Cuánto y cuándo. Series temporales para investigadores y curiosos.

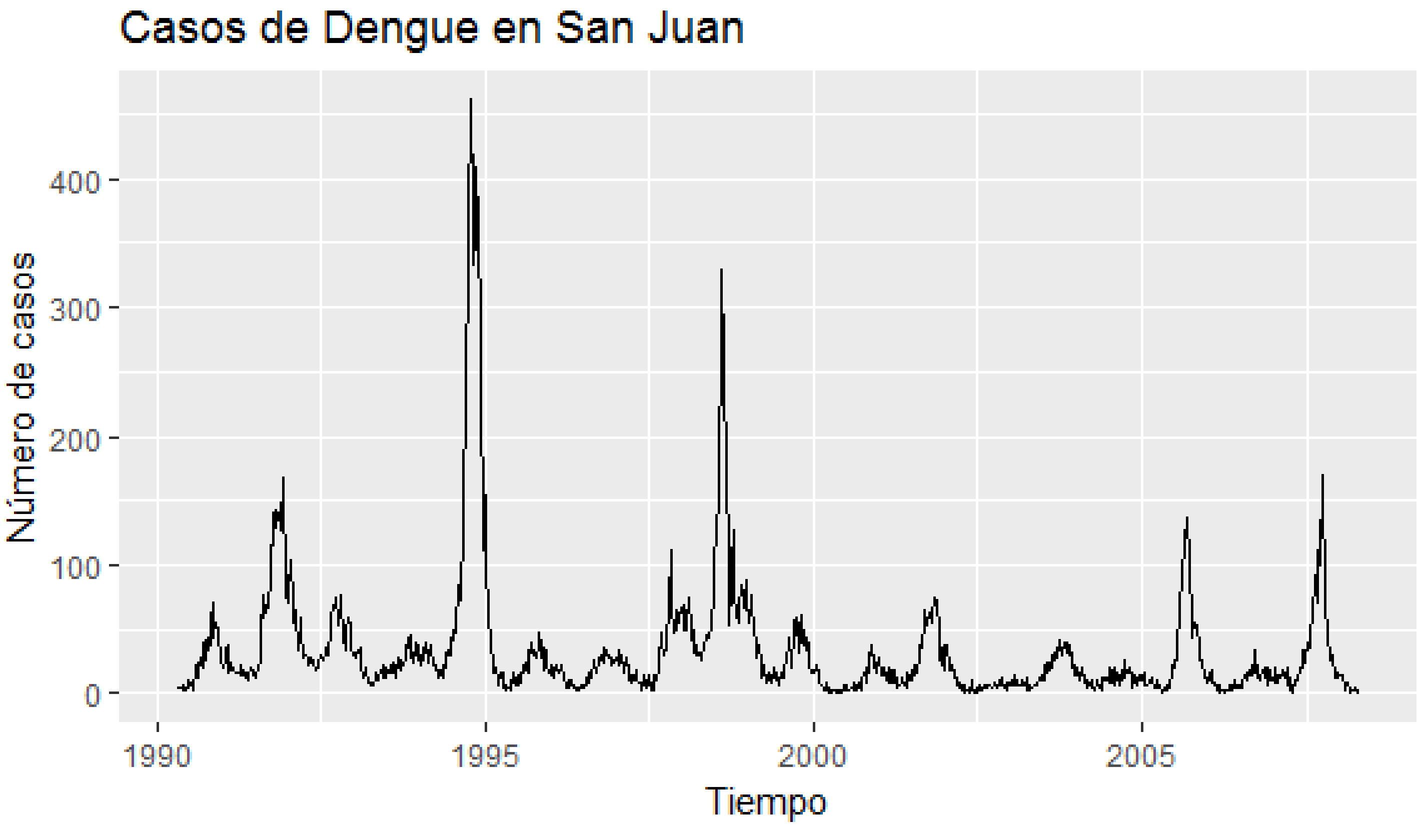
DATALAB
MEDIALAB USAL

OBJETIVOS

- SERIES TEMPORALES
- ANÁLISIS DE SERIES TEMPORALES
- APLICACIONES CON R
- ELABORACIÓN DE UNA METODOLOGÍA

¿Qué es una serie temporal?

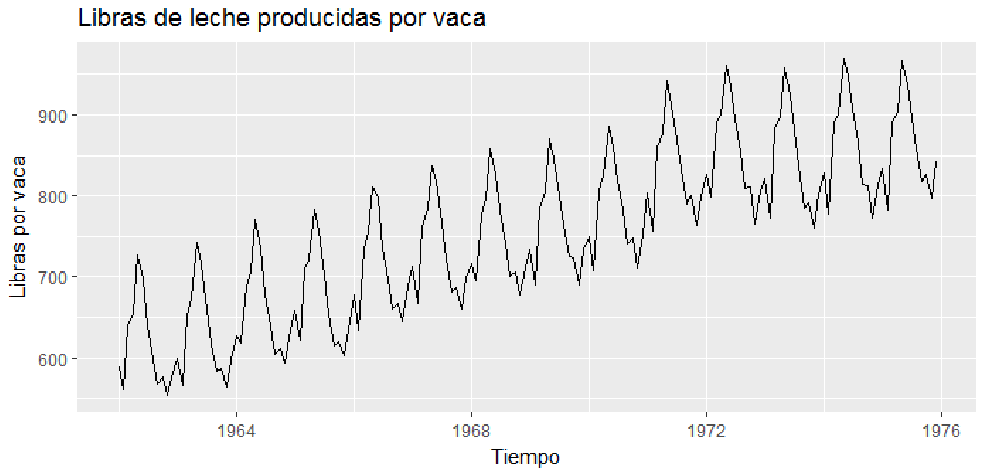
year	cases
1990	4
1990	5
1990	4
1990	3
1990	6
1990	2
.	.
.	.
.	.

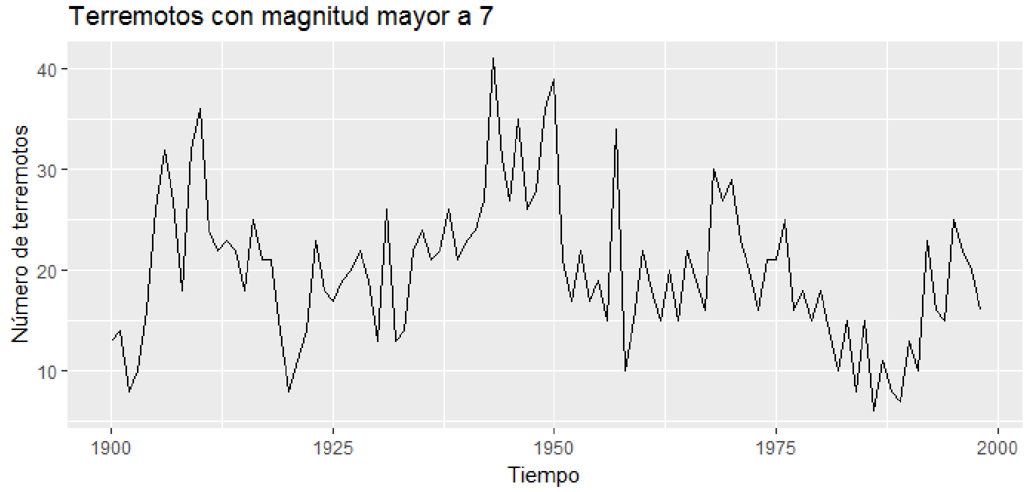


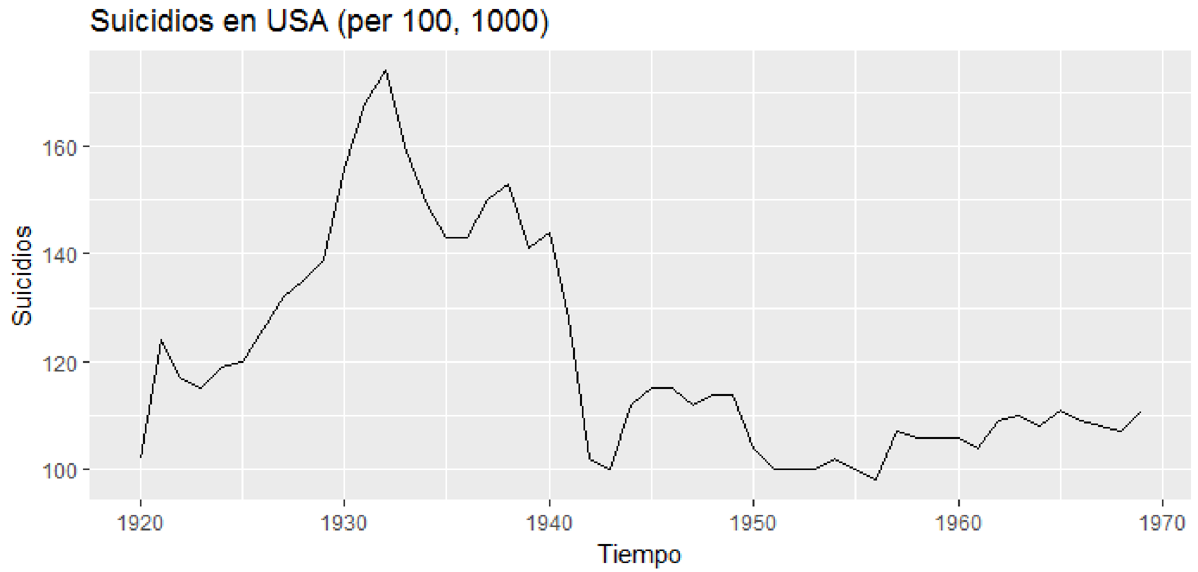
¿Qué es una serie temporal?

