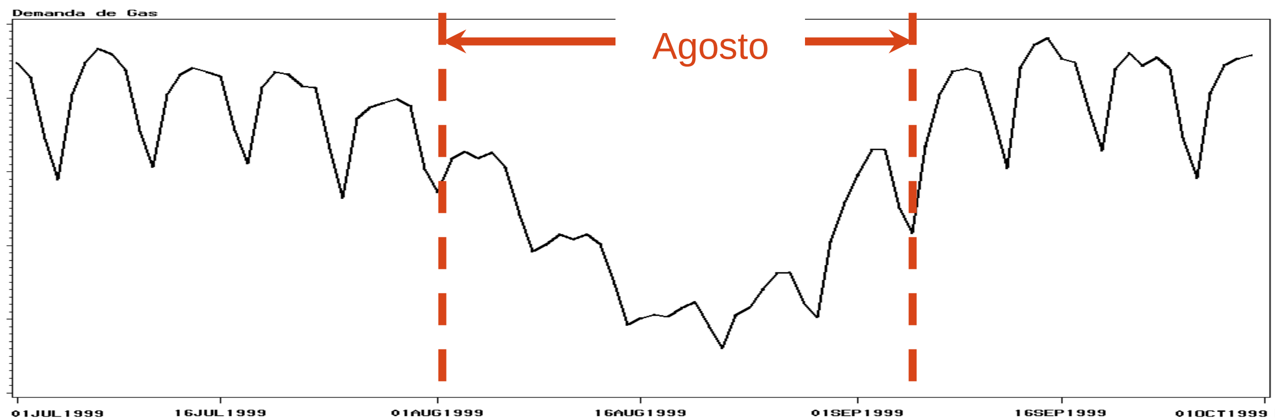
Análisis descriptivo de la serie

Algunas otras características de la serie son:



Presencias de datos atípicos tipo pulso: rompen de manera puntual el patrón de la serie.

Ejemplos: Festivos y puentes.



Presencia de datos atípico tipo escalón: rompen a lo largo de un periodo el patrón de la serie.

Ejemplos: Periodos vacacionales

2 Modelos ARMA

Procesos/Modelos ARMA

- Todo proceso estacionario se puede descomponer como suma de dos procesos mutuamente incorrelados: uno lineal determinístico y otro puramente indeterminístico":

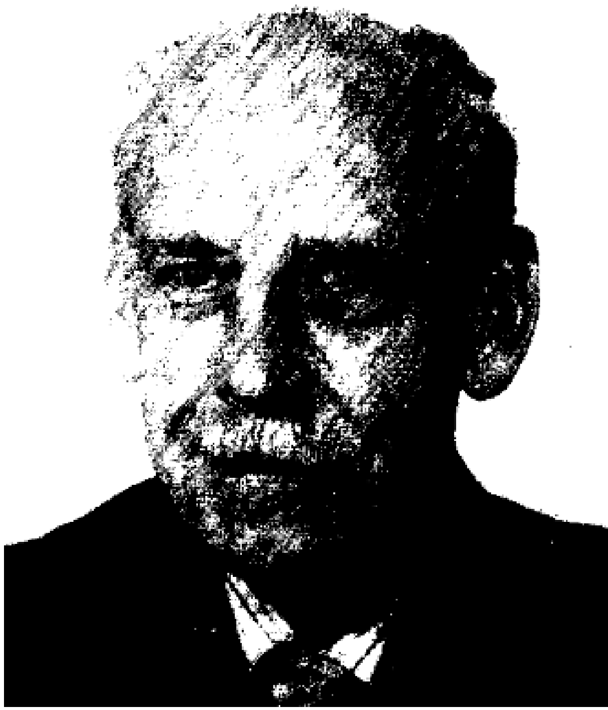
$$Y_t = D_t + X_t \text{ con } corr(D_t, X_{t'}) = 0$$

- Los **procesos determinísticos (o predecibles)** son aquéllos cuyos valores futuros se **pueden predecir de forma exacta** (sin duda, sin varianza) a partir de los anteriores (presente y pasado).

- Los **procesos indeterminísticos** son aquéllos que no se pueden predecir de forma exacta.

  Los modelos ARMA son un caso particular de ellos.

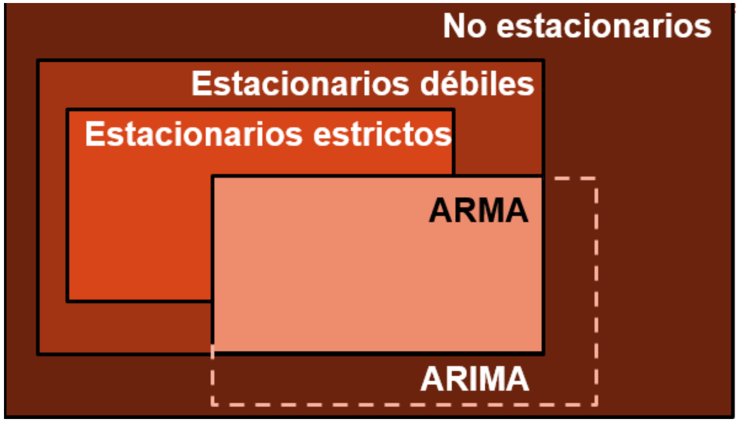
(Modelo ARMA: $X_t = \frac{\theta(B)}{\varphi(B)} a_t$)



Procesos/Modelos ARMA

- Dada una serie temporal, el objetivo es hacerla estacionaria para asumir esa estabilidad que permita hacer que todos los instantes sean comparables.

- Una vez que el proceso (serie) es estacionario, se busca algún tipo de modelo adecuado para su caracterización: **“los procesos ARMA estacionarios son modelizables mediante modelos ARMA”.**



ARMA(p,q): $$(1-\phi_1 B-\ldots-\phi_p B^p)X_t=(1-\theta_1 B-\ldots-\theta_q B^q)a_t$$

Procesos/Modelos AR (autorregresivos)

- **Parte AUTORREGRESIVA.-** muestra la dependencia del dato real con su propio pasado. Es una regresión de la variable en sí misma (autorregresión).

AR(p): $$X_t = \mu + \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + a_t$$

$$(\beta_0 + \beta_1 X_{t-1} + \ldots + \beta_p X_{t-p} + \varepsilon)$$

- La **condición de estacionariedad** se exige sobre la parte autorregresiva del modelo y es una condición necesaria para el ajuste de modelos ARMA. Establece que el polinomio

$$1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p = 0$$

debe tener sus raíces fuera del círculo unidad.

Supóngase un AR(1) **no estacionario** en el que $|\phi_1| > 1$, entonces las predicciones se dispararían:

dando lugar a un comportamiento **EXPLOSIVO:**

$$X_{t+1} = \phi_1 X_t, X_{t+1} = \phi_1 \; X_{t+1} = \phi_1^{\,2} X_t, \ldots$$

Procesos/Modelos I (integrados)

- Si el proceso no es estacionario se puede hacer estacionario a través del operador diferencia $(1\text{-}B)^d$.

- Este operador se incorpora sobre la propia serie (no sobre el proceso residual) y de hecho, puede verse como una versión extrema de los modelos AR.

- Los procesos **INTEGRADOS** son aquéllos que precisan de la realización de **DIFERENCIAS** para ser estacionarios.

Proceso Integrado de orden d: $$(1 - B)^d X_t = a_t$$

"precisa de d diferencias para ser estacionario".

- Si X_t es un proceso ARIMA(p,d,q), entonces $(1-B)^d X_t$ es un ARMA(p,q).

$$\phi(B)(1-B)^d X_t = \mu + \theta(B) a_t$$

Procesos/Modelos MA ("*moving averages*")

- **Parte de MEDIAS MÓVILES.-** muestra la dependencia del dato real con el pasado del proceso de error (media móvil de la serie de los errores). Permite al modelo "aprender de sus errores".

MA(q): $X_t = \mu - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} + a_t$

- Los procesos MA son siempre estacionarios.

$$\Rightarrow E[X_t] = E[a_t] - \theta_1 E[a_{t-1}] - \theta_2 E[a_{t-2}] - \ldots - \theta_q E[a_{t-q}] = 0$$

$$\Rightarrow V[X_t] = V[a_t] + \theta_1^2 V[a_{t-1}] + \theta_2^2 V[a_{t-2}] + \ldots + \theta_q^2 V[a_{t-q}] =$$

$$(1 + \theta_1^2 + \theta_2^2 + \ldots + \theta_q^2)\sigma_a^2, \, siendo \, V[a_t] = \sigma_a^2$$

$$\Rightarrow \gamma_k = E[X_t X_{t-k}] =$$

$$E[(a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q})(a_{t-k} - \theta_1 a_{t-k-1} - \ldots - \theta_q a_{t-k-q})] =$$

$$= \begin{cases} (-\theta_k + \theta_1\theta_{k+1} + \ldots + \theta_{q-k}\theta_q)\sigma_a^2 & k = 1, 2, \ldots, q \\ 0 & \forall k > q \end{cases}$$

Procesos/Modelos MA (“*moving averages*”)

- Existe otra condición, denominada **condición de invertibilidad** (deseable pero no necesaria para el ajuste de modelos ARMA). Su incumplimiento lleva a que **datos más alejados tengan más peso en la predicción que datos más recientes.**

- Esta condición la cumplen todos los procesos AR pero no así los modelos MA (ni en consecuencia los ARMA), siendo necesario para ello, que $1-\theta_1 z-\theta_2 z^2-\ldots-\theta_q z^q=0$ tenga todas sus raíces fuera del círculo unidad.

  Supóngase un MA(1) **no invertible**, es decir $|\theta_1|>1$, entonces **datos más alejados tendrían más peso en la predicción que datos más recientes.**

$$a_t = X_t + \theta_1 a_{t-1} = X_t + \theta_1\left(X_{t-1}+\theta_1 a_{t-2}\right) = X_t + \theta_1 X_{t-1} + \theta_1{}^2 a_{t-2} =$$
$$X_t + \theta_1 X_{t-1} + \theta_1^2\left(X_{t-2}+\theta_1 a_{t-3}\right) = X_t + \theta_1 X_{t-1} + \theta_1{}^2 X_{t-2} + \ldots$$
$$\Rightarrow X_t = a_t - \theta_1 X_{t-1} - \theta_1{}^2 X_{t-2} - \theta_1{}^3 X_{t-3} - \cdots$$

- Se establece una **dualidad** entre los procesos **AR y MA** respecto al cumplimiento de las **condiciones de estacionariedad e invertibilidad.**

Identificación de los órdenes de un ARMA

- El **objetivo** que se persigue es **identificar el proceso que subyace bajo la los datos**, lo cual consiste **en identificar los órdenes p y q** del modelo ARMA que generó la serie temporal.

- Las herramientas para identificar a estos procesos son las **funciones de autocorrelación simple (f.a.s) y parcial (f.a.p).**

- Los **correlogramas** permiten la representación de estas funciones que **solo tienen sentido dentro del ámbito de los procesos estacionarios** porque asumen que la correlación entre dos valores de la serie solo depende de su distancia, no del instante de tiempo al que van referidos.

- A partir de estos correlogramas, se puede **intuir los órdenes p y q** del modelo ARMA correspondiente. La **dualidad** entre los procesos **AR** y **MA** se vuelve a poner de manifiesto respecto al **patrón** que presentan dichos modelos en uno y otro gráfico.

Función de autocorrelación simple

- El **coeficiente de correlación simple** (y por tanto la f.a.s) refleja la correlación entre la variable X en un instante y el valor retardado de la misma en k instantes anteriores.

- Ejemplo: correlación de la serie de demanda de gas consigo misma retardada 7 unidades:

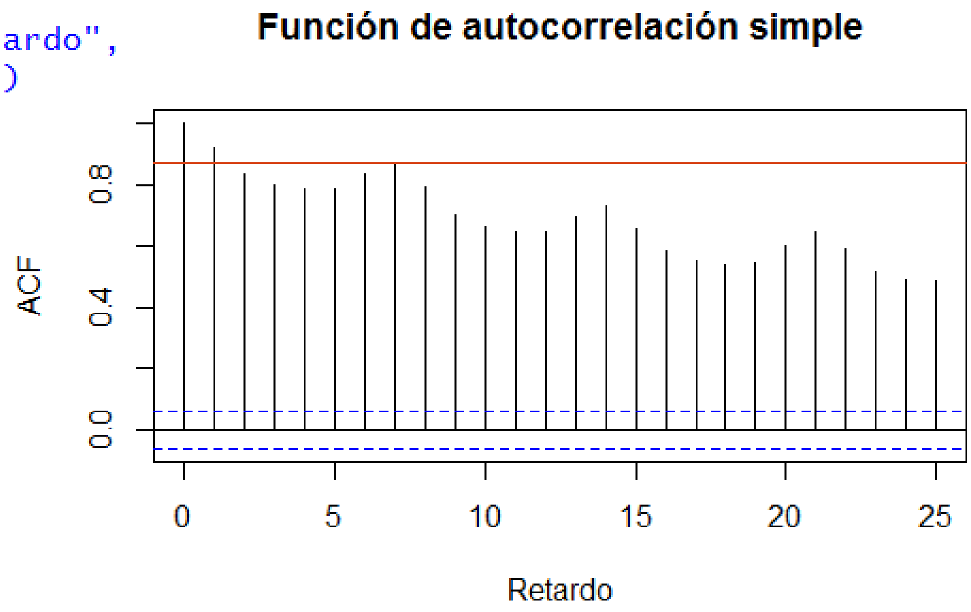
```
> DEMANDA_GAS <- ts(datos.train$DEMANDA_GAS)
> DEMANDA_GAS_8<-lag(ts(datos.train$DEMANDA_GAS), k = 7)
> comprobacionCorr <- as.data.frame(cbind(DEMANDA_GAS,DEMANDA_GAS_8))
> cor.test(comprobacionCorr$DEMANDA_GAS,comprobacionCorr$DEMANDA_GAS_8)

        Pearson's product-moment correlation

data:  comprobacionCorr$DEMANDA_GAS and comprobacionCorr$DEMANDA_GAS_8
t = 60.771, df = 1086, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8648198 0.8919041
sample estimates:
      cor
0.8790695
```

Función de autocorrelación simple

- El valor de la f.a.s. en el retardo 8 también se puede calcular con la función *acf*:

```
> comprobacionAcf <- acf(datos.train.ts, lag.max=25, plot=F)
> comprobacionAcf$acf[8]
[1] 0.8719139
```

```
> acf(datos.train.ts, lag.max = 25, xlab = "Retardo",
+      main= "Función de autocorrelación simple")
> abline(h=0.872,col='#d84519')
```



Función de autocorrelación simple

- **La máxima autocorrelación simple distinta de 0 determina el orden del MA.**

MA(1): $X_t = \theta_1 \boldsymbol{a}_{t-1} + \boldsymbol{a}_t$

$$\boldsymbol{cov}(\boldsymbol{X}_t, \boldsymbol{X}_{t-1}) = \boldsymbol{E}[(\theta_1 \boldsymbol{a}_{t-1} + \boldsymbol{a}_t)(\theta_1 \boldsymbol{a}_{t-2} + \boldsymbol{a}_{t-1})] = \theta_1 \boldsymbol{\sigma^2}_{\boldsymbol{a}}$$

$$\boldsymbol{cov}(\boldsymbol{X}_t, \boldsymbol{X}_{t-k}) = \boldsymbol{E}[(\theta_1 \boldsymbol{a}_{t-1} + \boldsymbol{a}_t)(\theta_1 \boldsymbol{a}_{t-k-1} + \boldsymbol{a}_{t-k})] = \boldsymbol{0}, \boldsymbol{k} > \boldsymbol{1}$$

MA(2): $X_t = \theta_1 \boldsymbol{a}_{t-1} + \theta_2 \boldsymbol{a}_{t-2} + \boldsymbol{a}_t$

$$\boldsymbol{cov}(\boldsymbol{X}_t, \boldsymbol{X}_{t-1}) = \boldsymbol{E}[(\theta_1 \boldsymbol{a}_{t-1} + \theta_2 \boldsymbol{a}_{t-2} + \boldsymbol{a}_t)(\theta_1 \boldsymbol{a}_{t-2} + \theta_2 \boldsymbol{a}_{t-3} + \boldsymbol{a}_{t-1})] = \theta_1 \boldsymbol{\sigma^2}_{\boldsymbol{a}} + \theta_1 \theta_2 \boldsymbol{\sigma^2}_{\boldsymbol{a}}$$

$$\boldsymbol{cov}(\boldsymbol{X}_t, \boldsymbol{X}_{t-2}) = \boldsymbol{E}[(\theta_1 \boldsymbol{a}_{t-1} + \theta_2 \boldsymbol{a}_{t-2} + \boldsymbol{a}_t)(\theta_1 \boldsymbol{a}_{t-3} + \theta_2 \boldsymbol{a}_{t-4} + \boldsymbol{a}_{t-2})] = \theta_2 \boldsymbol{\sigma^2}_{\boldsymbol{a}}$$

$$\boldsymbol{cov}(\boldsymbol{X}_t, \boldsymbol{X}_{t-k}) = \boldsymbol{E}[(\theta_1 \boldsymbol{a}_{t-1} + \theta_2 \boldsymbol{a}_{t-2} + \boldsymbol{a}_t)(\theta_1 \boldsymbol{a}_{t-k-1} + \theta_2 \boldsymbol{a}_{t-k-2} + \boldsymbol{a}_{t-k})] = \boldsymbol{0}, \boldsymbol{k} > \boldsymbol{2}$$

Función de autocorrelación parcial

- Debe tenerse en cuenta que, parte de la **correlación** entre la variable X en un instante y un instante anterior, **puede deberse a la correlación existente de la variable con ella misma en instantes intermedios.**

- Ejemplo: Sea un modelo AR(2)

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} = \phi_1(\phi_1 X_{t-2} + \phi_2 X_{t-3}) + \phi_2 X_{t-2} = {\phi_1}^2 X_{t-2} + \phi_1\phi_2 X_{t-3} + \phi_2 X_{t-2}$$

  - Existe un efecto directo de X_{t-2} sobre a través de ϕ_2.