- Secuencia de variables aleatorias Y_1, Y_2, Y_3 ... $\rightarrow \{Y_t\}$ con $t \in N$
- Estas variables están correlacionadas temporalmente
- Parte determinista y parte aleatoria

¿Qué es una serie temporal?







¿Cómo analizamos la serie?

- Enfoque determinista
- Metodología Box-Jenkins

ENFOQUE DETERMINISTA

ENFOQUE DETERMINISTA

Enfoque determinista: Se busca expresar la serie temporal como la combinación de varias componentes. Cada una de estas componentes recoge cierta información de la serie .

$$Y_t = f(t) + a_t \text{ con } a_t \sim RB(0, \sigma^2)$$

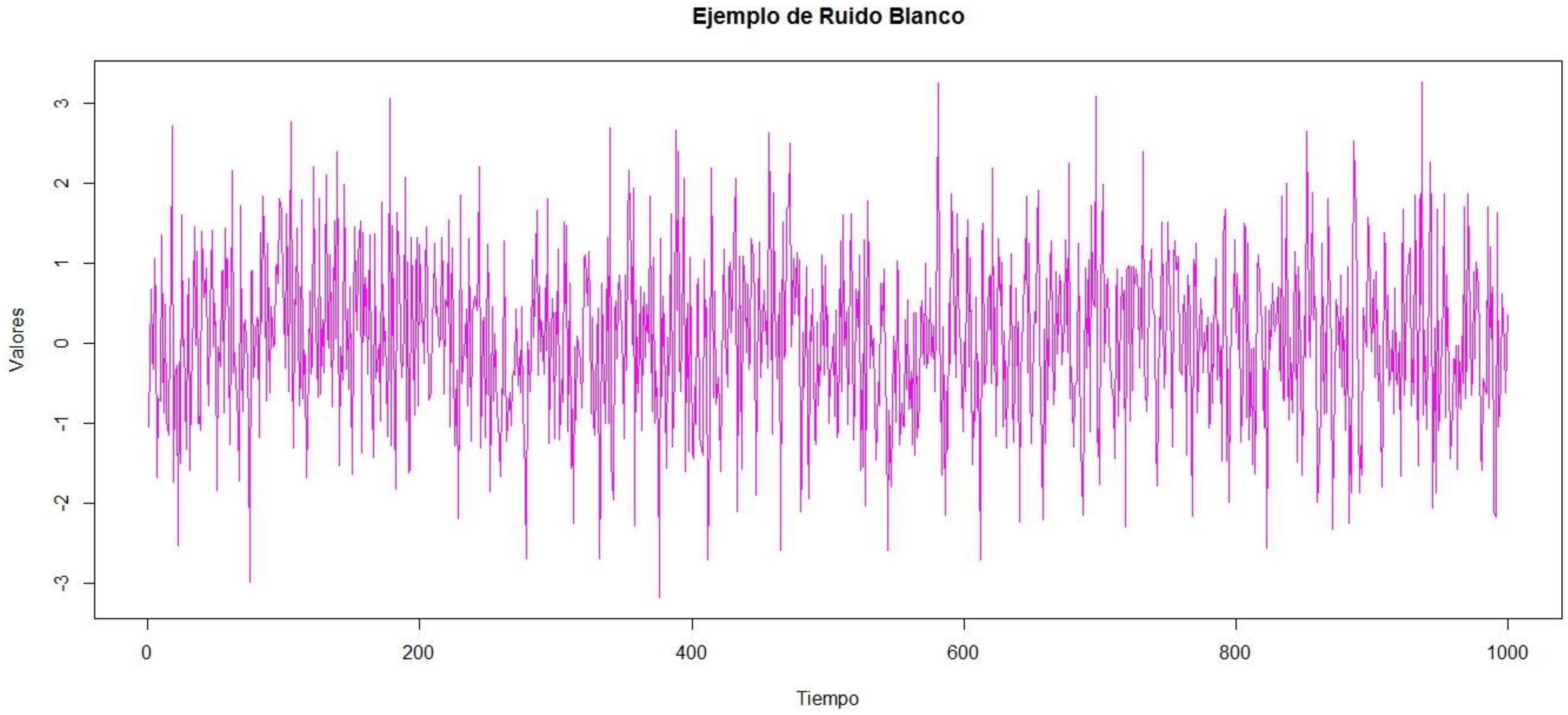
Tendencia

Estacionalidad

Ciclo

Fluctuaciones irregulares

ENFOQUE DETERMINISTA



ENFOQUE DETERMINISTA

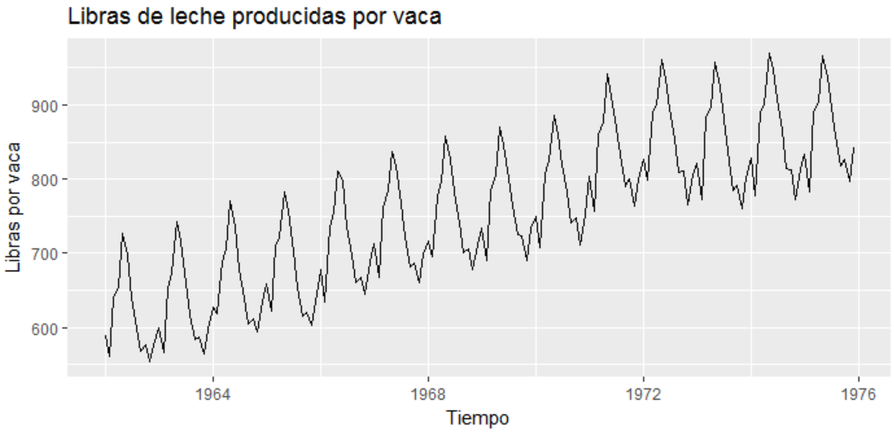
- **Tendencia (T_t):** muestra la evolución de la serie temporal a largo plazo.

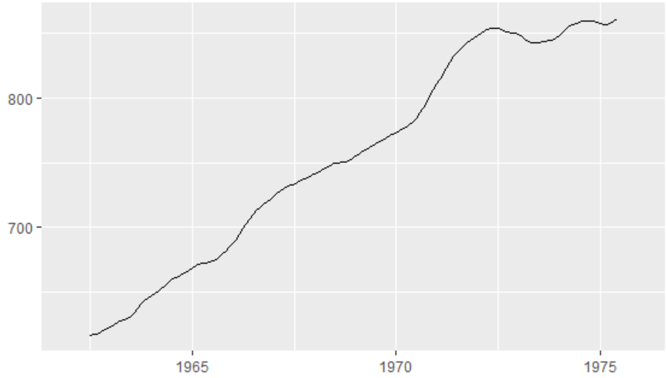
- **Estacionalidad (S_t):** variaciones que ocurren siguiendo un patrón basado en periodos temporales cortos y conocidos.

- **Ciclo (C_t):** variaciones similares a las de la estacionalidad pero que siguen periodos mucho mas largos.

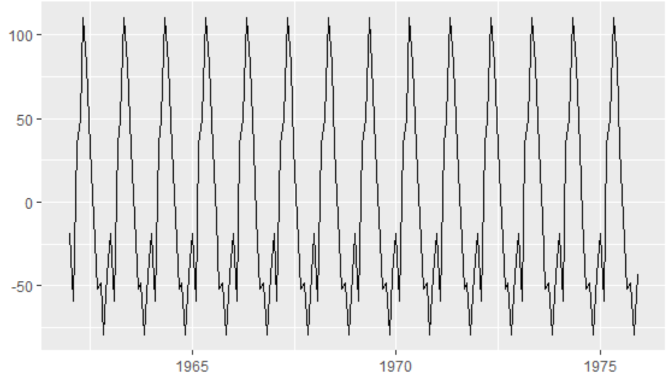
- **Fluctuaciones irregulares (a_t):** recoge las fluctuaciones de la serie que no siguen una pauta periódica ni tendencia reconocible.

ENFOQUE DETERMINISTA

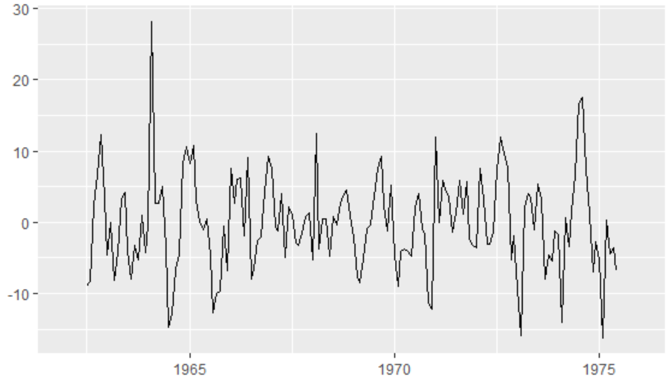




Tendencia



Estacionalidad



Fluctuaciones irregulares

Hay dos formas de combinar estas componentes:

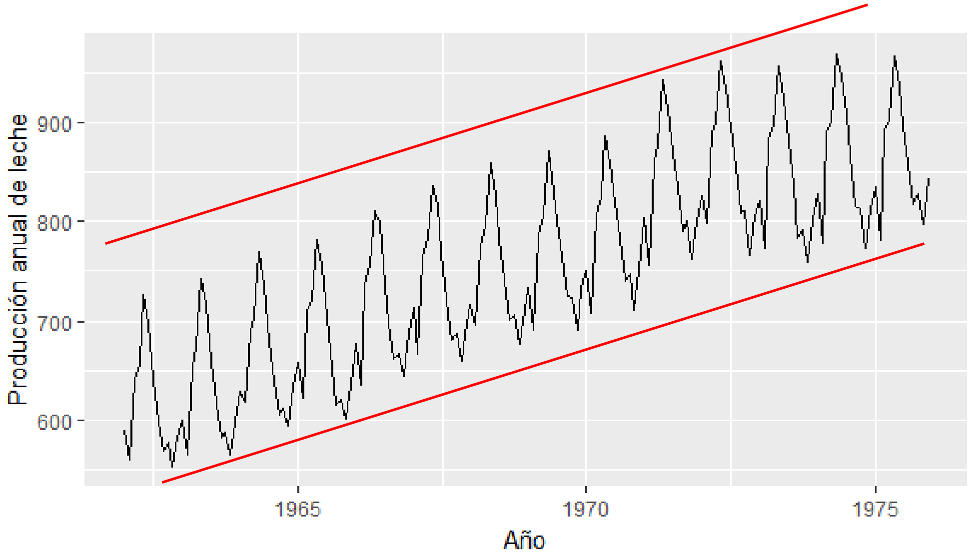
- **Modelo aditivo:** Propio de las series temporales que conservan una amplitud constante a lo largo del tiempo.

$$Y_t = T_t + S_t + C_t + \ a_t$$

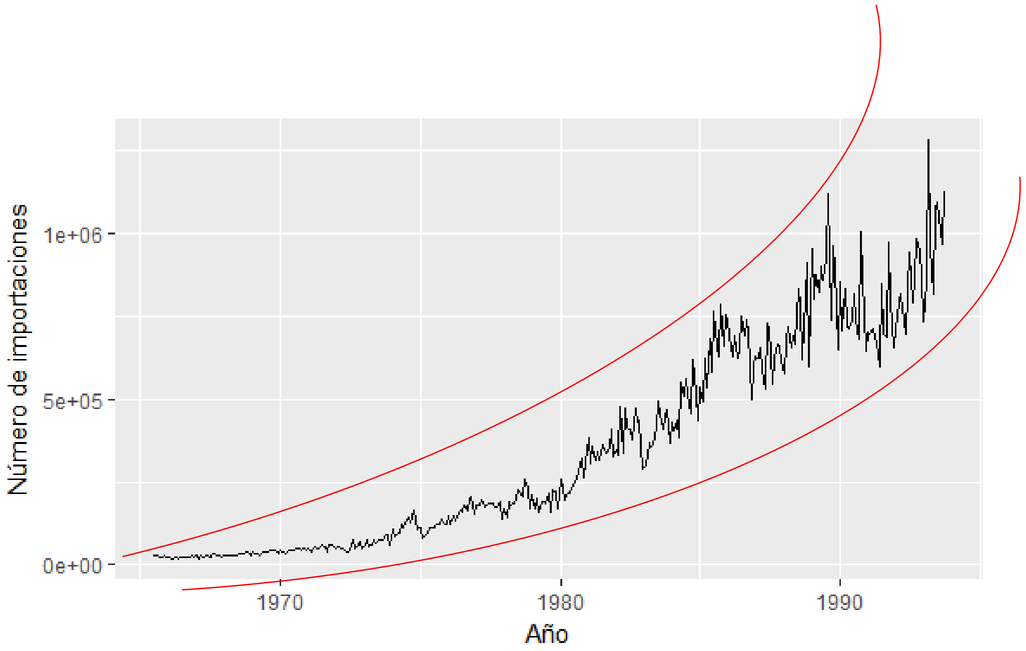
- **Modelo multiplicativo:** Este modelo se suele ajustar a aquellas series con una amplitud dependiente del tiempo

$$Y_t = T_t \times S_t \times C_t \times \ a_t$$

ENFOQUE DETERMINISTA



Estructura aditiva

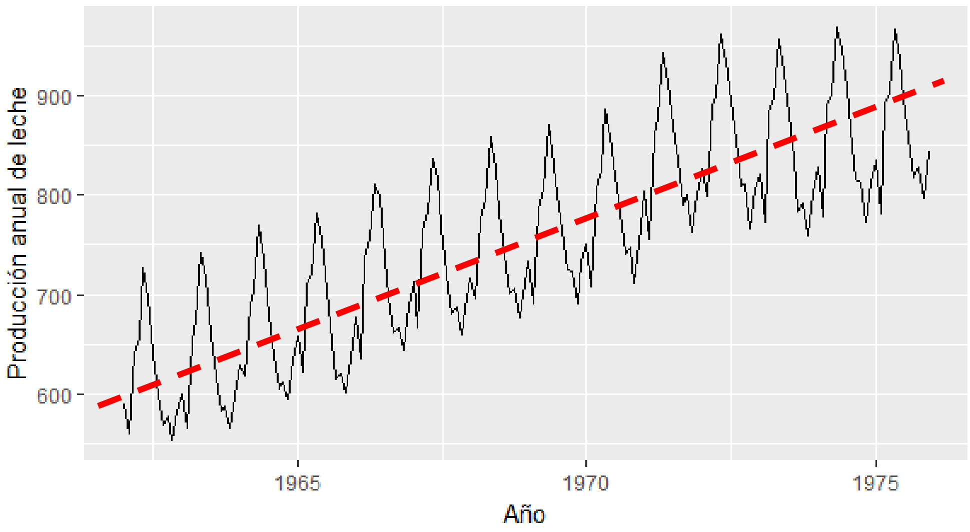


Estructura multiplicativa

¿Cómo estimo la tendencia?