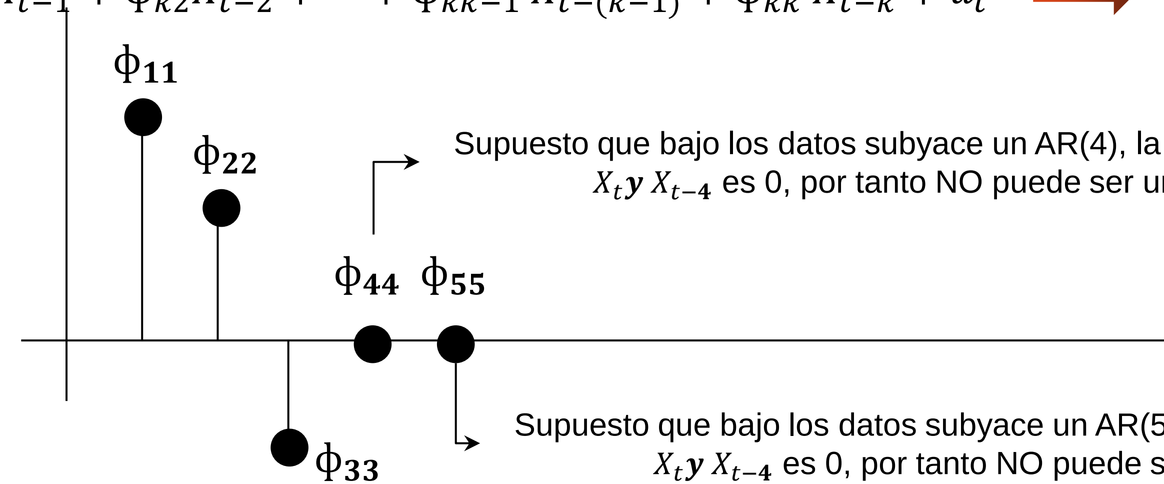
  - Existe un efecto indirecto de X_{t-2} sobre X_t a través de X_{t-1}, es decir, debido al hecho de que X_t y X_{t-1} están relacionados por ϕ_1. Si ϕ_1 =0, no existiría relación entre X_t y X_{t-1} y $X_t = 0^2 X_{t-2} + 0\phi_2 X_{t-3} + \phi_2 X_{t-2}$ por lo que solo existiría un efecto de X_{t-2} sobre X_t que sería el efecto directo.

  - Además, existe un efecto directo de X_{t-3} sobre X_t a través de X_{t-1} y X_{t-2}.

Función de autocorrelación parcial

- El **coeficiente de correlación parcial** (y así la f.a.p) calcula la correlación directa eliminando posibles dependencias asociadas a retardos intermedios.

$$X_t = \phi_{k1}X_{t-1} + \phi_{k2}X_{t-2} + \cdots + \phi_{kk-1}\, X_{t-(k-1)} + \phi_{kk}\, X_{t-k} + u_t \Rightarrow \rho_k^P = \phi_{kk}$$



Supuesto que bajo los datos subyace un AR(4), la correlación entre X_t **y** X_{t-4} es 0, por tanto NO puede ser un AR(4)

Supuesto que bajo los datos subyace un AR(5), la correlación entre X_t **y** X_{t-4} es 0, por tanto NO puede ser un AR(5)

- **La máxima correlación parcial distinta de 0 determina el orden del AR.**

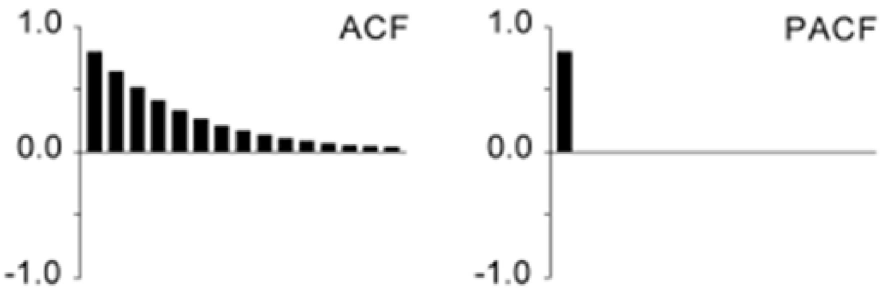
$$X_t = \phi_{11}X_{t-1} + \boldsymbol{u_t^1} \Rightarrow \rho_{\mathbf{1}}^P = \boldsymbol{\phi_{11}}$$

$$X_t = \phi_{21}X_{t-1} + \phi_{22}X_{t-2} + \boldsymbol{u_t^2} \Rightarrow \rho_{\mathbf{2}}^P = \boldsymbol{\phi_{22}}$$

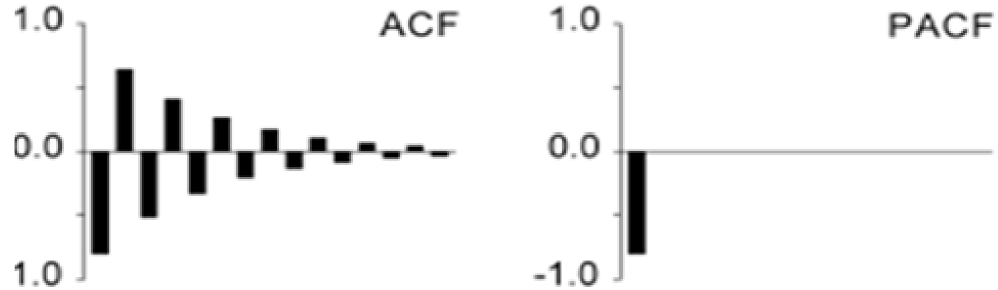
$$X_t = \phi_{31}X_{t-1} + \phi_{32}X_{t-2} + \boldsymbol{\phi_{33}}\, X_{t-3} + \boldsymbol{u_t^3} \Rightarrow \rho_{\mathbf{3}}^P = \boldsymbol{\phi_{33}}$$

Patrones característicos de un modelo AR

AR(1)



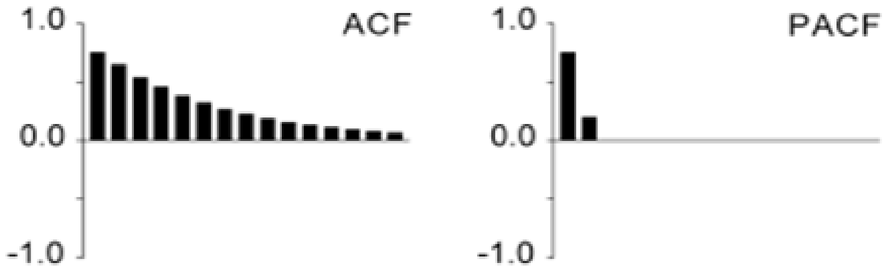
AR(1): $Y_t = 0.8Y_{t-1} + A_t$.



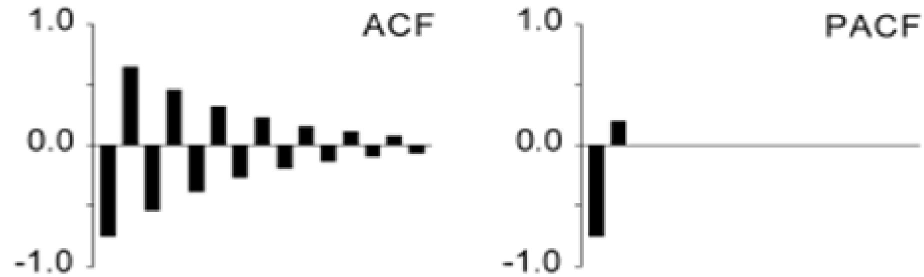
AR(1): $Y_t = -0.8Y_{t-1} + A_t$.

Patrones característicos de un modelo AR

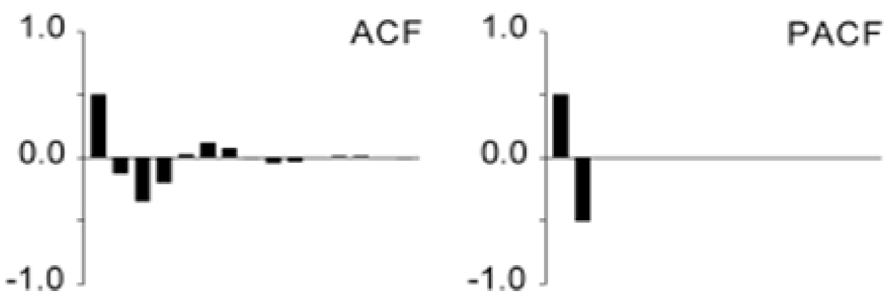
AR(2)



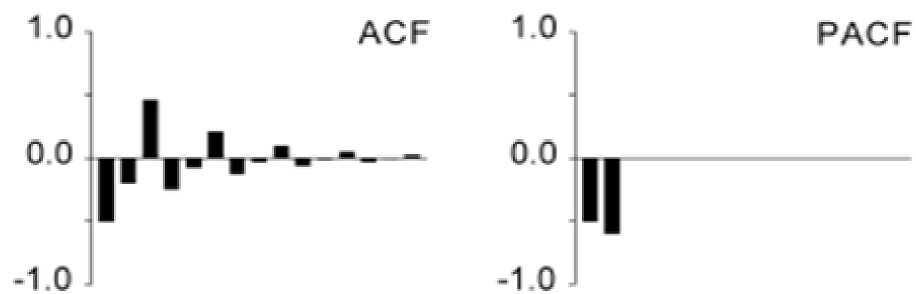
$$AR(2):\ Y_t = 0.6Y_{t-1} + 0.2Y_{t-2} + A_t\,.$$



$$AR(2):\ Y_t = -0.6Y_{t-1} + 0.2Y_{t-2} + A_t\,.$$



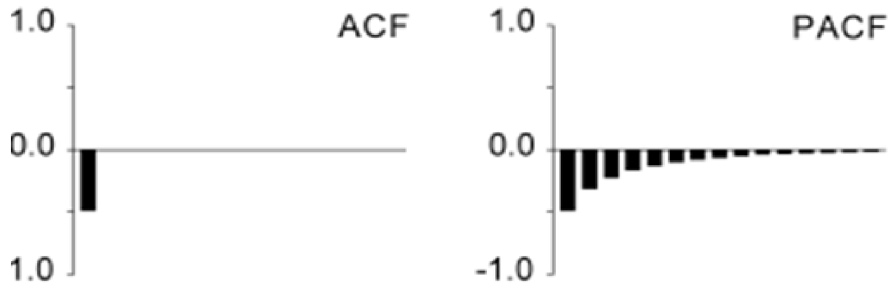
$$AR(2):\ Y_t = 0.75Y_{t-1} - 0.5Y_{t-2} + A_t\,.$$



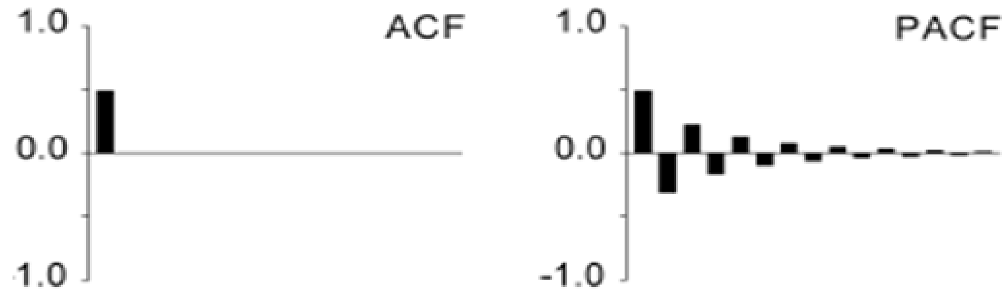
$$AR(2):\ Y_t = -0.8Y_{t-1} - 0.6Y_{t-2} + A_t\,.$$

Patrones característicos de un modelo MA

MA(1)



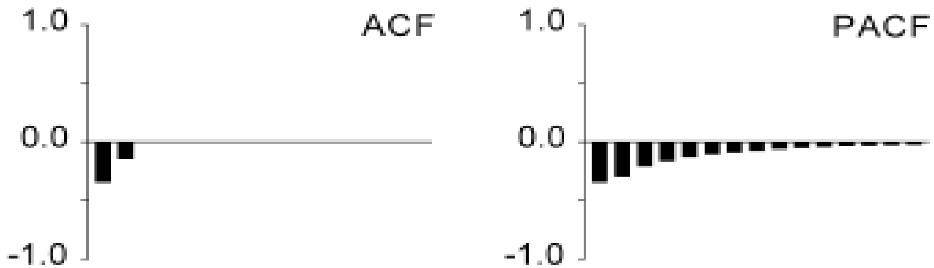
MA(1): $Y_t = A_t - 0.8A_{t-1}$.



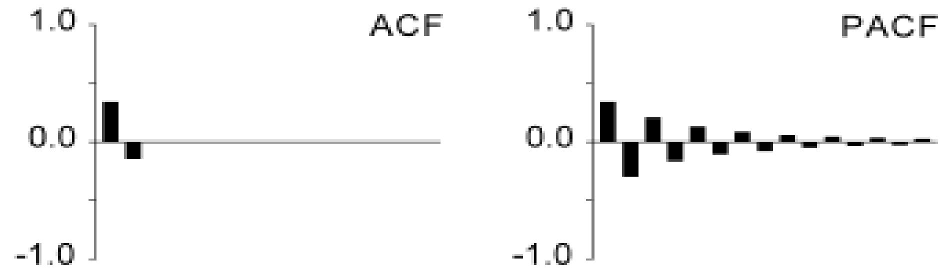
MA(1): $Y_t = A_t + 0.8A_{t-1}$.

Patrones característicos de un modelo MA

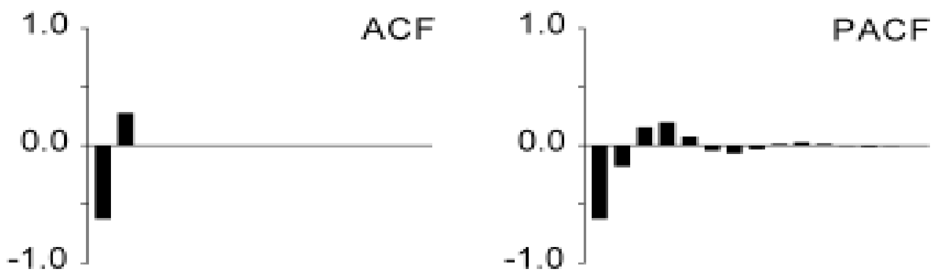
MA(2)



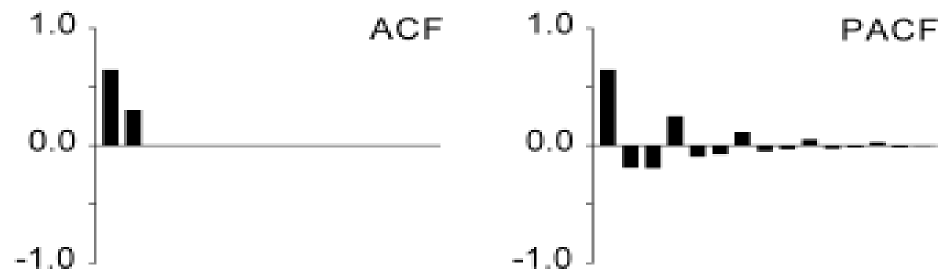
$$\text{MA}(2)\text{: } Y_t = A_t - 0.6A_{t-1} - 0.2A_{t-2}.$$



$$\text{MA}(2)\text{: } Y_t = A_t + 0.6A_{t-1} - 0.2A_{t-2}.$$



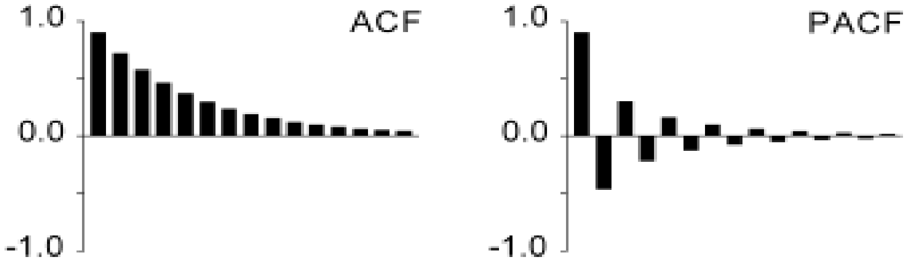
$$\text{MA}(2)\text{: } Y_t = A_t - 0.75A_{t-1} + 0.5A_{t-2}.$$



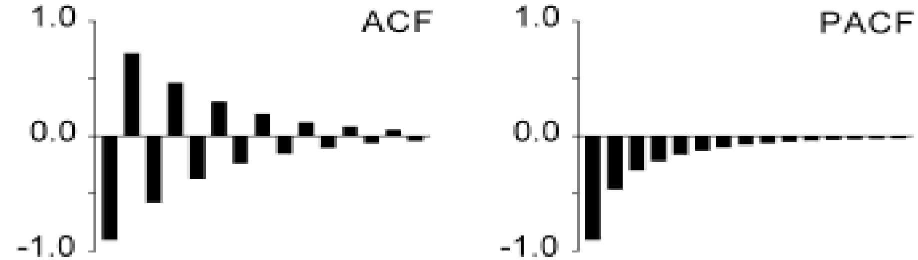
$$\text{MA}(2)\text{: } Y_t = A_t + 0.8A_{t-1} + 0.6A_{t-2}.$$

Patrones característicos de un modelo ARMA

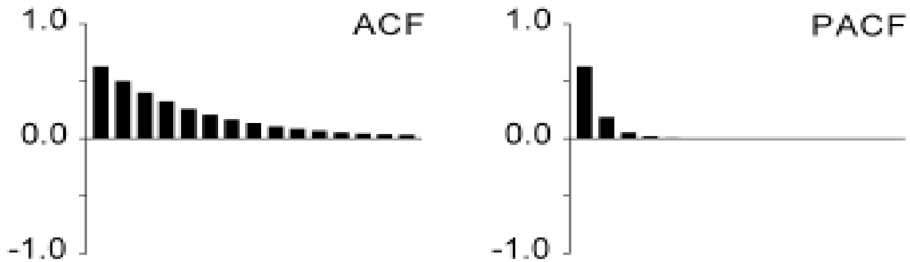
ARMA(1,1)



ARMA(1,1): $Y_t = 0.8Y_{t-1} + A_t + 0.8A_{t-1}$.



ARMA(1,1): $Y_t = -0.8Y_{t-1} + A_t - 0.8A_{t-1}$.