- Funciones conocidas
- Técnicas de suavizado

Estimación de la tendencia a través de funciones conocidas



- **Función lineal:**

$$h(t) = \alpha_0 + \alpha_1 t + a_t$$

- **Función cuadrática:**

$$h(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + a_t$$

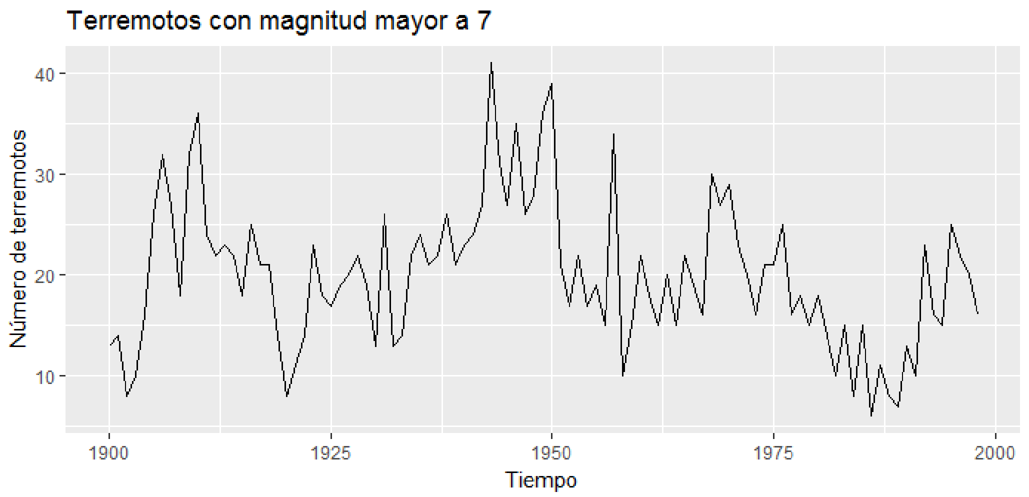
- **Función exponencial:**

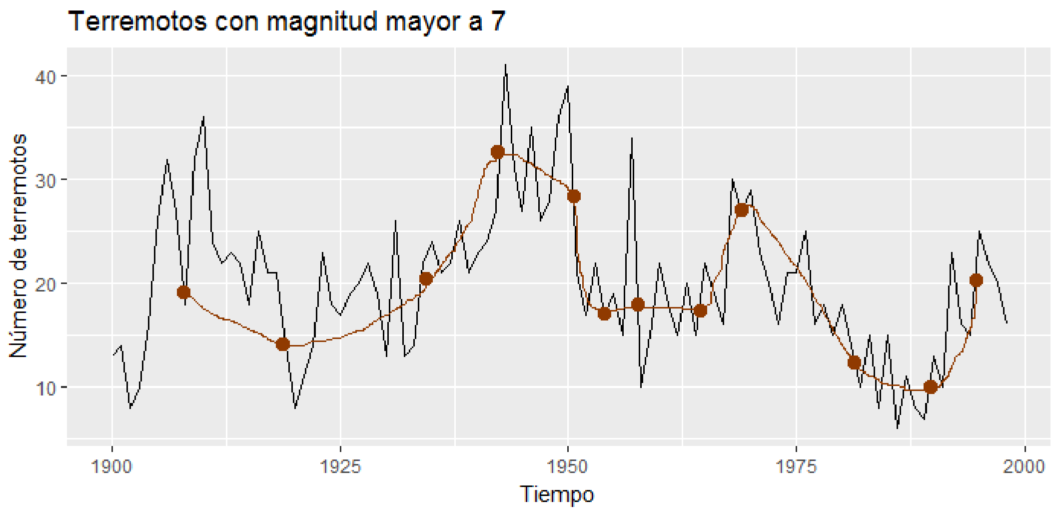
$$h(t) = e^{\alpha_0 + \alpha_1 t} + a_t$$

ENFOQUE DETERMINISTA

¿Y si mi serie no muestra comportamientos tan regulares?

Técnicas de suavizado





ENFOQUE DETERMINISTA

Suavizado de media móvil: La tendencia es el resultado de aplicar una media de orden p sobre la serie original.

$$Y_t' = \frac{1}{p}\left(Y_{t-\left(\frac{p-1}{2}\right)} + \cdots + Y_{t-1} + Y_t + Y_{t+1} + \cdots + Y_{t+\left(\frac{p-1}{2}\right)}\right)$$

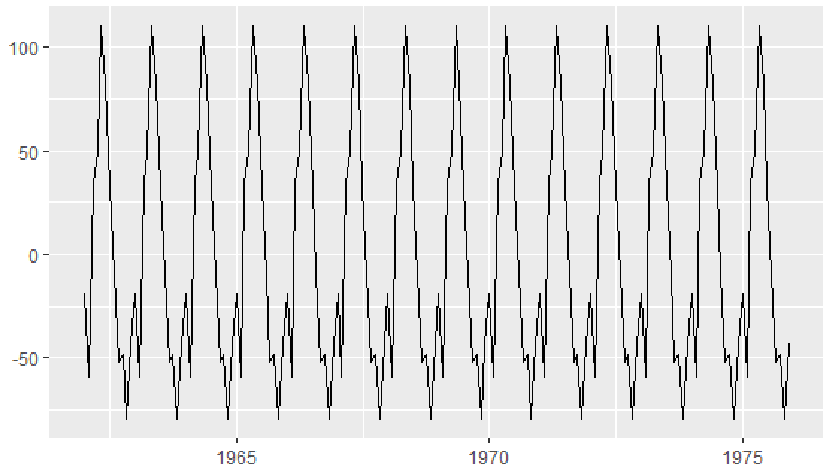
Con esta técnica conseguimos:

- Extraer la tendencia
- Eliminar ruido y fluctuaciones

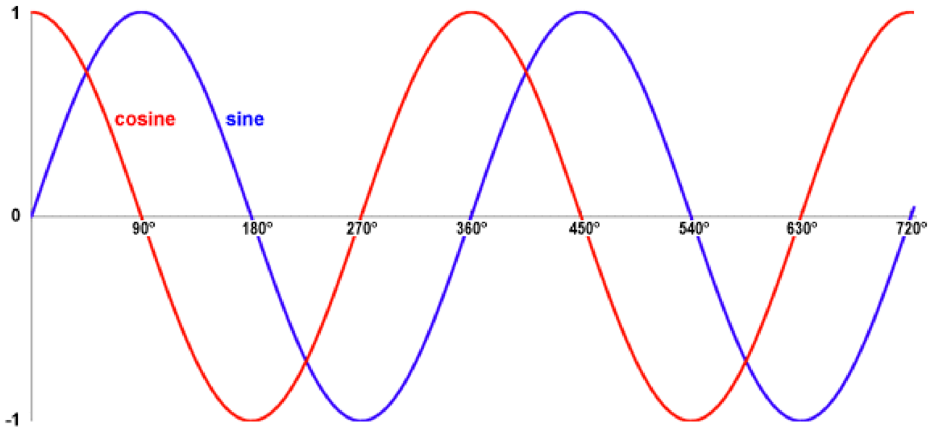
¿Cómo estimo la estacionalidad?

- Funciones periódicas
- Variables *dummy*
- “Jugando” con el resto de componentes

Estimación de la estacionalidad a través de funciones periódicas



$\approx$



$$g(t) = g(t+s) = g(t+2s)$$

Estimación de la estacionalidad a través de variables *dummy*

$$g(t) = \sum_{i=1}^{S} \lambda_i d_{it}$$

Donde:

- S es el número de periodos estacionales
- d_{it} son variables dummy que toman el valor 1 si el dato pertenece a ese periodo estacional y 0 en caso contrario
- El coeficiente λ_i indica la variación correspondiente al periodo i

Estimación de la estacionalidad jugando con las componentes

$$Y_t' = T_t + C_t \longrightarrow Y_t - Y_t' = Y_t - (T_t + C_t) = S_t + a_t$$

$$Y_t' = T_t \times C_t \longrightarrow \frac{Y_t}{Y_t'} = \frac{Y_t}{T_t \times C_t} = S_t \times a_t$$

¿Dónde esta la componente cíclica?

Es posible predecir con las técnicas de suavizado:

- La predicción $\hat{Y}_{T+1}$ se calcula a partir de una media de las observaciones anteriores
- Existen otras técnicas de suavizado más complejas que la de medias móviles

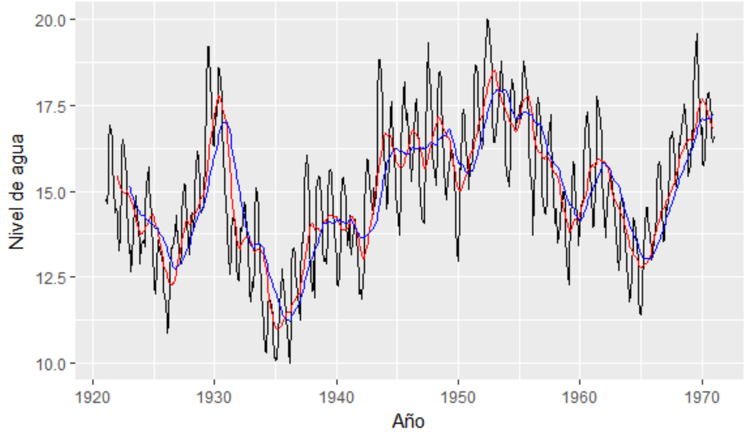
Suavizado exponencial

- El suavizado exponencial es una media móvil ponderada en la que los valores mas cercanos en el tiempo se ponderan más que los más alejados.