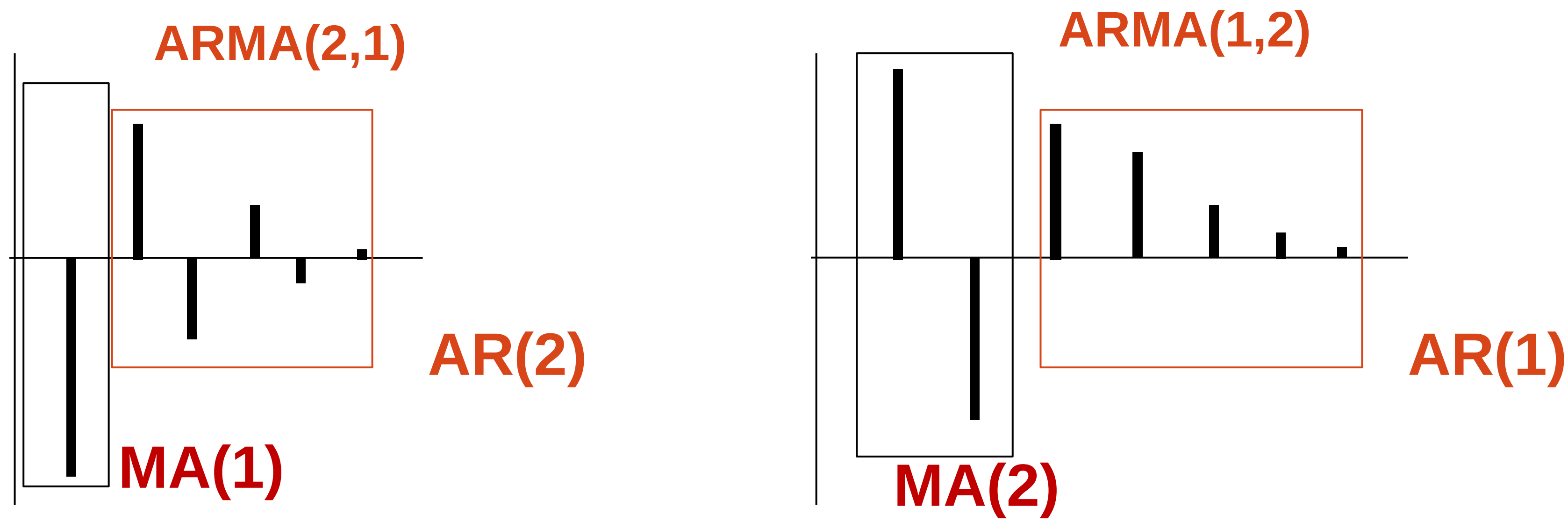
ARMA(1,1): $Y_t = 0.8Y_{t-1} + A_t - 0.3A_{t-1}$.

ARMA(1,1): $Y_t = 0.3Y_{t-1} + A_t - 0.8A_{t-1}$.

Patrones característicos de un modelo ARMA

- La **f.a.s. y la f.a.p. de los procesos ARMA** es el **resultado de la superposición** de sus propiedad **AR** y **MA**:

  - En la f.a.s ciertos coeficientes iniciales que dependen del orden de la parte MA y después un decrecimiento dictado por la parte AR.



  - En la f.a.p ciertos coeficientes iniciales que dependen del orden de la parte AR y después un decrecimiento dictado por la parte MA.

Identificación de los órdenes de un ARMA

- Esta estructura compleja hace que el **orden de un proceso ARMA sea difícil de identificar** en la práctica, al existir muchos procesos/modelos que generarían un mismo patrón es estas funciones.

- Por ello, la **estructura** del modelo se **va proponiendo paso a paso** de acuerdo a los gráficos f.a.s. y f.a.p. del residuo que queda tras cada paso.

  Sea el proceso ARMA(1,1):

$$(\mathbf{1} - \phi_1 \boldsymbol{B}) X_t = (\mathbf{1} - \boldsymbol{\theta}_1 \boldsymbol{B}) \boldsymbol{a}_t \quad \text{con} \quad \boldsymbol{a}_t \sim \boldsymbol{RB}$$

  Si en el proceso de ajuste, se comienza proponiendo un AR(1), entonces:

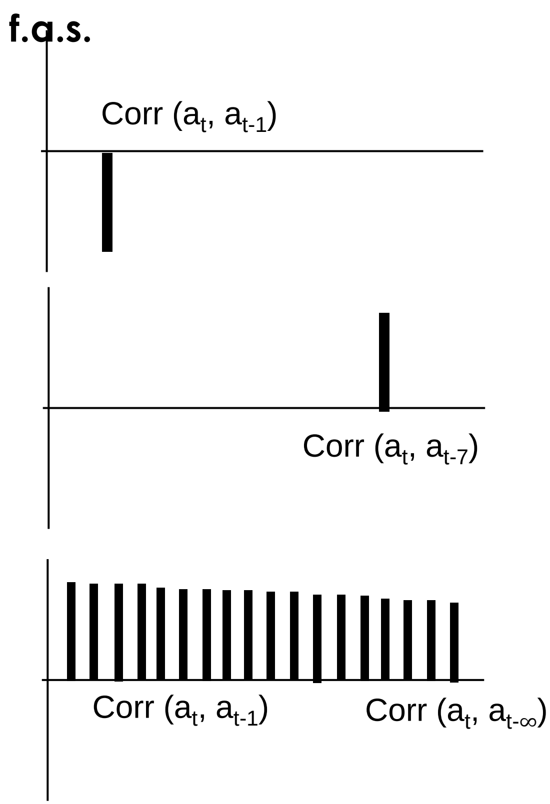
$$(\mathbf{1} - \boldsymbol{\delta} \boldsymbol{B}) X_t = \boldsymbol{e}_t \quad \text{donde } \boldsymbol{e}_t \text{ presentará estructura.}$$

  Suponiendo que ϕ_1 y $\boldsymbol{\delta}$ son similares: $\boldsymbol{e}_t = (\mathbf{1} - \boldsymbol{\theta}_1 \boldsymbol{B}) \boldsymbol{a}_t$

  $\boldsymbol{e}_t$ presentaría una estructura de tipo MA(1) que se identificaría en la f.a.s.

Identificación de los órdenes de un ARMA

- La naturaleza de este tipo de ajuste por pasos conduce a **proponer siempre una estructura de tipo MA** dado que el proceso consiste en observar la **f.a.s. asociada a un proceso residual.**



$$X_t = \boldsymbol{a_t - \theta_1 a_{t-1}}$$

$$X_t = \boldsymbol{a_t - \theta_7 a_{t-7}}$$

$$X_t = \boldsymbol{a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_\infty a_{t-\infty}}$$

¡ Principio de Parsimonia!

Identificación de los órdenes de un ARMA

- Se puede demostrar por sustitución recursiva que **un AR(p) equivale a un MA(∞).**

AR(1)

$$X_t = \phi_1 X_{t-1} + \boldsymbol{a}_t = \phi_1(\phi_1 X_{t-2} + \boldsymbol{a}_{t-1}) + \boldsymbol{a}_t = {\phi_1}^2 X_{t-2} + \phi_1 \boldsymbol{a}_{t-1} + \boldsymbol{a}_t =$$

$$= {\phi_1}^2(\phi_1 X_{t-3} + \boldsymbol{a}_{t-2}) + \phi_1 \boldsymbol{a}_{t-1} + \boldsymbol{a}_t = {\phi_1}^3 X_{t-3} + {\phi_1}^2 \boldsymbol{a}_{t-2} + \phi_1 \boldsymbol{a}_{t-1} + \boldsymbol{a}_t = \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\infty} {\phi_1}^{\boldsymbol{k}} \boldsymbol{a}_{t-\boldsymbol{k}}$$

AR(2)

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \boldsymbol{a}_t = \phi_1(\phi_1 X_{t-2} + \phi_2 X_{t-3} + \boldsymbol{a}_{t-1}) + \phi_2(\phi_1 X_{t-3} + \phi_2 X_{t-4} + \boldsymbol{a}_{t-2}) + \boldsymbol{a}_t =$$

$$= {\phi_1}^2 X_{t-2} + 2\phi_1\phi_2 X_{t-3} + \phi_1 \boldsymbol{a}_{t-1} + {\phi_2}^2 X_{t-4} + \phi_2 \boldsymbol{a}_{t-2} + \boldsymbol{a}_t =$$

$$= {\phi_1}^2(\phi_1 X_{t-3} + \phi_2 X_{t-4} + \boldsymbol{a}_{t-2}) + 2\phi_1\phi_2(\phi_1 X_{t-4} + \phi_2 X_{t-5} + \boldsymbol{a}_{t-3}) + {\phi_2}^2(\phi_1 X_{t-5} + \phi_2 X_{t-6} + \boldsymbol{a}_{t-4})$$

$$+ \phi_1 \boldsymbol{a}_{t-1} + \phi_2 \boldsymbol{a}_{t-2} + \boldsymbol{a}_t = \cdots = \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\infty} \boldsymbol{\delta}_{\boldsymbol{k}} \boldsymbol{a}_{t-\boldsymbol{k}}$$

- Cuando existen **muchas correlaciones significativas en la f.a.s.,** se propone una **estructura de tipo AR** y es la **f.a.p.** la que **ayuda a determinar el orden de la parte autorregresiva.**

Procesos/Modelos SARMA (ARMA estacionales)

- Una de las principales **causas** de la **no estacionariedad** de un proceso es la **presencia de un patrón estacional.**

- Una **serie** es **estacional** cuando su **valor esperado no** es **constante** (no es estacionario en media) pero varía con una pauta cíclica: $E[X_t] = E[X_{t-s}]$.

- **En el contexto de los modelos ARMA**, el concepto de **estacional** no se plantea en el sentido de algo periódico sino en el sentido de que **lo que ocurre en un instante "t" está correlacionado con lo que ocurre en el instante "t-s".**

$$(\mathbf{1} - \Phi_{\mathbf{1}}\boldsymbol{B}^{s} - \Phi_{\mathbf{2}}\boldsymbol{B}^{\mathbf{2}s} - \cdots - \Phi_{\boldsymbol{P}}\boldsymbol{B}^{\boldsymbol{P}s})X_t = (\mathbf{1} - \Theta_{\mathbf{1}}\boldsymbol{B}^{s} - \Theta_{\mathbf{2}}\boldsymbol{B}^{\mathbf{2}s} - \cdots - \Theta_{\boldsymbol{Q}}\boldsymbol{B}^{\boldsymbol{Q}s})\, \boldsymbol{a}_t$$

- De acuerdo a este concepto y aunque parezca contradictorio, existen **modelos ARMA estacionales estacionarios:** $\Phi(\boldsymbol{B}, \boldsymbol{B}^{s})X_t = \Theta(\boldsymbol{B}, \boldsymbol{B}^{s})\boldsymbol{a}_t$

  donde Φ verifica la cond. de estacionariedad

- El modelo es estacional no en el sentido de algo periódico sino en el sentido de que el operador retardo $\boldsymbol{B}^{s}$ aparece en alguna parte de la fórmula.

Procesos/Modelos SARMA (ARMA estacionales)

- El caso más habitual es que la **estacionalidad** de incorpore dentro del modelo ARMA **de forma multiplicativa** de manera que los polinomios $\Phi(\boldsymbol{B},\boldsymbol{B^s})\ y\ \Theta(\boldsymbol{B},\boldsymbol{B^s})$ se pueden factorizar en:

  - Un polinomio que depende de B y sus potencias: **PARTE REGULAR.**
  - Un polinomio que depende de Bs y sus potencias: **PARTE ESTACIONAL.**

SARMA(p,q)X(P,Q)$_s$: $\phi(\boldsymbol{B})\Phi(\boldsymbol{B^s})\,X_t = \theta(\boldsymbol{B})\Theta(\boldsymbol{B^s})\boldsymbol{a_t}$

$$\left.\begin{aligned} \phi(\boldsymbol{B}) &= \boldsymbol{1} - \phi_{\boldsymbol{1}}\boldsymbol{B} - \phi_{\boldsymbol{2}}\boldsymbol{B^2} - \cdots - \phi_{\boldsymbol{p}}\boldsymbol{B^p} \\ \theta(\boldsymbol{B}) &= \boldsymbol{1} - \theta_{\boldsymbol{1}}\boldsymbol{B} - \theta_{\boldsymbol{2}}\boldsymbol{B^2} - \cdots - \theta_{\boldsymbol{q}}\boldsymbol{B^q} \end{aligned}\right\} \Rightarrow \textbf{PARTE REGULAR}$$

$$\left.\begin{aligned} \Phi(\boldsymbol{B}) &= \boldsymbol{1} - \Phi_{\boldsymbol{1}}\boldsymbol{B^s} - \Phi_{\boldsymbol{2}}\boldsymbol{B^{2s}} - \cdots - \Phi_{\boldsymbol{P}}\boldsymbol{B^{Ps}} \\ \Theta(\boldsymbol{B}) &= \boldsymbol{1} - \Theta_{\boldsymbol{1}}\boldsymbol{B^s} - \Theta_{\boldsymbol{2}}\boldsymbol{B^{2s}} - \cdots - \Theta_{\boldsymbol{Q}}\boldsymbol{B^{Qs}} \end{aligned}\right\} \Rightarrow \textbf{PARTE ESTACIONAL}$$

Procesos/Modelos Integrados SARMA (SARIMAs)

- La posibilidad de incorporar una diferencia para hacer la serie estacionaria aplica ahora tanto a la parte regular como a la parte estacional.

- La estructura más general corresponde al modelo **SARIMA(p,d,q)X(P,D,Q)$_s$**:

$$\phi(\boldsymbol{B})\Phi(\boldsymbol{B}^s)(\mathbf{1}-\boldsymbol{B})^d(\mathbf{1}-\boldsymbol{B}^s)^D X_t = \theta(\boldsymbol{B})\Theta(\boldsymbol{B}^s)\boldsymbol{a}_t$$

siendo:

$$\left.\begin{array}{l}\phi(\boldsymbol{B}) = \mathbf{1} - \phi_1\boldsymbol{B} - \phi_2\boldsymbol{B}^2 - \cdots - \phi_p\boldsymbol{B}^p \\ \theta(\boldsymbol{B}) = \mathbf{1} - \theta_1\boldsymbol{B} - \theta_2\boldsymbol{B}^2 - \cdots - \theta_q\boldsymbol{B}^q\end{array}\right\} \Rightarrow \text{PARTE REGULAR}$$

$$\left.\begin{array}{l}\Phi(\boldsymbol{B}) = \mathbf{1} - \Phi_1\boldsymbol{B}^s - \Phi_2\boldsymbol{B}^{2s} - \cdots - \Phi_P\boldsymbol{B}^{Ps} \\ \Theta(\boldsymbol{B}) = \mathbf{1} - \Theta_1\boldsymbol{B}^s - \Theta_2\boldsymbol{B}^{2s} - \cdots - \Theta_Q\boldsymbol{B}^{Qs}\end{array}\right\} \Rightarrow \text{PARTE ESTACIONAL}$$

Identificación de los órdenes de un SARMA

- Estructura de la F.A.S y F.A.P. de un modelo SARMA(p,q)x(P,Q)$_s$:

F.A.S. (ACF)

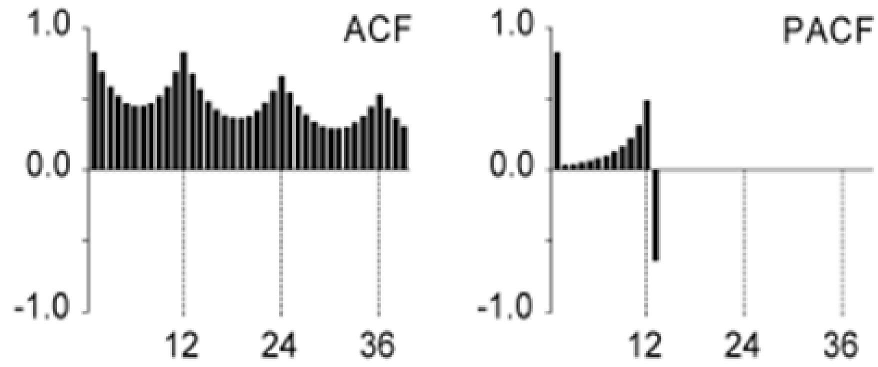
En los primeros retardos aparece la f.a.s. de la parte regular y en los retardos s,2s,… aparece la f.a.s. de la parte estacional acompañada a ambos lados de la f.a.s. regular.

F.A.P. (PACF)

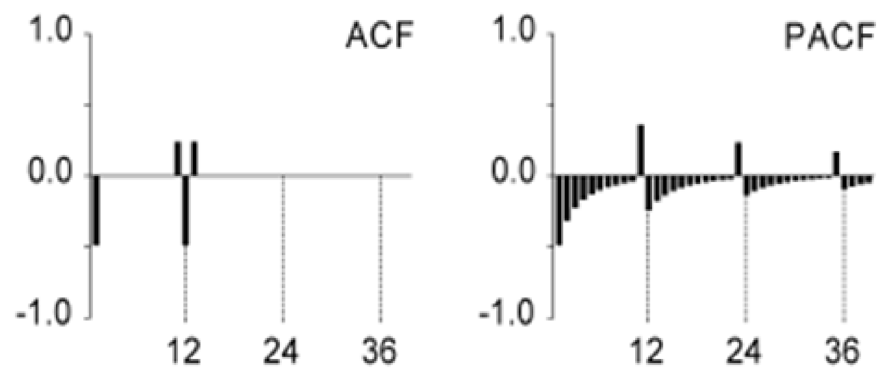
En los primeros retardos aparece la f.a.p. de la parte regular y en los retardos s,2s,… aparece la f.a.p. de la estacional.

A la derecha de cada coeficiente estacional aparece la f.a.p. de la parte regular. Si el coeficiente estacional es "+", la f.a.p. aparece invertida, mientras que si es "-", la f.a.p. aparece con su signo.

A la izquierda de los coeficientes estacionales, la f.a.s. de la parte regular.



$\mathrm{AR}(1) \times \mathrm{AR}(1)_{12}$, con $\phi_1 = 0.8$, $\Phi_1 = 0.8$.



$\mathrm{MA}(1) \times \mathrm{MA}(1)_{12}$, con $\theta_1 = 0.8$, $\Theta_1 = 0.8$.

¿Modelo SARMA multiplicativo o aditivo?

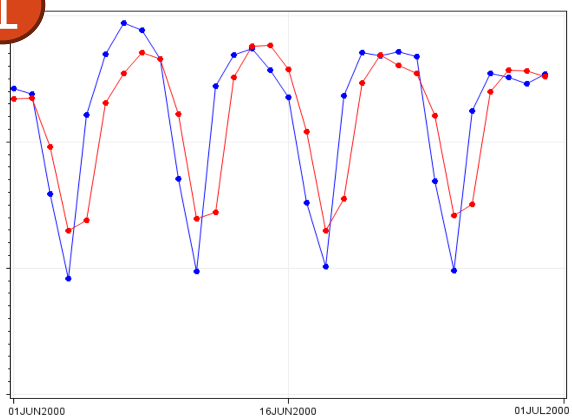
1 Los modelos multiplicativos suelen funcionar mejor:

AR(s) con $\phi_2 = \cdots = \phi_{s-1} = 0 \;\; (1 - \phi_1 B - \phi_s B^s) X_t = a_t \Leftrightarrow X_t = \phi_1 X_{t-1} + \phi_s X_{t-s} + a_t$

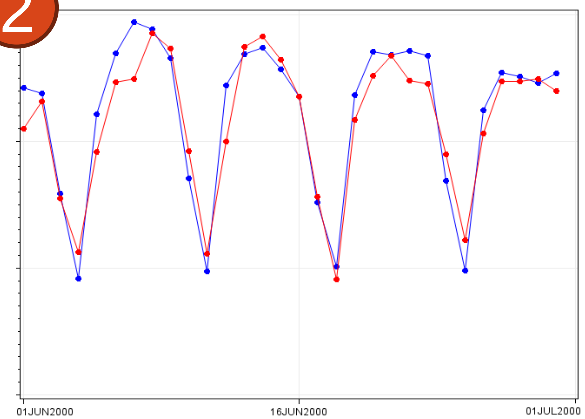
2 SAR(1)x(1) $(1 - \phi_1 B)(1 - \phi_s B^s) X_t = a_t \Leftrightarrow X_t = \phi_1 X_{t-1} + \phi_s X_{t-s} - \phi_1 \phi_s X_{t-s-1} + a_t \Leftrightarrow$

$$\Leftrightarrow X_t = \phi_1 X_{t-1} + \phi_s (X_{t-s} - \phi_1 X_{t-s-1}) + a_t$$

1



2



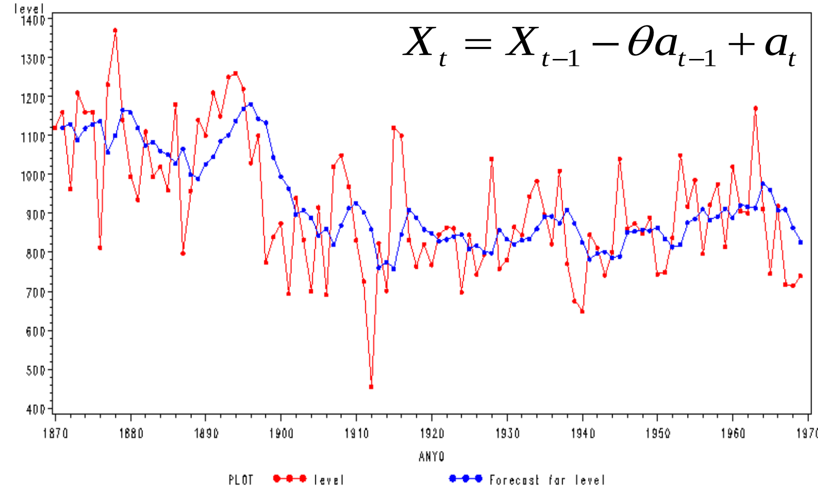
$\phi_1 = 0{,}60377; \phi_7 = 0{,}38170;$

$\phi_1 = 0{,}87375; \phi_7 = 0{,}81103; \phi_8 = -0{,}71041;$

Algunos modelos “de andar por casa”

- **Suavizado exponencial simple:** ARIMA(0,1,1)

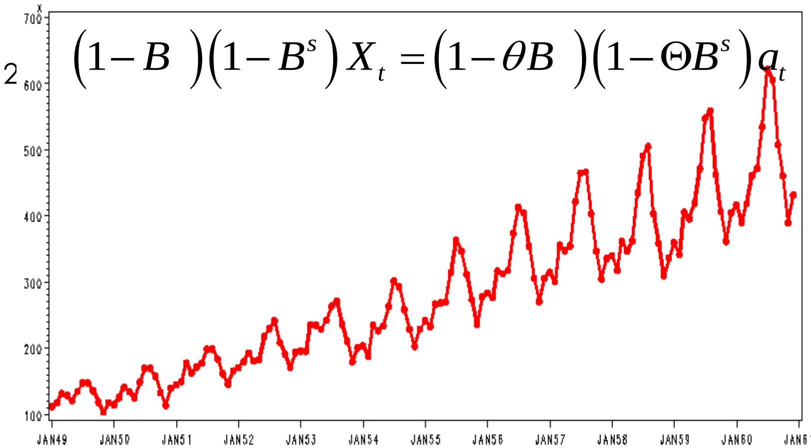
Para predecir, toma como base el último dato real disponible y se fija en el último error que cometió el modelo
El ARMA(1,1) se puede ver como una generalización suya



$$X_t = X_{t-1} - \theta a_{t-1} + a_t$$

- **Modelo de Líneas Aéreas:** SARIMA(0,1,1) x $(0,1,1)_{12}$

Para predecir, toma como base el último dato real disponible (regular y estacional) y se fija en el último error que cometió el modelo (regular y estacional)
Los SARIMA(1,0,1) x (1,0,1) son generalizaciones de este modelo



$$(1-B)(1-B^s)X_t = (1-\theta B)(1-\Theta B^s)a_t$$

3 Metodología *Box-Jenkins*

Pasos principales al ajustar series temporales

- Los pasos principales en el proceso de ajuste de una serie temporal son:

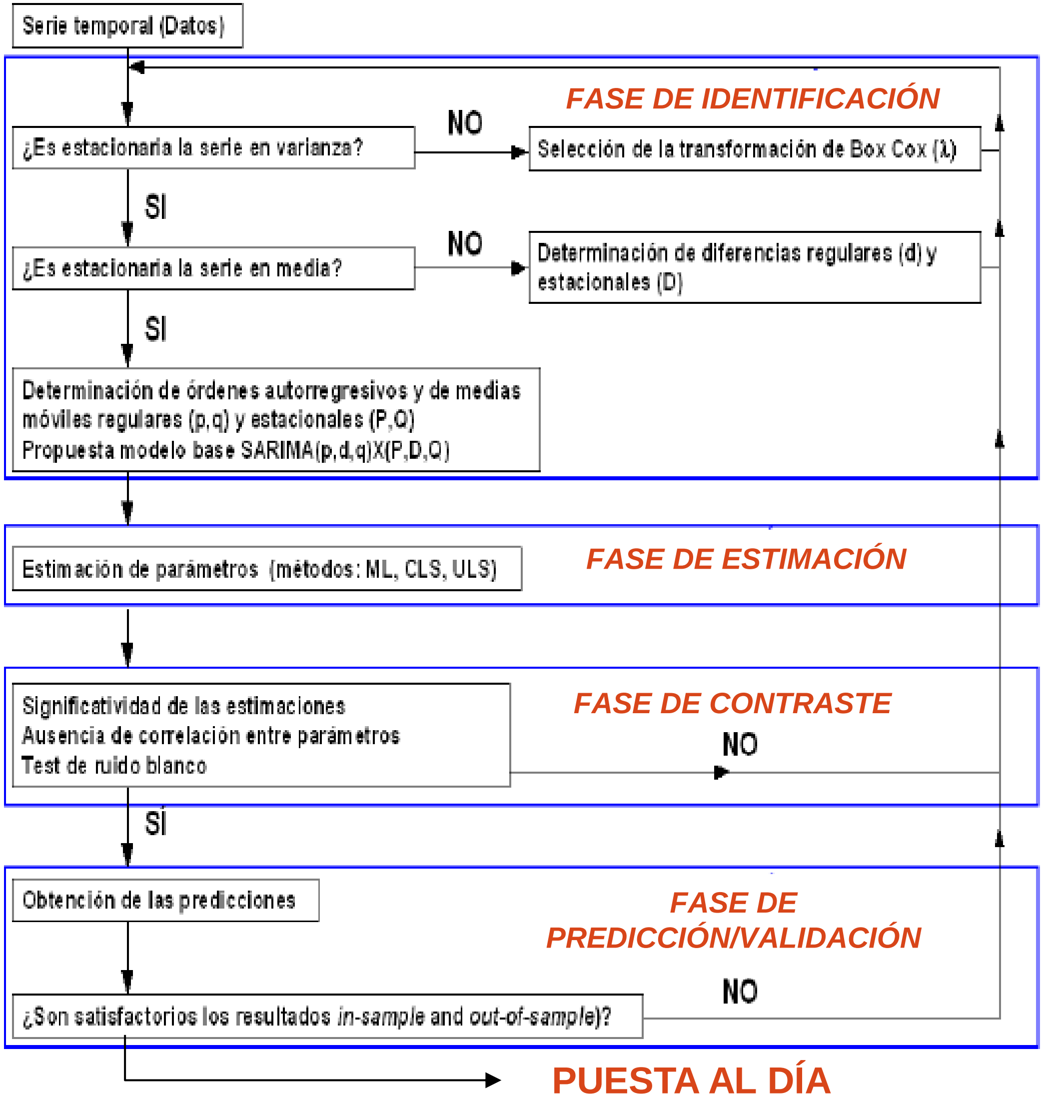
  1. Hacer estacionaria la serie.

  2. Identificar los órdenes del proceso ARMA subyacente bajo la serie estacionaria.

  3. Contrastar que el proceso residual resultante tras el ajuste es Ruido Blanco.

- Estos pasos constituyen la esencia de una metodología que goza de bastante popularidad en el ajuste de series temporales: Metodología *Box – Jenkins*.

Metodología *Box-Jenkins* (1976)



En esta fase se realizan las transformaciones necesarias para hacer que la serie sea estacionaria, se identifican los órdenes p,q,P y Q del SARIMA estacionario y se incluyen posibles variables input.

En esta fase se cuantifican los parámetros reflejados en la fórmula.

En esta fase se contrasta la validez del modelo desde el punto de vista estadístico: parámetros significativos, ausencia de correlaciones, RB, etc.

En esta última fase se contrasta la bondad del modelo desde el punto de vista de su calidad predictiva.

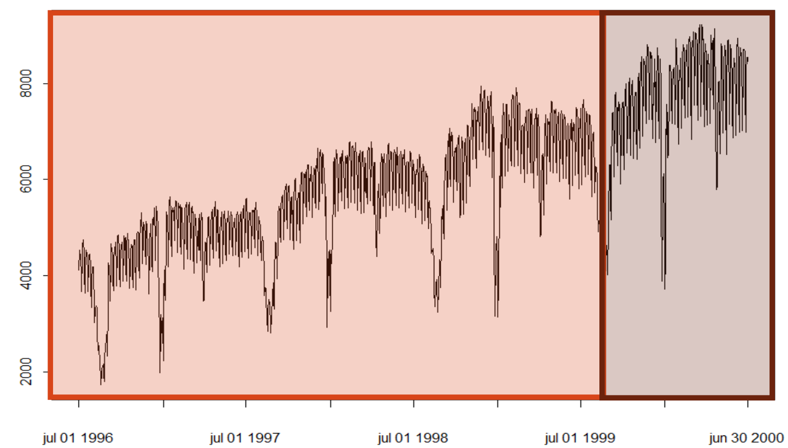
Selección de muestras de entrenamiento y validación

- Retomemos el ejercicio de predicción de demanda industrial de gas.

- Antes de empezar a hacer el estudio hay que distinguir la parte de los datos que se utilizará para construir la fórmula (para modelizar) de aquélla que se utilizará para validar los resultados y cuyos datos, no deben ser utilizados.

- En esta separación se ha tenido en cuenta la estacionalidad anual de la serie.

```
> datos.train <- subset(datos, "1996-07-01"<=FECHA & FECHA<="1999-06-30")
> datos.train.ts <- as.ts(datos.train$DEMANDA_GAS, frequency = 7)
>
> datos.validate <- subset(datos, FECHA>"1999-06-30")
> datos.validate.ts <- as.ts(datos.validate$DEMANDA_GAS, frequency=7)
```



Fase de identificación: Homocedasticidad

- A través de la función *boxcox* se puede valorar la conveniencia de realizar alguna transformación previa de los datos para conseguir que sea estacionaria en varianza.

```
> # Se evalúa la necesidad de transformar la serie para hacerla
> # estacionaria en varianza
> box_cox <- boxcox(DEMANDA_GAS ~ FECHA,
+                   data = datos.train,
+                   lambda = c(0, 0.5, 1))
> lambda <- box_cox$x[which.max(box_cox$y)]
> lambda
[1] 1
> # -> No es necesario transformar los datos
```

Nº de valores λ a contrastar

Valor de λ que maximiza la función de verosimilitud

$$Y^{(\lambda)}{}_t = \begin{cases} \dfrac{(X_t - min\{X_t\} + 1)^\lambda}{\lambda} & si\ \lambda \neq 0 \\ ln(X_t - min\{X_t\} + 1) & si\ \lambda = 0 \end{cases}$$

Metodología *Box-Jenkins*
Fases de identificación y estimación.

- Las funciones *acf* y *pacf* permiten visualizar las gráficas de la f.a.p. y la f.a.s. respectivamente. Esto permite saber los órdenes p, P, q y Q del modelo SARIMA a ajustar:

```
> acf(datos.train.ts, lag.max = 25, xlab = "Retardo",
+     main= "Función de autocorrelación simple")
>
> pacf(datos.train.ts, lag.max = 25, xlab = "Retardo",
+      main = "Función de autocorrelación parcial")
```

- Por su parte, la función *Arima* del paquete *forecast* permite el ajuste de modelos

```
> modeloArima <- Arima(datos.train.ts,
+                      order = c(p, d, q),
+                      seasonal = list (order = c(P, D, Q), period = 7),
+                      method = "ML")
```

Ajuste de parámetros de la parte regular

Especificación del periodo de estacionalidad de la serie

Método de estimación de los parámetros

Ajuste de parámetros de la parte estacional

NOTA: Por defecto, el término constante se suprime automáticamente cuando deja de ser significativo.

Metodología *Box-Jenkins*
Fase de estimación

- Una vez identificado el modelo, se procederá a la estimación de sus parámetros a través de alguno de los siguientes métodos:
  - **Método de mínimos cuadrados.-** este método realiza la estimación de los parámetros marcando el objetivo de minimizar el error cuadrático medio (diferencia al cuadrado entre dato real y dato predicho).
  - **Método de máxima verosimilitud (ML).-** los parámetros estimados por este método son aquéllos que maximizan la probabilidad de que la serie venga representada por dichos parámetros.

- Estos métodos presentan dos variantes:
  - Un **enfoque condicional** en el que se condiciona al conocimiento de los valores iniciales de la serie y el residuo.
  - Un **enfoque no condicional** o exacto en el que no se asume dicho conocimiento.

- Los **algoritmos de estimación** están **sujetos a restricciones sobre los parámetros**: las **raíces** de los polinomios asociados a las partes AR y MA **no pueden estar fuera del círculo unidad** (como máximo, sobre él).

Metodología *Box-Jenkins*

Fase de contraste

- En el proceso iterativo de identificación y estimación, se realiza constantemente **un análisis de las estimaciones**:

  - **Estimadores significativos** (p-valor/t-ratio)
    $$\begin{cases} |t-ratio(\phi)| = \left| \dfrac{\hat{\phi}}{\sqrt{Var(\hat{\phi})}} \right| >= 1.96 \\ \text{p-valor} < 0.05 \end{cases}$$

```
install.packages("lmtest")
library(lmtest)

coeftest(ajuste1)
```

  - Los parámetros cumplen la condición de **estacionariedad**.

  - **Ausencia de correlación entre parámetros** (<0,8 en valor absoluto) para **evitar** posibles efectos de **multicolinealidad**.

```
install.packages("caschrono")
library(caschrono)

cor.arma(ajuste1)
```

Metodología *Box-Jenkins*

Fase de contraste

- En el proceso iterativo de identificación y estimación, se observa de forma constante el resultado del **test de ruido blanco**. El estadístico de *Ljung-Box*, permite determinar si grupos de "h" correlaciones residuales son o no 0.

$$Q(h) = n(n+2)\sum_{j=1}^{h} \frac{r_{a_j}^2}{n-j} \approx \chi_{h-l}^2, l = n^o \; parámetros estimados$$

```
Box.test.2(residuals(ajuste1),
           nlag = c(6,12,18,24,30,36,42,48),
           type="Ljung-Box")
```

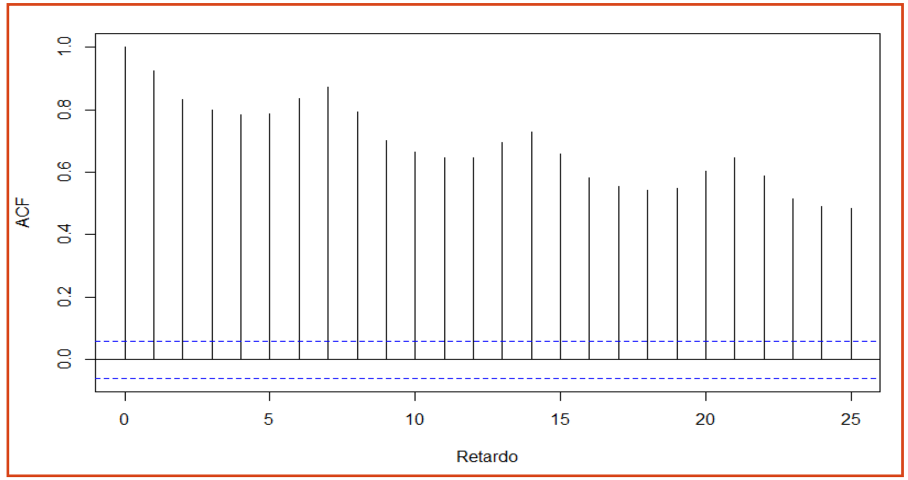
- Por otra parte, al final del ajuste, se debe contrastar:

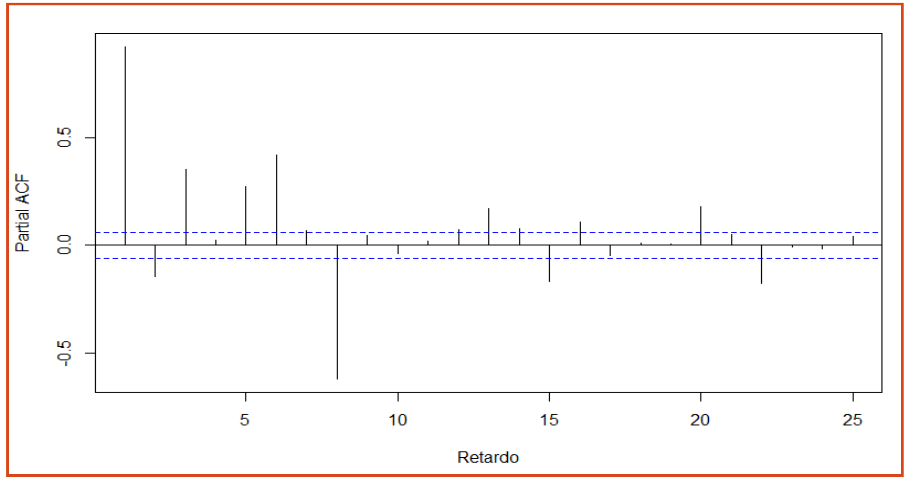
  - Que la **media** del **residuo es 0**. Esta condición está garantizada si el método de estimación es de tipo mínimo - cuadrático.
  - Que la **varianza** del **residuo** es **constante**. Aun cuando esta condición está prácticamente garantizada (la posible transformación de Box-Cox garantiza la estacionariedad en varianza de la serie y por tanto del residuo), se puede recurrir a un gráfico rango/std sobre el residuo.
  - Que los **residuos** son **normales**. Esta condición no es necesaria pero sí deseable (pues garantiza independencia del residuo).

Metodología *Box-Jenkins*

Proceso iterativo de identificación, estimación y contraste

- El gráfico f.a.s de la serie original sugiere el ajuste de un modelo AR.
- Lo habitual es empezar con una estructura AR(1). La fuerte correlación de orden 1 en la f.a.p parece reforzar esta decisión.
- También se observan ciertos "repuntes" en las correlaciones múltiplo de 7 de la f.a.s, que llevan a intuir una estructura estacional.