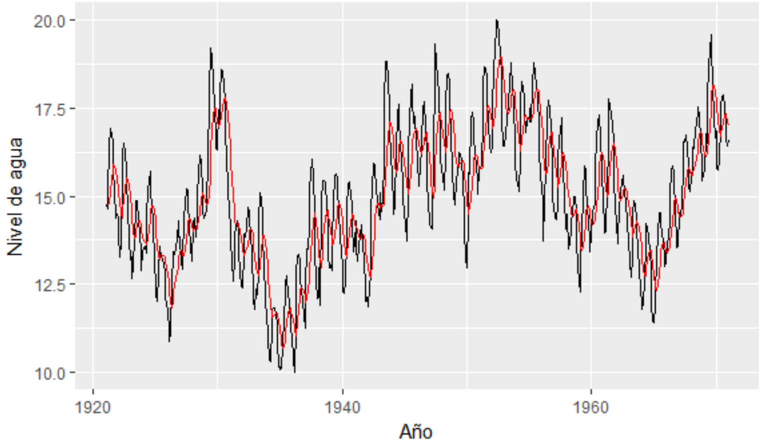
$$Y_t' = \alpha Y_t + (1-\alpha)Y_{t-1} + (1-\alpha)Y_{t-2} + \cdots \quad con \quad 0 < \alpha < 1$$

ENFOQUE DETERMINISTA

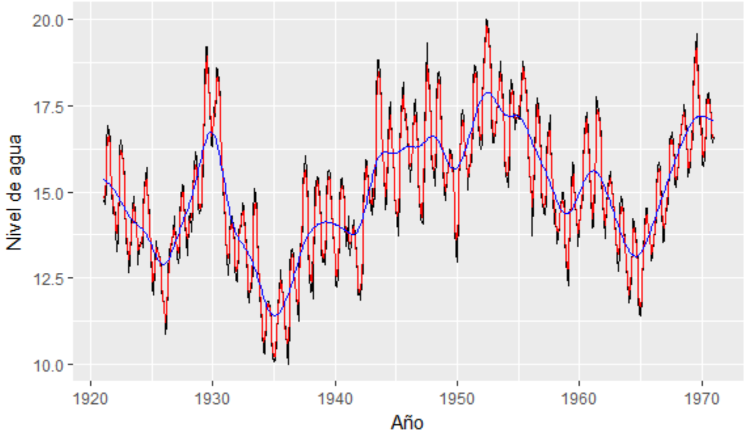
Medias móviles



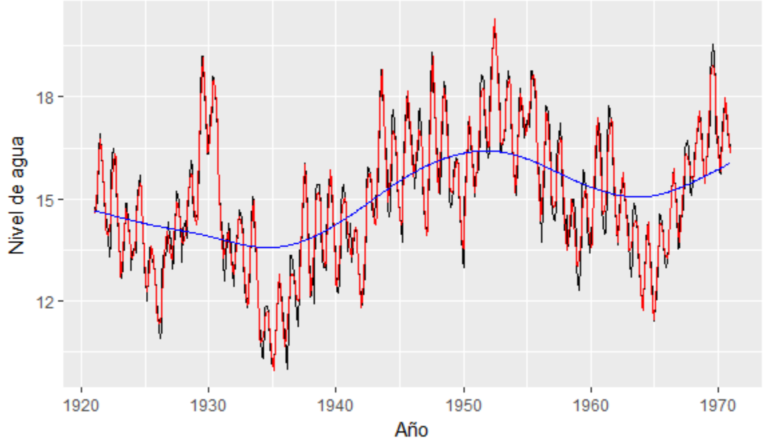
Exponencial



Kernel

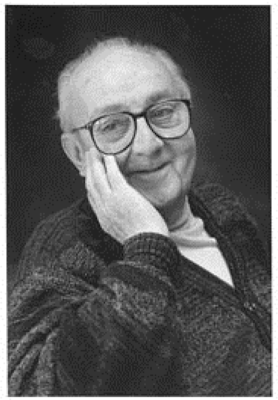


Splines



METODOLOGÍA BOX-JENKINS

- Esta metodología fue desarrollada por E. P. Box y G. Jenkins en 1976.
- Supone que la serie temporal son realizaciones de un proceso estocástico modelizable.
- Se sustenta en modelos ARIMA



E. P. Box

G. Jenkins

Conceptos básicos a definir:

- Función de autocorrelación
- Estacionariedad débil
- Modelos ARIMA

Función de autocorrelación:

- Nos ayuda a visualizar las relaciones entres observaciones

$$\rho_k = \frac{\gamma_k}{\gamma_0} \quad con \ \gamma_k = Cov(Y_t, Y_{t+k})$$

- ρ_k es la correlación existente entre dos observaciones separadas por un periodo de tiempo k.

Estacionariedad débil:

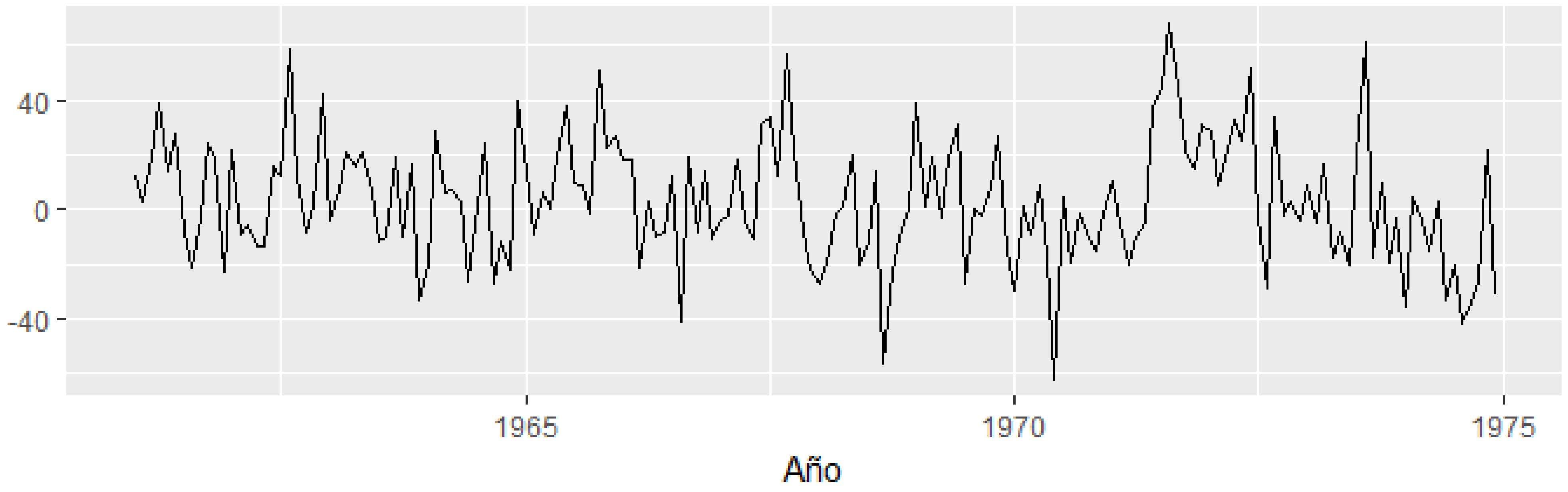
- Nuestra serie debe cumplir lo siguiente:

$$E[Y_t] = \mu < \infty \qquad \forall t$$

$$Var(Y_t) = \sigma^2 < \infty \qquad \forall t$$

$$Cov(Y_t, Y_{t+k}) < \infty \qquad \forall t$$

Ejemplo de serie estacionaria



Modelos ARIMA

- Procesos autorregresivos integrados de medias móviles

ARIMA

Procesos autorregresivos AR

Integrados

Procesos de medias móviles MA

Parte AR

- La variable se ve influenciada por sus valores pasados hasta el retardo $t - p$

- AR(p)

$$Y_t = \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \cdots + \emptyset_p Y_{t-p} + a_t$$

Parte MA

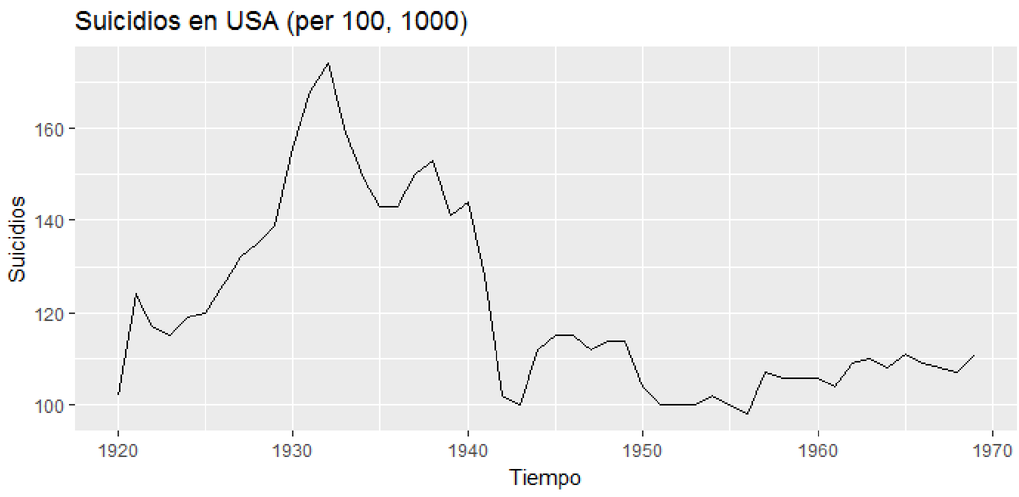
- La variable se ve influenciada por sus innovaciones contemporáneas y pasadas hasta el retardo $t-q$