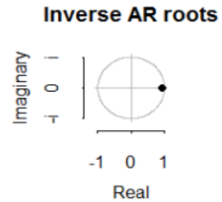




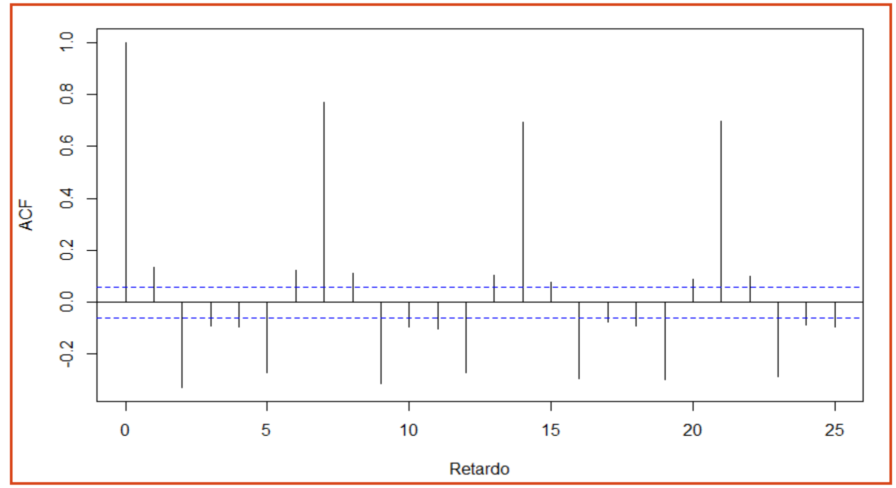
Metodología *Box-Jenkins*

Proceso iterativo de identificación, estimación y contraste

- El ajuste del modelo AR(1) + μ proporciona los siguientes resultados:



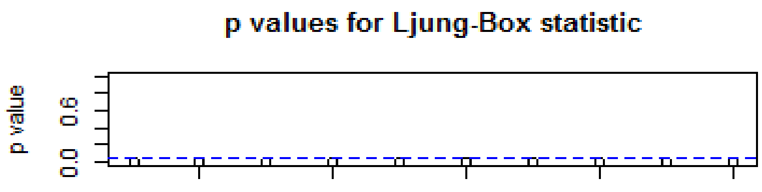

```
Coefficients:
         ar1  intercept
      0.9256  5532.6121
s.e.  0.0114   187.7531

sigma^2 estimated as 218754:  log likelihood = -8286.61,
aic = 16579.21
```



Matriz de correlación de parámetros estimados

```
                   ar1     intercept
ar1        1.0000000000 -0.0007726061
intercept -0.0007726061  1.0000000000
```

Test de ruido blanco sobre los residuos (Ljung-Box)



Retard	p-value
6	0
12	0
18	0
24	0
30	0
36	0

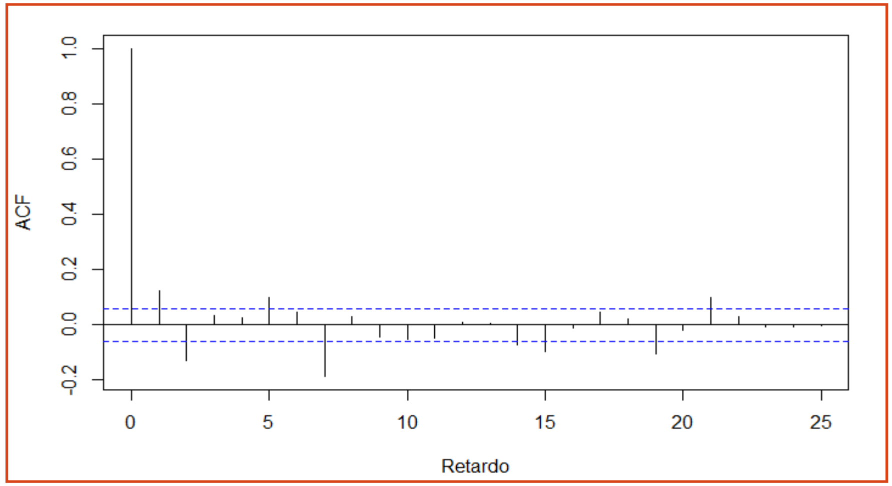
- El parámetro asociado al AR(1) se mueve cerca de 1: podría sugerirse una diferencia I(1).
- El nuevo gráfico f.a.s corresponde a un AR(7). Se aprecia estructura multiplicativa.

Metodología *Box-Jenkins*

Proceso iterativo de identificación, estimación y contraste

- El ajuste del modelo $SAR(1)x(1)_7 + \mu$ proporciona los siguientes resultados:

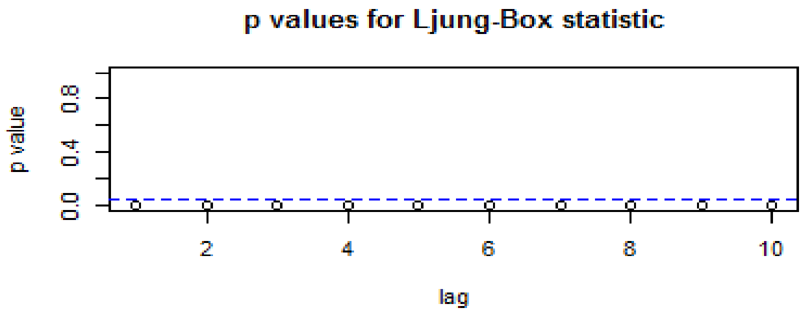
```
Coefficients:
         ar1     sar1       mean
      0.8727   0.7869  5532.3225
s.e.  0.0147   0.0187   315.6007

sigma^2 estimated as 85291:  log likelihood=-7772.91
AIC=15553.83   AICc=15553.86   BIC=15573.82
```



Matriz de correlación de parámetros estimados

```
                   ar1        sar1   intercept
ar1        1.000000000 -0.13981597 0.003511229
sar1      -0.139815967  1.00000000 0.001928590
intercept  0.003511229  0.00192859 1.000000000
```

Test de ruido blanco sobre los residuos (Ljung-Box)




```
     Retard p-value
[1,]      6   1e-08
[2,]     12   0e+00
[3,]     18   0e+00
[4,]     24   0e+00
[5,]     30   0e+00
```

- Se observa una correlación fuerte de orden 7 que justifica la propuesta de un MA(7).

Metodología *Box-Jenkins*

Proceso iterativo de identificación, estimación y contraste

- El modelo SARIMA(1,0,0)x$(1,0,1)_7$ + μ es inestable, violándose la condición de estacionariedad.

```
Coefficients:
         ar1     sar1     sma1      mean
      0.9490   0.9994  -0.9402  5531.626
s.e.  0.0098   0.0004   0.0128  3066.838

sigma^2 estimated as 63809:  log likelihood=-7621.53
AIC=15253.06   AICc=15253.11   BIC=15278.05
```

- El polinomio AR(7) tiene una raíz unitaria que justifica reemplazar el AR(7) por una diferencia estacional.

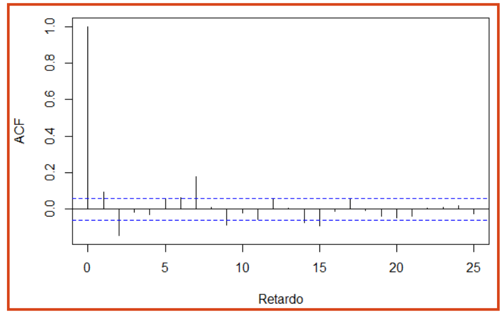
- Al hacer este cambio, la pendiente deja de ser significativa, por lo que deja de estar incluida en el modelo.

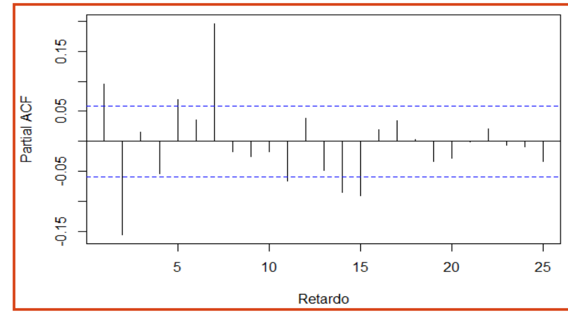
Metodología *Box-Jenkins*

Proceso iterativo de identificación, estimación y contraste

- Se ajusta un SARIMA(1,0,0)x(0,1,1)$_7$ que proporciona los siguientes resultados:

```
Coefficients:
         ar1      sma1
      0.9509   -0.9411
s.e.  0.0099    0.0125
```

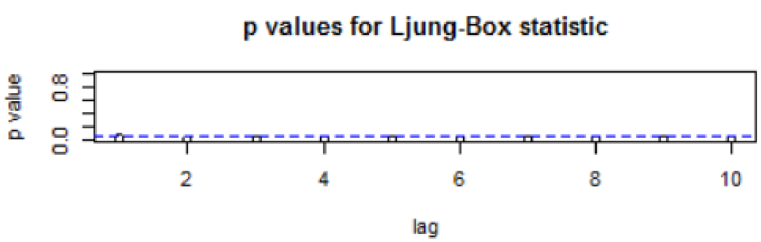




<u>Matriz de correlación de parámetros estimados</u>

```
            ar1       sma1
ar1   1.0000000 -0.3043618
sma1 -0.3043618  1.0000000
```

<u>Test de ruido blanco sobre los residuos (Ljung-Box)</u>




```
     Retard p-value
[1,]      6 1.7e-07
[2,]     12 0.0e+00
[3,]     18 0.0e+00
[4,]     24 0.0e+00
[5,]     30 0.0e+00
[6,]     36 0.0e+00
[7,]     42 0.0e+00
[8,]     48 0.0e+00
```

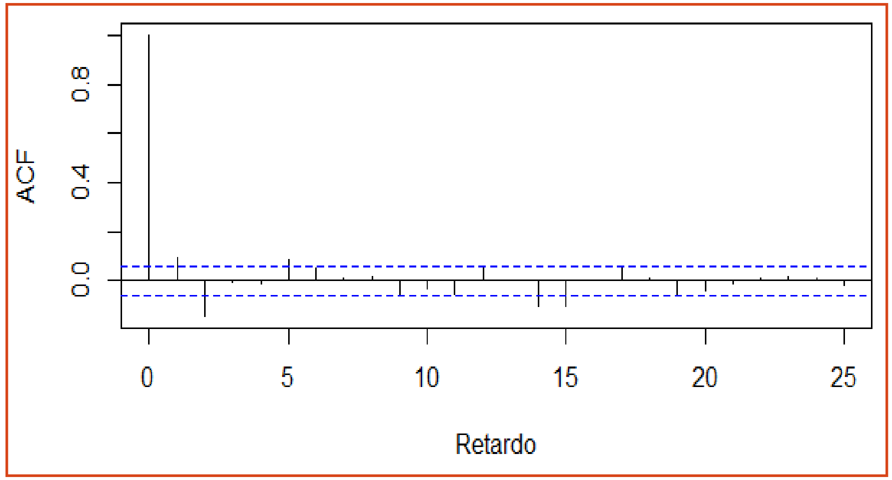
- Hay una correlación fuerte de orden 7 en la f.a.s.
- Ya existe una estructura MA(7) en el modelo, pero la f.a.p permite justificar un AR(7).

Metodología *Box-Jenkins*

Proceso iterativo de identificación, estimación y contraste

- Se ajusta un SARIMA(1,0,0)x$(1,1,1)_7$ que proporciona los siguientes resultados:

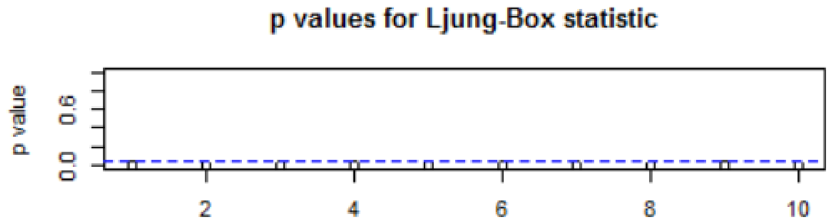
```
Coefficients:
         ar1     sar1     sma1
      0.9417   0.2065  -0.9574
s.e.  0.0112   0.0316   0.0092

sigma^2 estimated as 61336:  log likelihood=-7546.97
AIC=15101.95   AICc=15101.98   BIC=15121.92
```



Matriz de correlación de parámetros estimados

```
            ar1        sar1        sma1
ar1   1.0000000 -0.1698264 -0.3101595
sar1 -0.1698264  1.0000000 -0.2191124
sma1 -0.3101595 -0.2191124  1.0000000
```

Test de ruido blanco sobre los residuos (Ljung-Box)




```
     Retard  p-value
[1,]      6 1.40e-07
[2,]     12 3.30e-07
[3,]     18 0.00e+00
[4,]     24 1.00e-08
[5,]     30 3.00e-08
[6,]     36 1.40e-07
[7,]     42 5.20e-07
[8,]     48 1.73e-06
```

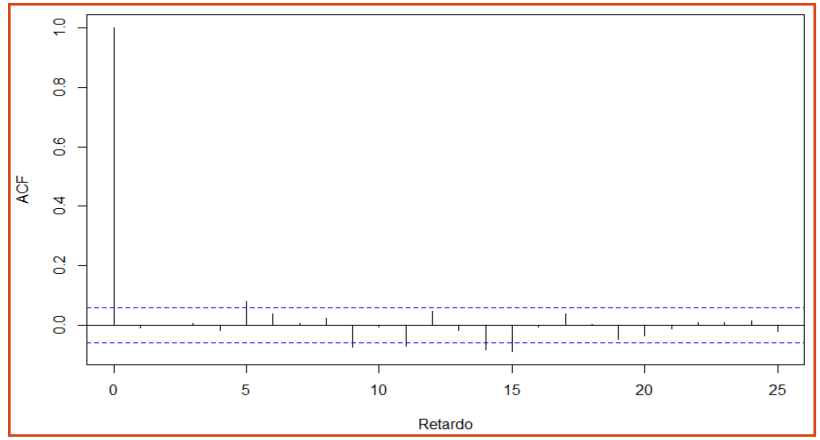
- Las dos primeras autocorrelaciones de la f.a.s son significativas. Se propone un MA(2).

Metodología *Box-Jenkins*

Proceso iterativo de identificación, estimación y contraste

- Se ajusta un SARIMA(1,0,2)x(1,1,1)$_7$ que proporciona los siguientes resultados:

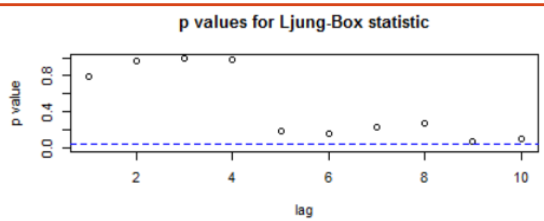
```
Coefficients:
         ar1     ma1      ma2    sar1     sma1
      0.9477  0.1199  -0.1573  0.2021  -0.9586
s.e.  0.0121  0.0322   0.0323  0.0316   0.0093

sigma^2 estimated as 59280:  log likelihood=-7527.53
AIC=15067.05   AICc=15067.13   BIC=15097.01
```



Matriz de correlación de parámetros estimados

```
            ar1         ma1         ma2        sar1        sma1
ar1   1.0000000 -0.36968806 -0.36665150 -0.16772306 -0.33435838
ma1  -0.3696881  1.00000000  0.20030598  0.02336780  0.08807173
ma2  -0.3666515  0.20030598  1.00000000  0.06242113  0.08670483
sar1 -0.1677231  0.02336780  0.06242113  1.00000000 -0.20897105
sma1 -0.3343584  0.08807173  0.08670483 -0.20897105  1.00000000
```

Test de ruido blanco sobre los residuos (Ljung-Box)




```
     Retard    p-value
[1,]      6 0.16668483
[2,]     12 0.02348665
[3,]     18 0.00141793
[4,]     24 0.00496902
```

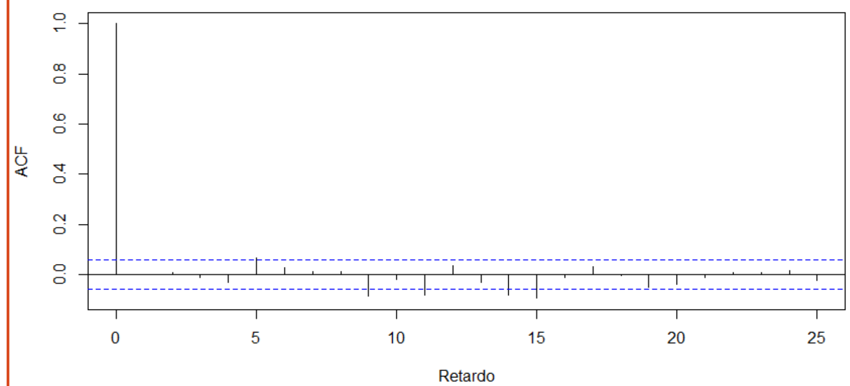
- Se observa como empiezan a aumentar los p-valores asociados al test de ruido blanco.
- No es posible identificar más estructura a partir de la f.a.s / f.a.p.

Metodología *Box-Jenkins*

Proceso iterativo de identificación, estimación y contraste

- Alternativamente, se podría haber ajustado un SARIMA(0,1,2)x(1,1,1)$_7$ (diferencia regular en lugar de AR(1)):

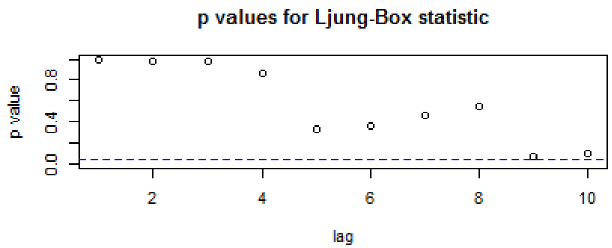
```
Coefficients:
         ma1      ma2    sar1     sma1
      0.0905  -0.1875  0.1904  -0.9683
s.e.  0.0299   0.0306  0.0310   0.0084

sigma^2 estimated as 60595:  log likelihood=-7533.84
AIC=15077.68   AICc=15077.73   BIC=15102.64
```



Matriz de correlación de parámetros estimados

	ma1	ma2	sar1	sma1
ma1	1.00000000	0.08822634	-0.01275039	0.01666388
ma2	0.08822634	1.00000000	0.03738667	0.03127935
sar1	-0.01275039	0.03738667	1.00000000	-0.25929401
sma1	0.01666388	0.03127935	-0.25929401	1.00000000

Test de ruido blanco sobre los residuos (Ljung-Box)



	Retard	p-value
[1,]	6	0.36535389
[2,]	12	0.01633023
[3,]	18	0.00052101
[4,]	24	0.00177601
[5,]	30	0.00099253
[6,]	36	0.00195558
[7,]	42	0.00467881
[8,]	48	0.00954918

- Empiezan a aumentar los p-valores asociados al test de RB y ausencia de estructura en la f.a.s / f.a.p.

Metodología Box-Jenkins
Fase de predicción/validación

- En esta fase se trata de utilizar el modelo para hacer predicciones. Es importante validar la bondad de las mismas.

- Para ello se suelen utilizar medidas tales como:

  - **MAPE** (Error Porcentual Absoluto Medio).- calculado como la media de: $APE_t = \frac{|Real - Prediccion|}{Real}$

  - **MedAPE** (Error Porcentual Absoluto Mediano): es una medida más robusta que el MAPE y resulta especialmente útil si en el periodo de predicción contempla datos atípicos difíciles de justificar.

  - **SMAPE** (Error Porcentual Absoluto Medio Simétrico) .- media de: $\frac{|Real - Prediccion|}{\frac{|Real + Prediccion|}{2}}$

- En cualquier caso, se recomienda:

  - **Utilizar el conjunto de validación/test:** Debe utilizarse un conjunto de datos distinto al que ha servido para construir el modelo. Un buen modelo, tiene errores de magnitud parecida sobre los conjuntos de entrenamiento y validación/test.
  - **Validar que la magnitud del error crece conforme lo hace el horizonte de predicción:** Conforme crece el horizonte, también lo hace la incertidumbre y por tanto debe disminuir la calidad de la predicción.

Metodología Box-Jenkins
Fase de predicción

- En esta fase se trata de utilizar el modelo para hacer predicciones. Es importante validar la bondad de las mismas.

  - **Fase de Predicción:** La función *predict* admite los siguientes parámetros:

```
> prediccion <- predict(modelo, n.ahead = , newxreg = )
```

Modelo ARIMA ajustado

Número de unidades temporales a las que se va a realizar la predicción

Parámetro opcional. Matriz con las variables explicativas informadas a futuro

- En esta fase se trata de utilizar el modelo para hacer predicciones. Es importante validar la bondad de las mismas.

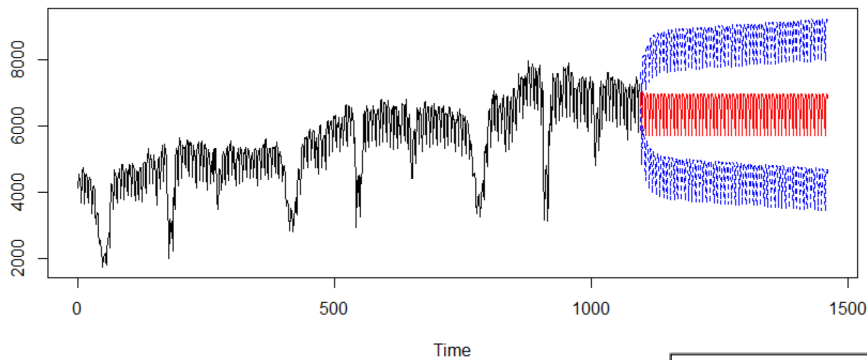
Metodología Box-Jenkins
Fase de predicción

- En caso de que la variable a predecir hubiese sido previamente transformada, es necesario destransformar la predicción.

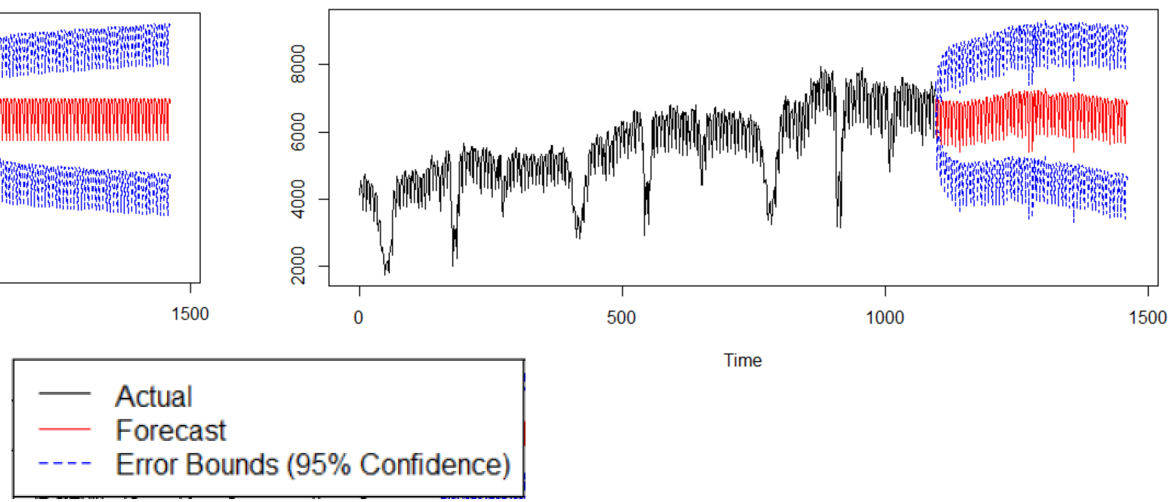
```
> predRaizDestransformada <- predConRaiz**2
> predLogDestransformada <- exp(predConLog+0.5*std**2)
```

- Es importante observar que los modelos SARIMA son apropiados para una predicción a corto plazo, no siendo fiables sus predicciones a largo plazo.

Sin festivos



Con festivos



4 Análisis de intervenciones y detección de *outliers*

Análisis de intervenciones y *outliers*

- En ocasiones el patrón esperado de la serie, se rompe por la presencia de efectos:
  - **Puntuales** (Atípicos de tipo "pulso").
  - **Permanentes** en el tiempo (Atípicos de tipo "escalón").
  - **Transitorios** (Atípicos de tipo "transitorio").
- La causa por la que un dato puede ser atípico puede:
  - Conocerse de antemano, siendo por tanto predecible.- festivos, puentes, periodos vacacionales, huelgas, etc.
  - Conocerse, pero no de antemano.- avería del sistema de medición de datos, atentado terrorista, etc.
  - No conocerse.- datos *missings*, datos injustificadamente altos o bajos.
- Las dos primeras se clasificarían dentro del denominado **análisis de intervenciones** y la tercera dentro del denominado proceso de **detección de *outliers*.**
- Esta información puede ser incorporada al modelo a través de variables explicativas binarias identificadas a partir de la visualización del histórico de la propia serie.

Análisis de intervenciones y *outliers*

A partir de una variable binaria se pueden reproducir diferentes efectos a través de **modelos de función de transferencia** que reflejan como se transfiere el efecto de la variable exógena a la serie.

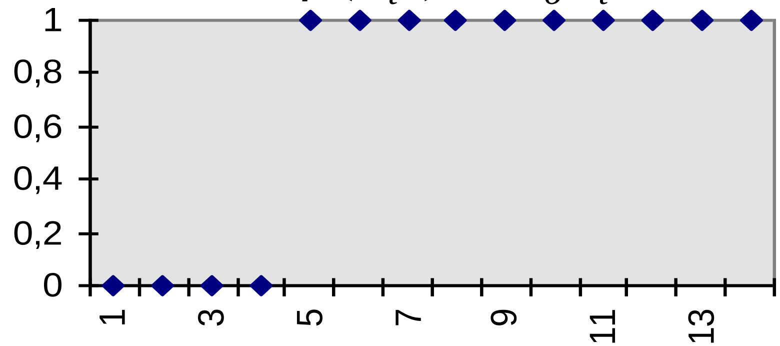
$$S_{t*}(t) = \begin{cases} 1 & si\ t \geq t^* \\ 0 & si\ t < t^* \end{cases}$$

	PERMANENTE	TRANSITORIO

INSTANTÁNEO

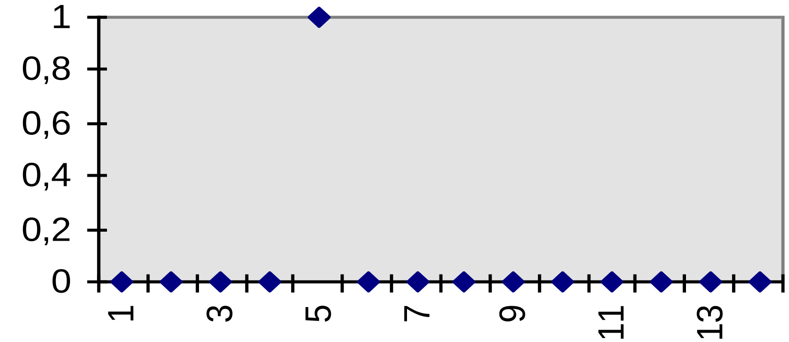
PERMANENTE

$$f(S_{t*}) = w_o S_{t*}$$



TRANSITORIO

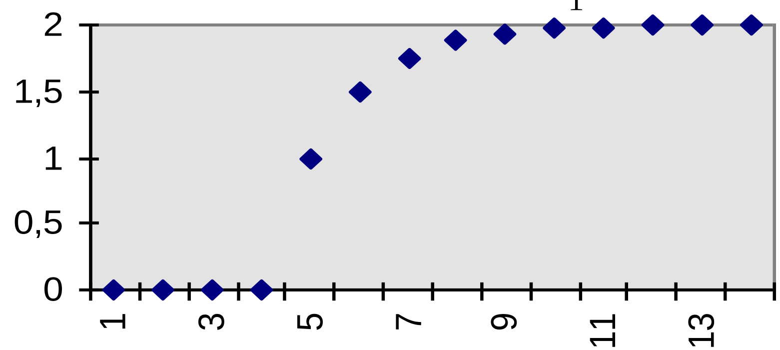
$$f(S_{t*}) = w_o (1-B) S_{t*}$$



GRADUAL

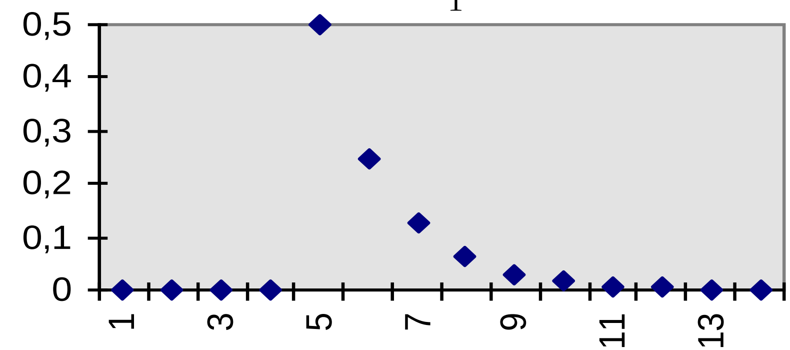
PERMANENTE

$$f(S_{t*}) = \frac{w_o}{1-\delta_1 B} S_{t*}$$



TRANSITORIO

$$f(S_{t*}) = \frac{w_o}{1-\delta_1 B} (1-B) S_{t*}$$



$\omega_0=1, \delta_0=0,5$

Generación de variables de intervención

- Se genera una variable binaria asociada a cada una de las festividades. Previamente es necesario obtener variables asociadas a la fecha (día del mes, mes, etc) para definir los festivos. Ejemplo:

```r
library(lubridate)

calendario$diaSemana <- wday(calendario$FECHA)
calendario$diaMes <- day(calendario$FECHA)
calendario$mes <- month(calendario$FECHA)

calendario$p_01ene <- ifelse(calendario$diaMes==1 & calendario$mes==1, 1, 0)
```

- La palabra clave *xreg* permite la inclusión de estas variables en el ARIMA a través un objeto de tipo matriz o vector:

```r
modeloArima <- Arima(datos.train.ts,
                     order = c(1,0,2),
                     seasonal = list(order = c(1,1,1), period=7),
                     method="ML",
                     xreg = calendario.train)
```

Estimación de parámetros asociados a intervenciones

- Las estimaciones de los parámetros asociados a las variables de intervención son negativos, lo cual concuerda, con el efecto que se espera tengan éstas.

```
Series: datos.train.ts
Regression with ARIMA(1,0,2)(1,1,1)[7] errors

Coefficients:
         ar1     ma1      ma2    sar1     sma1
      0.9504  0.1888  -0.0631  0.1806  -0.9511
s.e.  0.0111  0.0345   0.0322  0.0325   0.0102
       p_01ene    p_06ene    p_19mar    p_01may    p_15ago    p_12oct    p_01nov    p_08dic   p_06dic    p_25dic
     -964.1202  -923.9061  -465.3183  -368.5304  -217.5530  -324.2367  -162.3516  -348.1607 -228.4235  -941.2842
s.e.   75.3830    72.1291    72.0330    71.4878    71.3898    71.9102    71.2979    72.5343   72.4790    80.6290

sigma^2 estimated as 39447:  log likelihood=-7300.61
AIC=14633.21   AICc=14633.72   BIC=14713.09
```

Fase de contraste: análisis del residuo

- Se presenta el test de ruido blanco una vez introducidas todas estas variables para los dos modelos candidatos:

SARIMA(1,0,2)x(1,1,1)$_7$

	Retard	p-value
[1,]	6	0.16668483
[2,]	12	0.02348665
[3,]	18	0.00141793
[4,]	24	0.00496902
[5,]	30	0.00238685
[6,]	36	0.00493395
[7,]	42	0.00991395
[8,]	48	0.02016342

→

	Retard	p-value
[1,]	6	0.26783113
[2,]	12	0.06479063
[3,]	18	0.04728734
[4,]	24	0.10887151
[5,]	30	0.20118318
[6,]	36	0.17537176
[7,]	42	0.11347993
[8,]	48	0.13102472

SARIMA(0,1,2)x(1,1,1)$_7$

	Retard	p-value
[1,]	6	0.36535389
[2,]	12	0.01633023
[3,]	18	0.00052101
[4,]	24	0.00177601
[5,]	30	0.00099253
[6,]	36	0.00195558
[7,]	42	0.00467881
[8,]	48	0.00954918

→

	Retard	p-value
[1,]	6	0.67623553
[2,]	12	0.05109384
[3,]	18	0.02568330
[4,]	24	0.05918747
[5,]	30	0.11680779
[6,]	36	0.09623856
[7,]	42	0.06119261
[8,]	48	0.07062617

- En ambos casos, el proceso residual se puede considerar prácticamente RB.

Tratamiento calendario en series mensuales

- En las series mensuales existen ciertas definiciones que permiten reflejar en el modelo el efecto del calendario (días laborables y festivos), de la Semana Santa (no siempre cae en el mismo mes) y de los años bisiestos (cada 4 años):

  - **Efecto calendario.-** se puede reflejar mediante alguna de las siguientes definiciones:

    - Una variable continua: D_t = nº días laborables - $\frac{5}{2}$(nº sábados + nº domingos)
    - 7 variables, asociadas cada una de ellas a cada día de la semana:

$$\boldsymbol{D_t}^{\boldsymbol{i}} = \boldsymbol{W_t}^{\boldsymbol{i}} - \boldsymbol{W_t}^{\boldsymbol{7}} \quad \forall \boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{6} \qquad \boldsymbol{D_t}^{\boldsymbol{7}} = \boldsymbol{W_t}^{\boldsymbol{1}} + \cdots + \boldsymbol{W_t}^{\boldsymbol{7}}$$

    siendo $\boldsymbol{W_t}^{\boldsymbol{i}}$ el número de días de tipo "i" en el mes "t"

  - **Semana Santa.-** se puede reflejar mediante alguna de las siguientes definiciones:

    - Una variable tipo pulso que toma el valor 1 el mes del año en el que cae la mayor proporción de la Semana Santa
    - Una variable cuantitativa que marca, por mes, el número de días de la Semana Santa que tiene

  - **Año bisiesto.-** variable binaria que toma el valor 1 los febreros de 29 días

Identificación de *outliers*

- Como norma general, **debe recurrirse a la inclusión de *outliers* en una serie:**

  - **Una vez que se hayan introducido intervenciones justificadas.**

  - **Una vez que se hayan introducido variables exógenas justificadas.**

  - **Si su influencia en la gráfica de la serie es evidente.**
    Ejemplo: escalón resultante de la crisis económica en una serie de dicha naturaleza, pico de ausencia derivado de una noticia de impacto, etc.

  - **Cuando** habiendo contemplado los tres puntos anteriores, **no se tenga ruido blanco y, la inclusión de algunos (pocos) de ellos, permite conseguirlo.**

Identificación de *outliers*

- Para la **identificación de *outliers***, la sintaxis es la siguiente:

```
> listaOutliers <- locate.outliers(ajuste6ConFestivos$residuals,
+                                  pars = coefs2poly(ajuste6ConFestivos),
+                                  types = c("AO", "LS", "TC"),cval=5)
> listaOutliers
   type  ind   coefhat     tstat
2    AO  270  476.0949  5.701096
5    AO  651  483.0757  5.784617
6    AO  652 -704.7499 -8.439067
7    AO  834 -427.7705 -5.121532
8    AO  906  638.6052  7.643194
11   AO 1008  522.4300  6.242095
12   AO 1009 -874.6113 -10.450009
16   LS  275  622.4003  5.025186
19   LS  653  791.4321  6.389528
20   LS  907 -1497.4387 -12.059899
21   LS  918  650.9131  5.239840
24   LS 1010  999.6574  7.980121
25   TC  177 -731.6190 -6.269150
26   TC  271 -957.8952 -8.208081
27   TC  274 -802.7676 -6.878813
28   TC  542 -742.9889 -6.366561
30   TC  831 -612.3866 -5.246577
32   TC 1006 -938.6909 -8.026283
```

t-ratio a partir del cual se consideran significativos los *outliers*. Cuanto mayor es su valor, mayor es el grado de significatividad que se exige

```
> outliers <- outliers(c("AO", "AO", "LS", "TC"), c(270, 651, 275, 177))
> outliersVariables <- outliers.effects(outliers, length(ajuste6ConFestivos$residuals))
> calendarioMasOutliers <- cbind(calendario.train,outliersVariables)
> ajusteConOutliers <- Arima(datos.train.ts,
+                            order = c(1,0,2),
+                            seasonal = list(order = c(1,1,1), period=7),
+                            method="ML",
+                            xreg = calendarioMasOutliers)
```

Inclusión de variables explicativas

- Para la inclusión de variables explicativas cualesquiera (de intervención / *outliers* / otras series), se debe:

1. Incluir dichas variables en una **estructura de tipo matriz.**

```
calendarioTrain <-
  as.matrix(
    explicativasCalendarioTrain[,c("semanaSanta", "dt", "bisiesto")]
    )
```

2. Incluir dicha estructura matricial en la función *Arima* a través del **parámetro *Xreg*.**

```
ajusteconCalendario <- Arima(datos.train.ts,
                             order = c(0,1,1),
                             seasonal = list(order = c(0,1,1),
                                             period = 12),
                             xreg = calendarioTrain,
                             method = "ML")
```

3. Para realizar una **predicción a futuro**, es preciso declarar **otra estructura de tipo matriz e incluirla en la función *predict* a través del parámetro *newxreg*.**

```
calendarioTest <-
  as.matrix(
    explicativasCalendarioTest[,c("semanaSanta", "dt", "bisiesto")]
    )
```

5 Modelos de función de transferencia

Modelos de función de transferencia

- El análisis presentado aprovecha la información histórica de la serie X_t para explicar su comportamiento y predecir su valor esperado. Sin embargo, en ocasiones, existen otras variables exógenas Z_t que pueden condicionar dicho comportamiento.