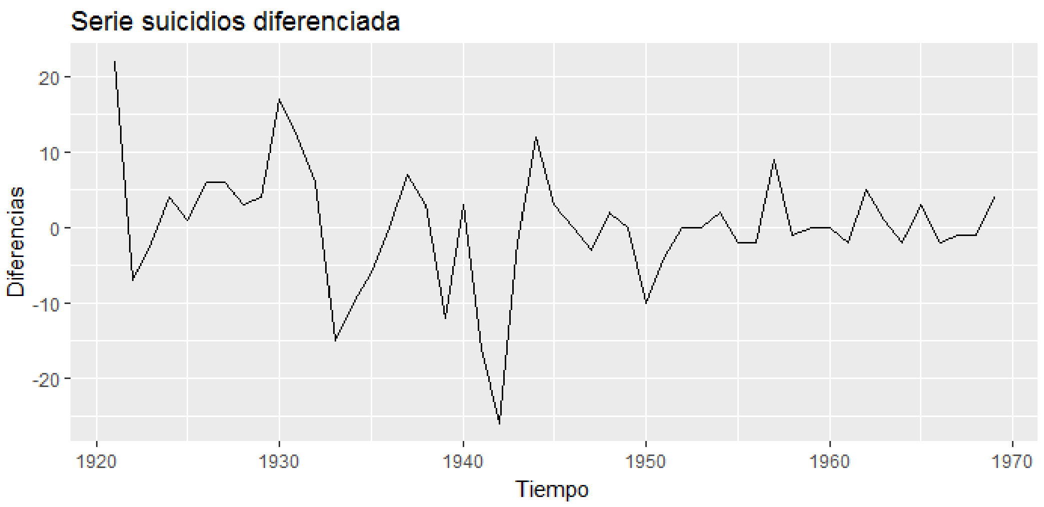
- MA(q)

$$Y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$$

Integración

- En ocasiones, es posible hacer una serie estacionaria diferenciándola





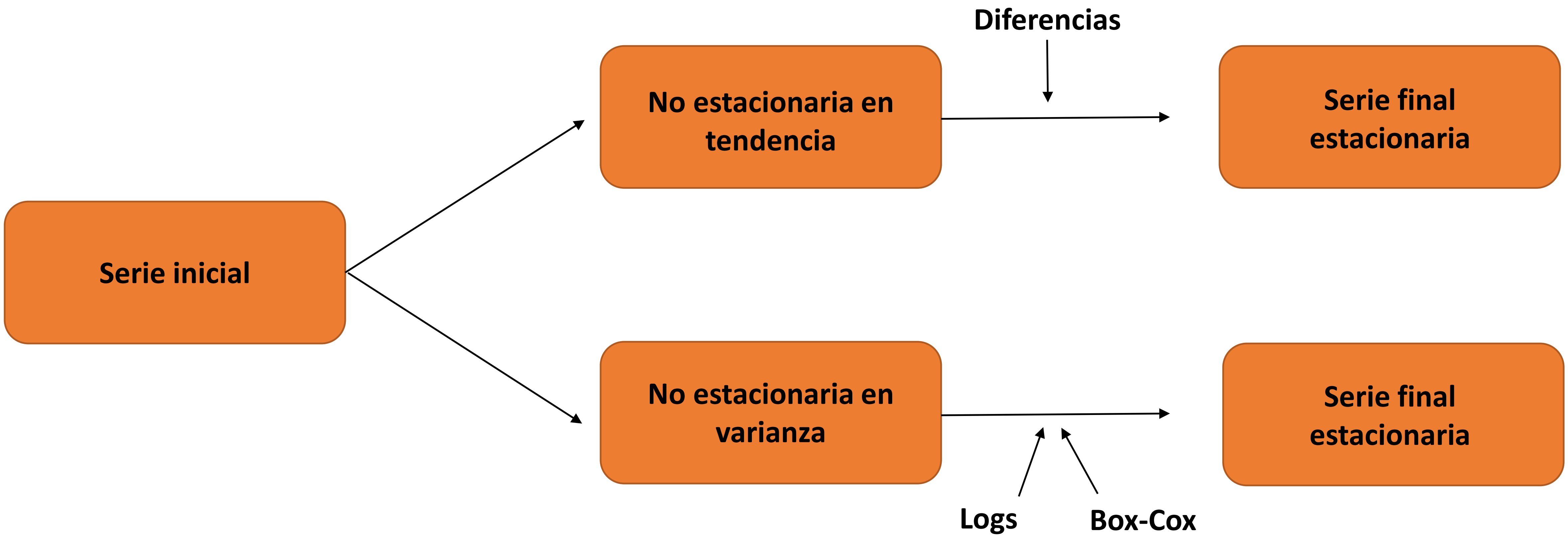
ARIMA$(\boldsymbol{p}, \boldsymbol{d}, \boldsymbol{q})$

- p $\rightarrow$ número de retardos de la parte AR
- q $\rightarrow$ número de retardos de la parte MA
- d $\rightarrow$ número de diferencias aplicadas a la serie

Etapas de la metodología Box-Jenkins:

- Preparación de los datos
- Identificación y selección del modelo
- Estimación de los parámetros
- Validación del modelo
- Predicción

Preparación de los datos:

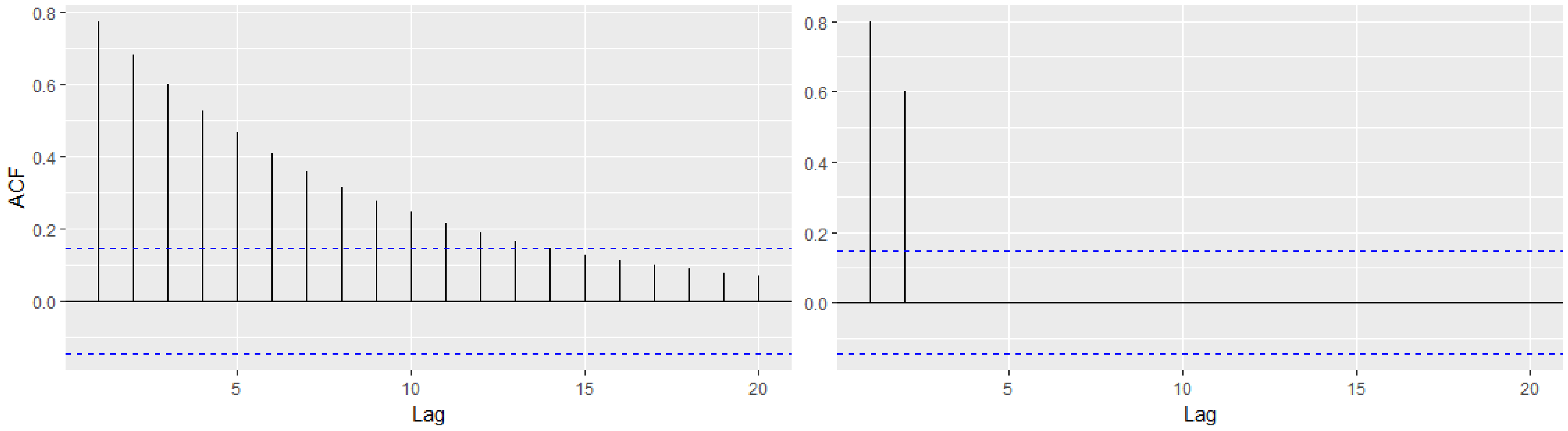


Identificación y selección del modelo:

	FAC	FACP
AR(p)	Decrece sin anularse	Se anula para $j>p$
MA(q)	Se anula para $j>q$	Decrece sin anularse
ARMA(p,q)	Decrece sin anularse	Decrece sin anularse

Identificación y selección del modelo:

$\mathrm{AR}(p = 2) \rightarrow$



Estimación de los parámetros:

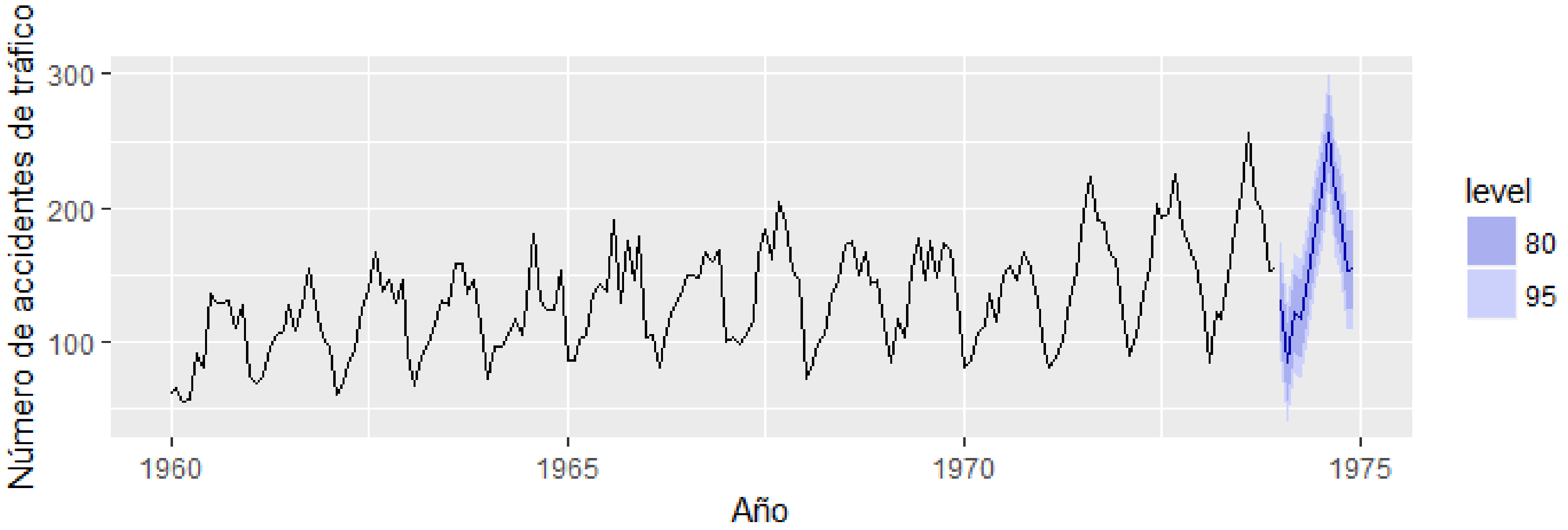
- Máxima verosimilitud
- Mínimos cuadrados no lineales
- Método de los momentos
- …

Validación del modelo:

- Los residuos son ruido
- Significatividad de los coeficientes
- Poder predictivo
- …

Predicciones:

- Una vez ajustado y validado el modelo toca predecir



Modelo lineal con tendencia y estacionalidad

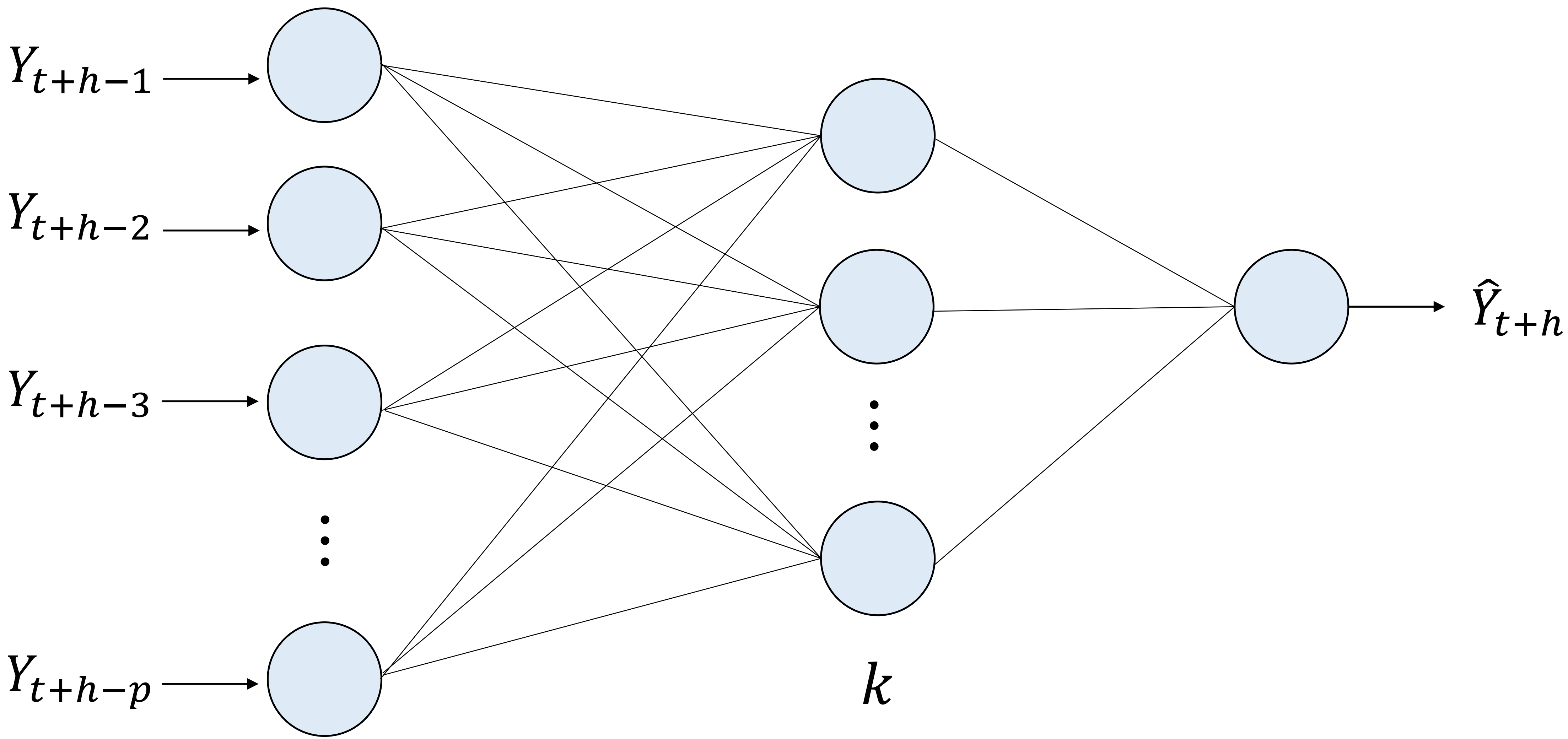
$$\boldsymbol{Y_t = \beta_0 + \beta_1 t + \sum_{i=2}^{M} \lambda_i d_i}$$

Triple suavizado exponencial (Holt-Winters)

$$\hat{Y}_{t+h} = l_t + hb_t + S_{t-M+h_M^+}$$
$$l_t = \alpha(Y_t - S_{t-M}) + (1-\alpha)(l_{t-1} + b_{t-1})$$
$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$$
$$S_t = \gamma(Y_t - l_{t-1} + b_{t-1}) + (1-\gamma)S_{t-M}$$

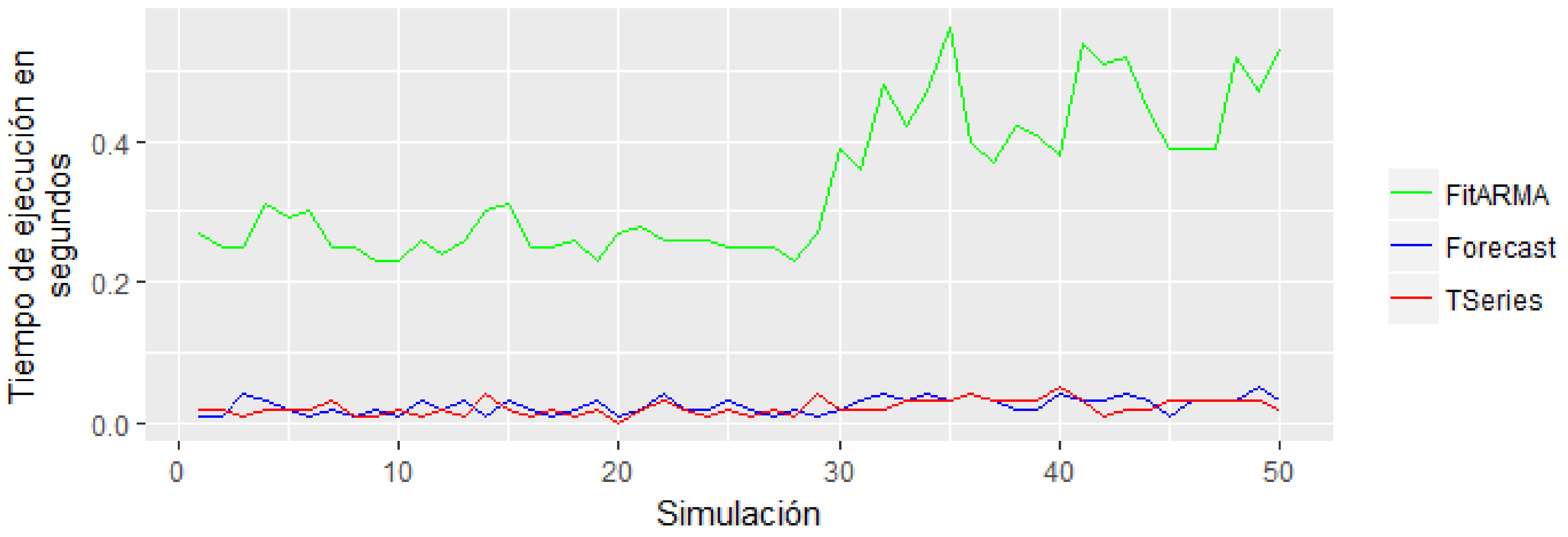
$$\hat{Y}_{t+h} = l_t + (\emptyset + \emptyset^2 + \cdots + \emptyset^h)b_t + S_{t-M+h_M^+}$$
$$l_t = \alpha(Y_t - S_{t-M}) + (1-\alpha)(l_{t-1} + \emptyset b_{t-1})$$
$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)\emptyset b_{t-1}$$
$$S_t = \gamma(Y_t - l_{t-1} + \emptyset b_{t-1}) + (1-\gamma)S_{t-M}$$