HIPÓTESIS DE PARTIDA

- Se supondrá que entre Z_t y X_t existe una relación de **causalidad unidireccional**, es decir, Z_t condiciona el valor de X_{t+k}, pero no al revés.

- Las series X_t y Z_t son **estacionarias**. De no ser así, se realizarán las transformaciones necesarias.

- Bajo dichas hipótesis, se plantea el modelo de **función de transferencia** dado por:

$$X_t = \frac{(w_0 + w_1 B + w_2 B^2 + \cdots + w_m B^m) B^b}{(1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_n B^n)} Z_t + \varepsilon_t = V(B) Z_t + \varepsilon_t$$

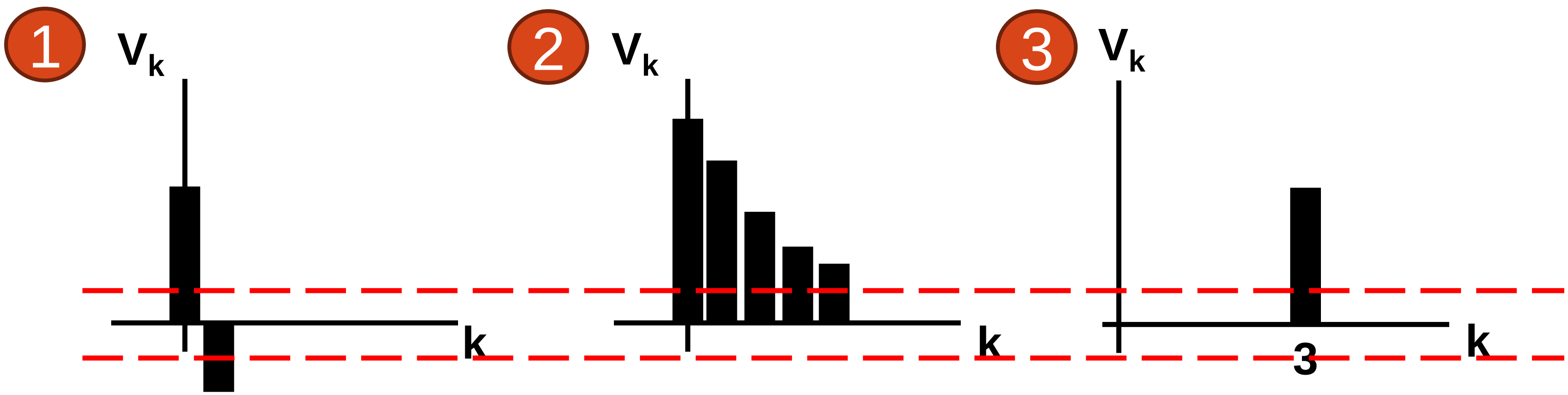
- Se trata de identificar los órdenes m, n y b de este modelo que reflejan **cómo se transfiere el efecto** de la variable Z_t a la variable X_t.

Modelos de función de transferencia

- El método más popular es el de aproximación finita, consistente en ajustar un modelo genérico del tipo:

$$X_t = V_0 Z_t + V_1 Z_{t-1} + V_2 Z_{t-2} + \cdots + V_h Z_{t-h} + a_t = V_h(B) Z_t + a_t$$

donde "h" es suficientemente grande y observar el patrón al que responden los coeficientes V's teniendo en cuenta su significatividad.



m = 1, n = 0, b = 0

m = 0, n = 1, b = 0

m = 0, n = 0, b = 3

Modelos de función de transferencia
Método de aproximaciones finitas

- Para aplicar el método de aproximaciones finitas, basta declarar retardos de la variable explicativa en cuestión con la función LAG:

```
datos.train.ts$TEMPERATURA_MAXIMA_1=lag(dplyr::lag(datos.train.ts$TEMPERATURA_MAXIMA,1))
datos.train.ts$TEMPERATURA_MAXIMA_2=lag(dplyr::lag(datos.train.ts$TEMPERATURA_MAXIMA,2))

datos.train.ts$TEMPERATURA_MAXIMA_1[is.na[datos$datos.train.ts$TEMPERATURA_MAXIMA_1]<-0
datos.train.ts$TEMPERATURA_MAXIMA_2[is.na[datos$datos.train.ts$TEMPERATURA_MAXIMA_2]<-0
```

  e incluirla en la función *Arima* a través de una estructura matricial con el parámetro *xreg*.

- Obsérvese la necesidad de reemplazar por 0 los valores NA generados por la función LAG.

- Finalmente, en función de los p-valores asociados a los parámetros de las distintas variables, se irían excluyendo aquellas variables que no resultaran significativas.

Modelos de función de transferencia
Método de aproximaciones finitas

- Además de la significatividad, es importante observar si el parámetro responde al sentido de negocio esperado.

```
Series: datos.train.ts
Regression with ARIMA(1,0,2)(1,1,1)[7] errors

Coefficients:
         ar1     ma1      ma2     sar1      sma1  TEMPERATURA_MAXIMA
      0.9471  0.1945  -0.0659   0.1931   -0.9504             -9.8478
s.e.  0.0115  0.0347   0.0324   0.0325    0.0103              1.8505
      TEMPERATURA_MINIMA    p_01ene    p_06ene    p_19mar    p_15ago
                 -9.3881  -994.3874  -933.1004  -446.9589  -219.6346
s.e.              2.4392    74.0880    71.0320    70.4150    69.7590
        p_01may    p_12oct    p_01nov    p_06dic    p_08dic    p_25dic
      -382.5814  -301.3256  -141.5147  -220.2918  -344.0241  -942.8754
s.e.    69.9698    70.4888    69.7438    70.9681    71.1909    79.4482

sigma^2 estimated as 38105:  log likelihood=-7280.67
AIC=14597.33   AICc=14597.97   BIC=14687.19
```

Fase de contraste: análisis del residuo

- Se presenta el test de ruido blanco una vez incluidos las funciones de transferencia para los dos modelos candidatos:

SARIMA(1,0,2)x(1,1,1)$_7$

	Retard	p-value
[1,]	6	0.26783113
[2,]	12	0.06479063
[3,]	18	0.04728734
[4,]	24	0.10887151
[5,]	30	0.20118318
[6,]	36	0.17537176
[7,]	42	0.11347993
[8,]	48	0.13102472

	Retard	p-value
[1,]	6	0.21338292
[2,]	12	0.13189153
[3,]	18	0.05569565
[4,]	24	0.11705951
[5,]	30	0.27349217
[6,]	36	0.35129454
[7,]	42	0.31161552
[8,]	48	0.31202151

SARIMA(0,1,2)x(1,1,1)$_7$

	Retard	p-value
[1,]	6	0.67623553
[2,]	12	0.05109384
[3,]	18	0.02568330
[4,]	24	0.05918747
[5,]	30	0.11680779
[6,]	36	0.09623856
[7,]	42	0.06119261
[8,]	48	0.07062617

	Retard	p-value
[1,]	6	0.70147491
[2,]	12	0.11811131
[3,]	18	0.03307123
[4,]	24	0.06776147
[5,]	30	0.17699596
[6,]	36	0.23309023
[7,]	42	0.22225774
[8,]	48	0.21827476

- El primer modelo supera, para un nivel de significación de 0,05, el test de RB

6 Ajuste masivo de series temporales

Ajuste masivo de series temporales

- En ocasiones el objetivo es **obtener predicciones para un volumen alto de series:**

  - Series de demanda energética a nivel de CUPS en el sector "*utilities*".
  - Series de ventas a nivel de referencia en el sector "*retails*", etc.

- Si las series tienen un **comportamiento independiente,** se puede aplicar un procedimiento de **ajuste automático** sobre cada una de ellas. Dicho ajuste es **paralelizable** y por tanto abordable bajo la óptica de **Big Data.**

- Más allá del **nivel de detalle temporal** (hora, día, semana, mes, etc.) al que interesa hacer las predicciones y el horizonte de predicción (número de unidades temporales), es importante saber también a qué **nivel físico** se pretende hacer:

  - En un problema de demanda energética, puede interesar la predicción a nivel global de cara a comprar energía en el mercado o a nivel de cliente (clientes industriales de entidad).
  - En un problema de demanda de productos ("*retails*"), interesa la predicción a nivel de referencia de cara a hacer una buena gestión del stock y evitar la venta perdida.

- Independientemente del nivel al que interesa obtener las predicciones, la manera de llegar a él puede ser diferente sin que tenga que existir un criterio claramente mejor.

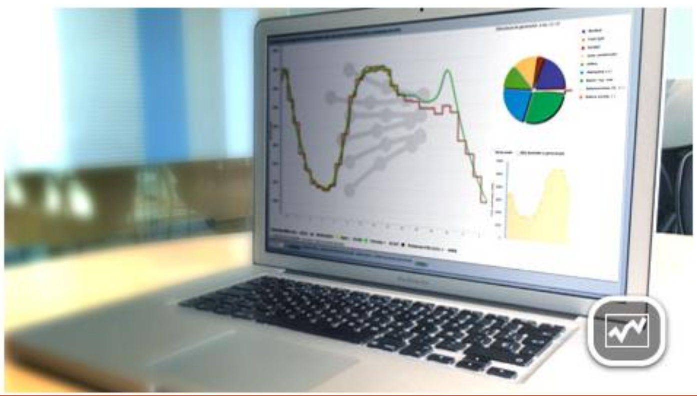
Ajuste masivo de series temporales

Caso de estudio: estimación de la demanda eléctrica de una cartera

- Las compañías eléctricas compran electricidad diariamente para suministrar a su conjunto de clientes (CUPS). La compra de esta energía se hace a Red Eléctrica de España (REE, http://www.ree.es/es/), que recibe los datos de demanda esperada a nivel diario de todas las compañías del sector y es la encargada de gestionarla contactando con las generadoras.

Demanda y producción en tiempo real

Red Eléctrica representa en los siguientes gráficos la demanda de energía que se está produciendo en el sistema eléctrico peninsular en tiempo real. Estos gráficos se actualizan cada diez minutos e incluyen datos de la demanda real, prevista y programada, así como los valores de máximos y mínimos de la demanda diaria.



RED ELÉCTRICA DE ESPAÑA

Balance diario

Es el detalle diario de la producción y del consumo de energía eléctrica en los sistemas peninsular y no peninsulares (programación para el día en curso y cierre de los días anteriores).

El balance incluye gráficos que muestran la estructura de generación necesaria para la cobertura de la demanda y el porcentaje de participación de fuentes de energía renovable y no renovable.

Además contiene el dato de la demanda de energía eléctrica corregida por temperatura y laboralidad, es decir, eliminando la influencia que el calendario laboral y las temperaturas ejercen en la demanda energética, así como los máximos de demanda horaria y diaria.

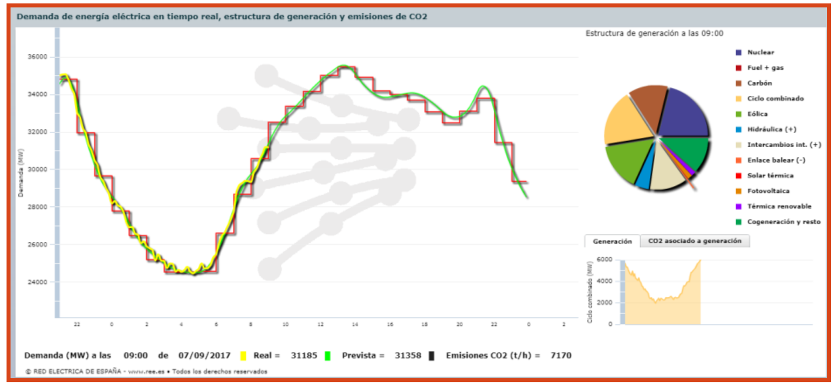
Desde los siguientes formularios se accede a los balances diarios nacional, peninsular, de Ceuta y de Melilla. Los balances correspondientes a Islas Baleares e Islas Canarias están disponibles desde las correspondientes subsecciones de balance ubicadas en las respectivas secciones de Sistema eléctrico balear y Sistema eléctrico canario.

Ajuste masivo de series temporales

Caso de estudio: estimación de la demanda eléctrica de una cartera

- El error en las estimaciones de la demanda diaria puede conllevar penalizaciones. Si una compañía se queda por debajo del dato real de consumo, y alguna o varias de las empresas de la competencia les ha sobrado energía, ésta se revende a precio de mercado. En caso contrario, REE suministrará esa energía a los clientes, pero el coste asociado a esa energía tendrá un precio más elevado. Del mismo modo, si la compañía da una estimación superior al dato real y la energía sobrante puede recolocarse, no habrá penalización económica. Por ello cobra especial importancia que **la estimación sea lo más precisa posible.**

- Por otro lado, aunque la compra de energía se hace a nivel diario, las compañías pueden **corregir su estimación en base al comportamiento observado de la demanda en algunos momentos puntuales del día a través del mercado intradiario** (http://www.omel.es/inicio/mercados-y-productos/mercado-electricidad/nuestros-mercados-de-electricidad/mercado-intradiar).



- Cobra por ello importancia realizar una estimación horaria de la demanda.

Ajuste masivo de series temporales

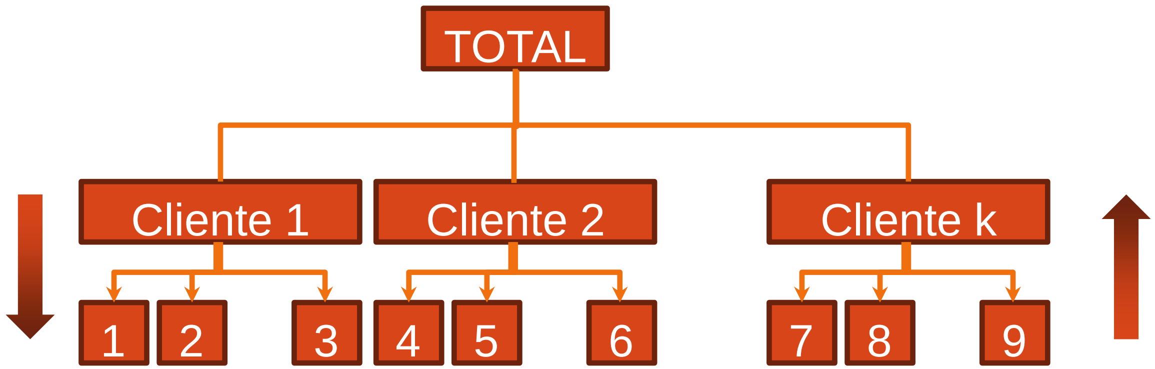
Caso de estudio: estimación de la demanda eléctrica de una cartera

- Existen diferentes **estrategias de predicción**, pero en todas ellas se ajustan modelos al nivel más bajo, para obtener la predicción global por agregación o promedio de ellas o desagregar predicciones a niveles superiores de acuerdo a distribuciones porcentuales.

1

TOP - DOWN

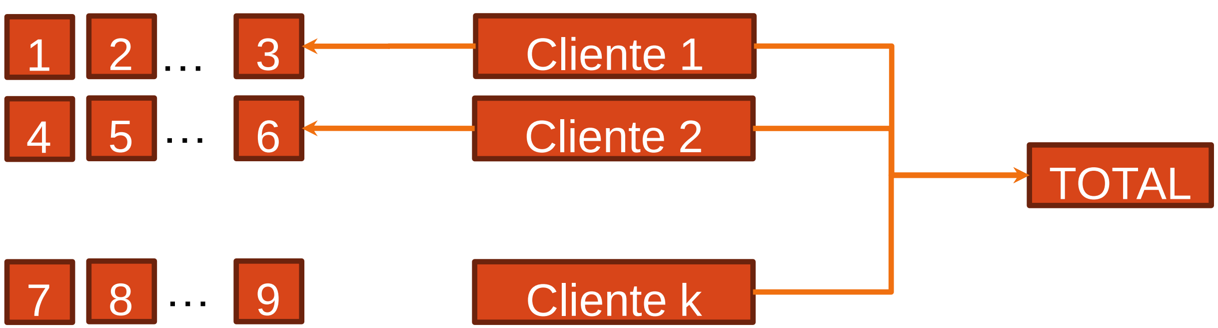
Permite eliminar el ruido excesivo de los datos de los niveles más bajos de la jerarquía



2

BOTTOM - UP

Permite observar patrones que solo son observables en los niveles más bajos de la jerarquía



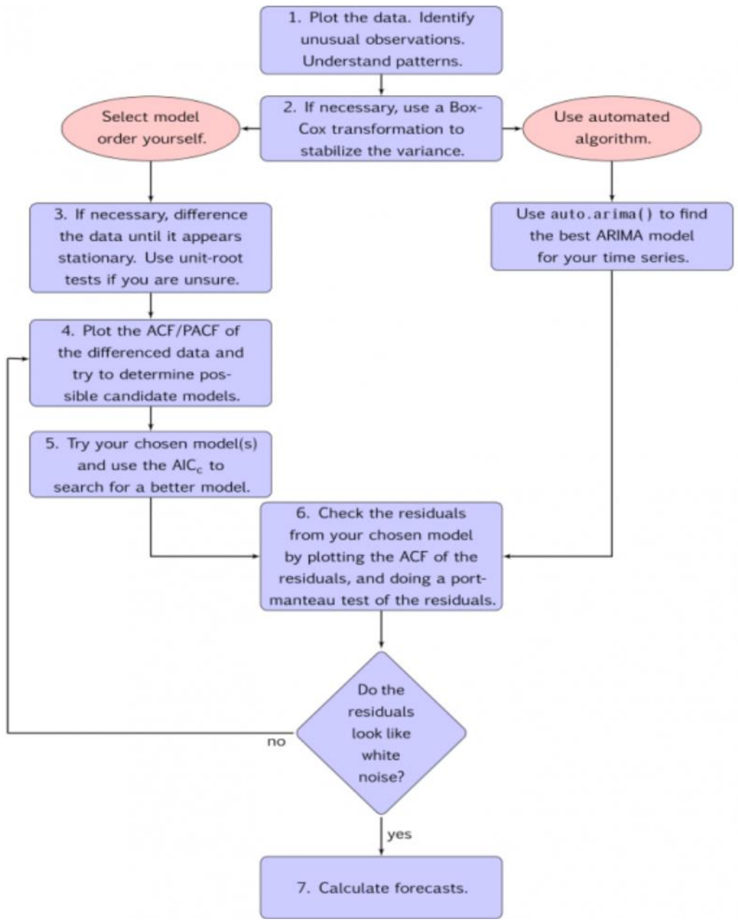
3

MIDDLE - OUT

Centra la predicción en un nivel intermedio de la jerarquía que resulta de interés para la compañía o agrega series con características similares

Ajuste masivo de series temporales

Caso de estudio: estimación de la demanda eléctrica de una cartera



- El proceso de ajuste de una serie temporal lleva asociadas las fases de la figura de la izquierda.
- En caso de predecir una única serie, el ajuste manual es lo más apropiado.
- Sin embargo, si interesan las predicciones a bajo nivel, el elevado número de series lleva a **aplicar algún tipo de ajuste automático.**
- Al realizar un ajuste automático, se contrastan diferentes modelos y **se selecciona aquél que, respecto de alguna métrica, tiene mejor valor.**
- **Se descuidan los contrastes de bondad** asociados a los parámetros (correlaciones, estacionariedad, etc.) y al residuo (ruido blanco).

Ajuste masivo de series temporales

Caso de estudio: estimación de la demanda eléctrica de una cartera

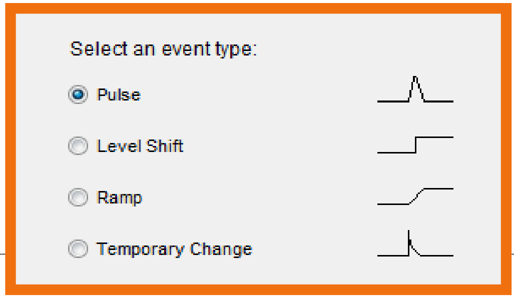
- Algunas opciones que es preciso configurar en el proceso de ajuste automático de modelos son:

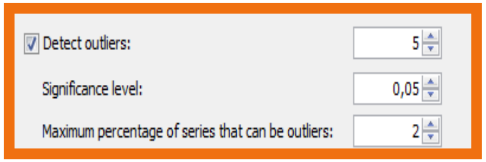
  1. **Depuración de datos:**

     - Se pueden considerar los tramos de ceros como valores "*missing*".
     - Se debe especificar cómo realizar imputaciones de datos "*missing*" para la variable que se desea predecir. Lo más sensato es utilizar la predicción del propio modelo.

  2. **Identificación automática de transformación**

  3. **Tipología de eventos a identificar, cantidad máxima de ellos, nivel de significatividad**





Ajuste masivo de series temporales
Caso de estudio: estimación de la demanda eléctrica de una cartera

- Algunas opciones que es preciso configurar en el proceso de ajuste automático de modelos son:

4. Estrategia de inclusión de variables exógenas e intervenciones:

- Identificarlos inputs antes de identificar las componentes (p,q) del modelo ARMA.
- Identificarlos inputs después de identificar las componentes (p,q) del modelo ARMA.
- Contrastar los dos métodos y seleccionar el mejor.
- Incluir las variables en función de su nivel de significatividad y/o su sentido de negocio.

5. Criterio de selección del modelo:

- Minimizar MAPE, MAE, AIC, BIC, SBC, etc.
- Maximizar R^2, verosimilitud, etc.

Dicha métrica puede ser evaluada *in-sampling* (a histórico) o *out-of-sampling* (a futuro).

Ajuste masivo de series temporales

Caso de estudio: estimación de la demanda eléctrica de una cartera

- La función de R que permite el ajuste automático de series temporales es *auto.arima*:

```
> ajusteTOP <- auto.arima(serieTemporal,
+                         max.d=1, max.D=1,
+                         max.p=2, max.P=2,
+                         max.q=2, max.Q=2,
+                         seasonal=TRUE,
+                         ic="aic",
+                         allowdrift=FALSE,
+                         xreg=inputs.train,
+                         stepwise = TRUE)
```

← ***Diferencias***
← ***Parte AR***
← ***Parte MA***
← ***Ajusta estacionales (*)***
← ***Métrica***
← ***Intercept***
← ***Stepwise regresores***

(*) Precisa de la definición de la opción frequency en el objeto TS

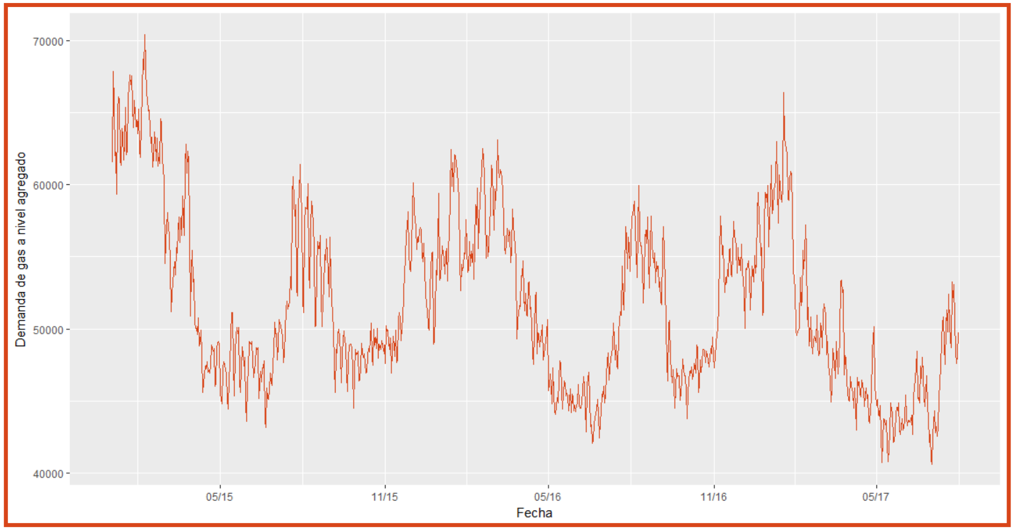
Ajuste masivo de series temporales
Caso de estudio: estimación de la demanda eléctrica de una cartera

- <u>Ejemplo:</u> predicción de consumo horario de la demanda eléctrica del total de los clientes de una compañía.

- Se ajustará la serie agregada de demanda diaria, obtenida como suma de todos los consumos de cada uno de los CUPS (puntos de suministro), con datos entre enero de 2015 y julio de 2017.

- En un paso posterior se desagregará dicha predicción a nivel horario mediante un modelo lineal generalizado (GLM).

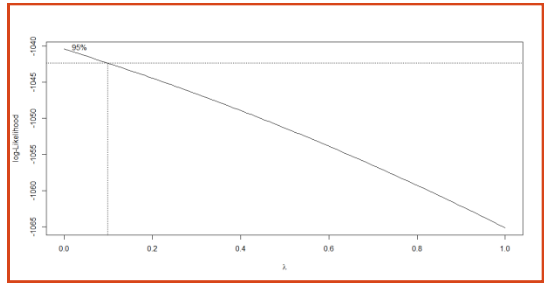
- El último mes se utilizará para validar el resultado.



Ajuste masivo de series temporales

Caso de estudio: estimación de la demanda eléctrica de una cartera

- La serie diaria no es estacionaria en varianza. El test de *Box-Cox* indica que hay que realizar una transformación logarítmica
- El modelo seleccionado automáticamente es un



```
> lambda
[1] 0
```

```
> ajusteTOP
Series: serieTOP.train.ts
Regression with ARIMA(1,1,1)(2,0,0)[7] errors

Coefficients:
         ar1      ma1    sar1    sar2  p_01ene  p_06ene  p_19mar  p_01may  p_15ago
      0.8259  -0.9585  0.3327  0.2247  -0.0649  -0.0403  -0.0156  -0.0455  -0.0241
s.e.  0.0403   0.0241  0.0340  0.0345   0.0128   0.0117   0.0115   0.0114   0.0140
      p_12oct  p_01nov  p_06dic  p_08dic  p_25dic  escalonJulio  TempMaxNac TempMinNac
      -0.0173   -0.022  -0.0355   -0.007  -0.0554        0.0268     -0.0044      1e-03
s.e.   0.0140    0.014   0.0140    0.014   0.0142        0.0131      0.0006      8e-04

 sigma^2 estimated as 0.0008075:  log likelihood=1958.91
 AIC=-3881.82   AICc=-3881.05   BIC=-3795.15
```

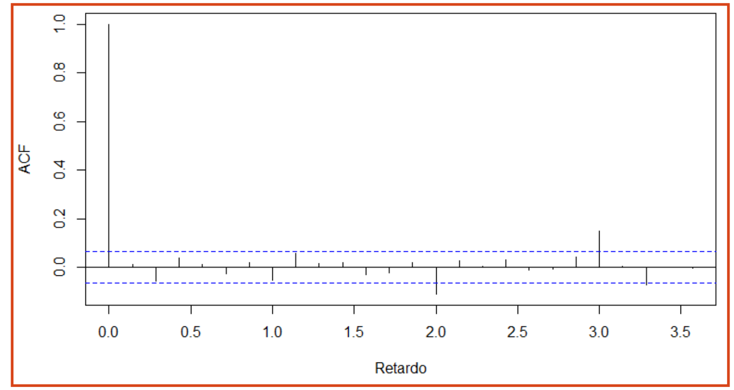
El efecto de los festivos sobre la serie tiene el sentido de negocio esperado (las estimaciones de sus parámetros son negativas)

Ajuste masivo de series temporales

Caso de estudio: estimación de la demanda eléctrica de una cartera

- La **correlación entre los parámetros es alta** para el AR(1) con el MA(1):

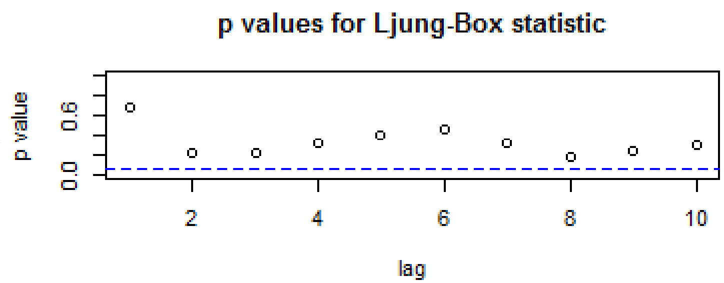
```
> cor.arma(ajusteTOP)
                       ar1           ma1          sar1         sar2
ar1           1.000000e+00 -0.8795268250  0.0800836540  0.258754375
ma1          -8.795268e-01  1.0000000000 -0.1929951228 -0.273820291
sar1          8.008365e-02 -0.1929951228  1.0000000000 -0.370901881
sar2          2.587544e-01 -0.2738202911 -0.3709018808  1.000000000
p_01ene       5.584448e-06 -0.0055241693 -0.0224652919  0.031120937
p_06ene       2.841779e-02 -0.0228930248 -0.0399438863  0.054904384
p_19mar      -2.978507e-02  0.0235138547  0.0906752890 -0.137330333
p_01may      -2.116488e-02  0.0261897347 -0.0253658261 -0.014698223
p_15ago       2.355060e-02 -0.0249497775  0.0416906249 -0.020546410
p_12oct      -5.895752e-03  0.0086720335 -0.0212114929  0.007969367
p_01nov       2.939743e-02 -0.0264077619 -0.0519458564  0.098072215
p_06dic      -3.928521e-03 -0.0003168365  0.0206084541 -0.012625227
p_08dic      -7.904404e-03  0.0038621405  0.0375116374 -0.042947732
p_25dic       2.796196e-02 -0.0247249298 -0.0024659585  0.009128911
escalonJulio  2.229233e-02 -0.0310886252 -0.0605841561  0.062091904
TempMaxNac   -2.047045e-02  0.0401695470 -0.0008437512 -0.065460579
TempMinNac    6.605266e-03  0.0217852025 -0.0363077919  0.005822575
```



- En la f.a.s. del residuo no se aprecia estructura

- Se obtiene **ruido blanco** para los primeros retardos para un nivel de significación 0.05:

```
     Retard    p-value
[1,]      6 0.40790039
[2,]     12 0.34745074
[3,]     18 0.10121766
[4,]     24 0.00059031
[5,]     30 0.00087015
```



Ajuste masivo de series temporales

Caso de estudio: estimación de la demanda eléctrica de una cartera

- Las curvas de **consumo doméstico horario son muy estables**, diferenciándose su comportamiento fundamentalmente entre los días laborables y los no laborables.

- De hecho:

  - No suelen existir grandes diferencias entre un sábado y un domingo.
  - Existen diferencias entre algunos meses, pero no entre todos: agrupar por estación.

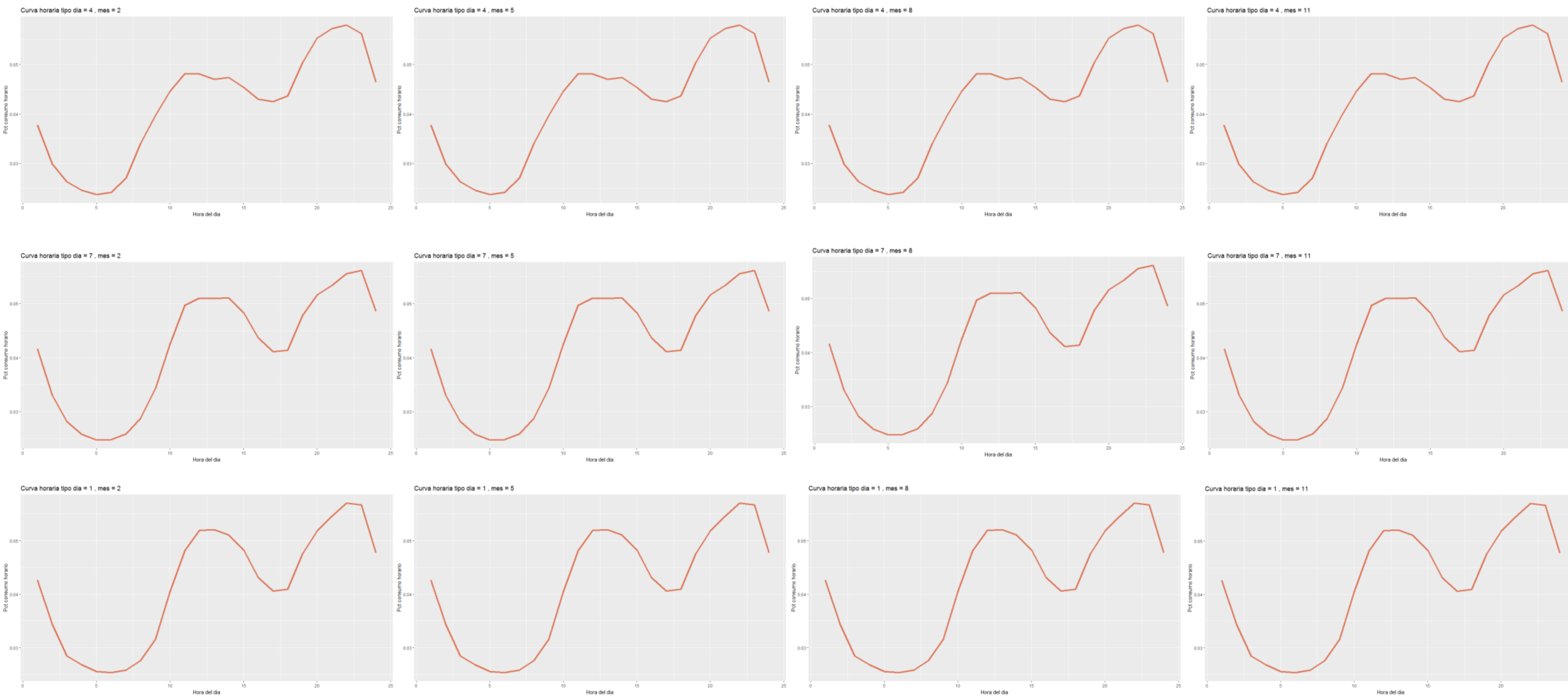
- Por ello, estas curvas de consumo horario son fácilmente predecibles teniendo en cuenta dichos factores, resultando preferible realizar una predicción a nivel diario con un modelo tipo ARIMA que sea desagregada que una predicción a nivel horario.

- Para realizar esta predicción, se puede utilizar **Modelos Lineales Generales (GLM).**

Ajuste masivo de series temporales

Caso de estudio: estimación de la demanda eléctrica de una cartera

	Febrero	Mayo	Agosto	Noviembre
Miércoles				
Sábado				
Domingo				



Ajuste masivo de series temporales

Caso de estudio: estimación de la demanda eléctrica de una cartera

```
> ajusteGLM <- glm(pctHoraio~diaSemana+mes+TempMaxNac+TempMinNac+festivo,
+                  data = datosConsumo)
```

```
Call:  glm(formula = form, data = consumoHorarioCalendario)

Coefficients:
(Intercept)   diaSemana2   diaSemana3   diaSemana4   diaSemana5
  0.0481268   -0.0010332   -0.0011279   -0.0012344   -0.0004876
 diaSemana6   diaSemana7         mes2         mes3         mes4
  0.0003331    0.0008756    0.0004265    0.0012306    0.0042055
       mes5         mes6         mes7         mes8         mes9
  0.0031815    0.0003729    0.0018121    0.0019399   -0.0004576
      mes10        mes11        mes12   TempMaxNac   TempMinNac
  0.0002591   -0.0012489   -0.0001393    0.0000688   -0.0002850
    festivo
  0.0014188

Degrees of Freedom: 942 Total (i.e. Null);  922 Residual
Null Deviance:      0.00755
Residual Deviance: 0.004166     AIC: -8907
```

7 Práctica:
Predicción de contratación de hipotecas

Descripción de la práctica

- El objetivo de la práctica es:
  - Demostrar que has adquirido los conocimientos necesarios para ajustar una Serie Temporal a través de modelos ARIMA.
  - Demostrar que podrías ajustar modelos de este tipo de forma masiva.

- Para ello, os asignaré una comunidad autónoma que tenga al menos 2 provincias (Andalucía, Aragón, Castilla y León, Castilla La Mancha, Cataluña, Comunidad Valenciana, Extremadura, Galicia, País Vasco) y debéis:
  - Construir la serie de la comunidad autónoma.
  - Especificar la secuencia de pasos que darías para ajustar un modelo ARMA a dicha serie, mostrando los modelos intermedios y los razonamientos que condujeron a proponer cada uno de ellos (ejemplos: retardos observados en el gráfico de la f.a.s / f.a.p, posible violación de la condición de estacionariedad, algún contraste asociado a los parámetros, etc.). Reflejar alguna intervención sí se considera necesario.
  - Comparar las predicciones con las que se obtendrían ajustando automáticamente un modelo a cada una de las provincias que integran la comunidad autónoma (mediante la función auto.arima) y sumando las predicciones que cada uno de ellos generaría.
  - Emplear el año 2019 para testear los resultados. Dicho año no puede participar en el proceso de ajuste de los modelos (no siendo necesario emplear todos los años desde 2003).

Afi Escuela de Finanzas

danielvelezserrano@mat.ucm.es