Red Neuronal (NNAR)



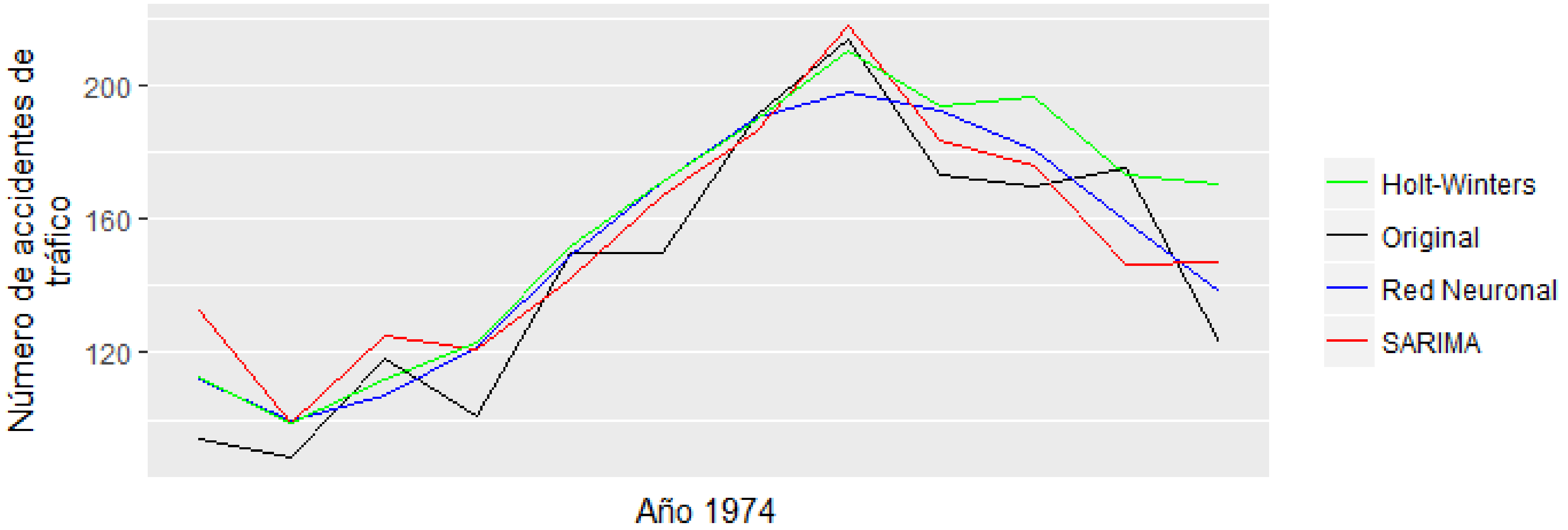
	`include.drift = FALSE`	`include.drift = TRUE`
AIC	1367.68	**1359.82**
AICc	1368.25	**1360.58**
BIC	1385.98	**1381.17**
MAE en el test	19.22651	**17.0999**
P valor (Ljung Box)	0.04321	**0.2748**

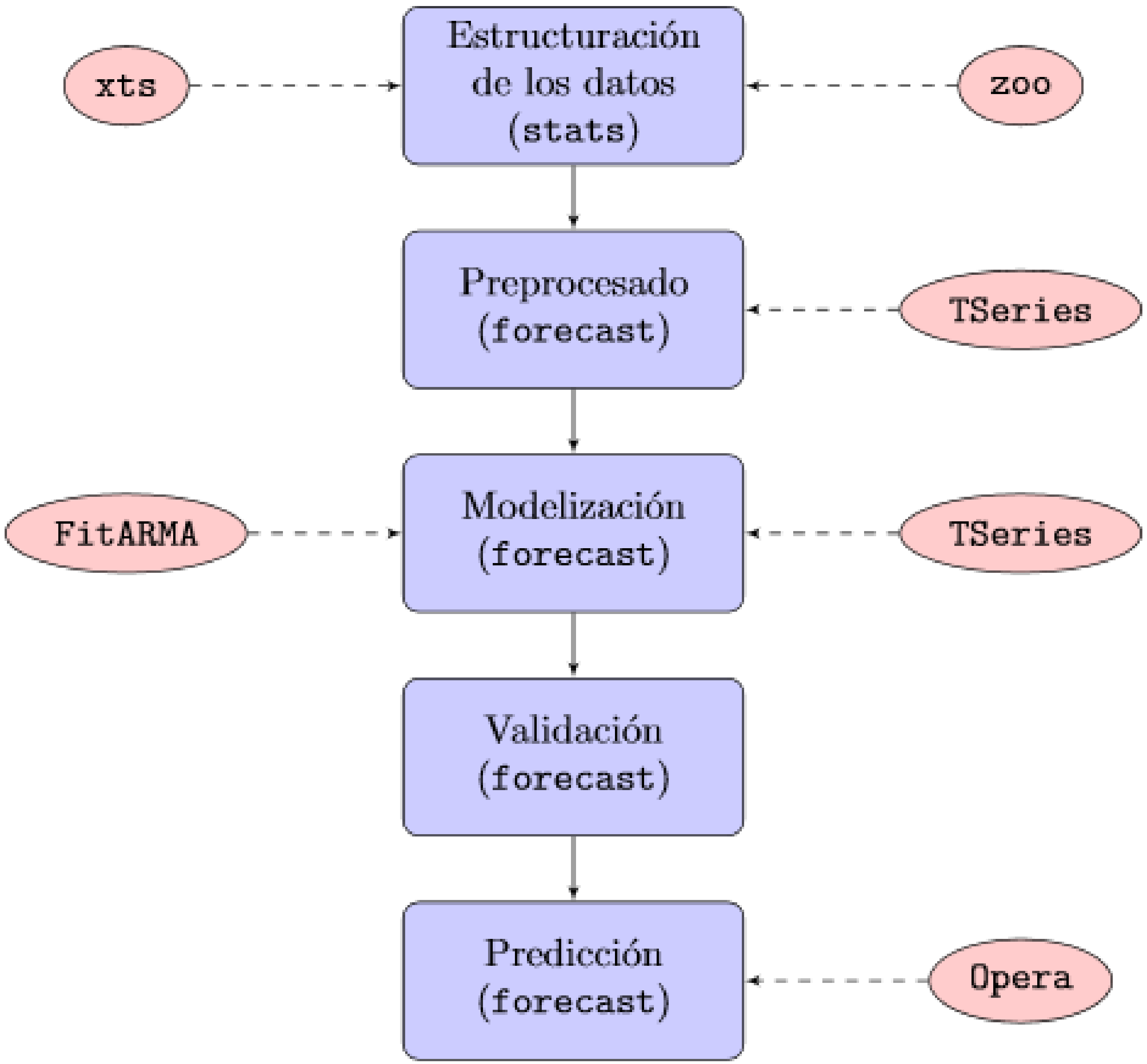
	AIC	AICc	BIC	MAE	Ljung Box
$(0,1,0)(0,0,2)_{12}$	1550.38	1550.53	1559.73	22.19949	~ 0
$(1,0,1)(2,1,1)_{12\, sin\, int}$	1367.68	1368.25	1385.98	19.22651	0.04321
$(1,0,1)(2,1,1)_{12\, con\, int}$	**1359.82**	**1360.58**	**1381.17**	17.09992	**0.2748**
$(1,0,0)(1,0,0)_{12}$	1514.22	1514.47	1526.72	**14.86938**	0.00085
$(2,0,0)(2,0,0)_{12}$	1499.24	1499.77	1517.99	**14.96979**	0.002708
$(1,1,4)(2,0,0)_{12}$	1485.29	1486.2	1510.24	23.44705	0.01353


Tiempo de ejecución en segundos
0.0
0.2
0.4
0
10
20
30
40
50
Simulación
FitARMA
Forecast
TSeries

k **modelos** $\longrightarrow$ $x_{k,\,t+h}$

$$\widehat{Y}_{t+h} = \sum_{i=2}^{M} p_{k,\,t+h}\, x_{k,\,t+h}$$


Número de accidentes de tráfico
200
160
120
Año 1974
Holt-Winters
Original
Red Neuronal
SARIMA


xts
Estructuración de los datos (stats)
zoo
Preprocesado (forecast)
TSeries
FitARMA
Modelización (forecast)
TSeries
Validación (forecast)
Predicción (forecast)
Opera