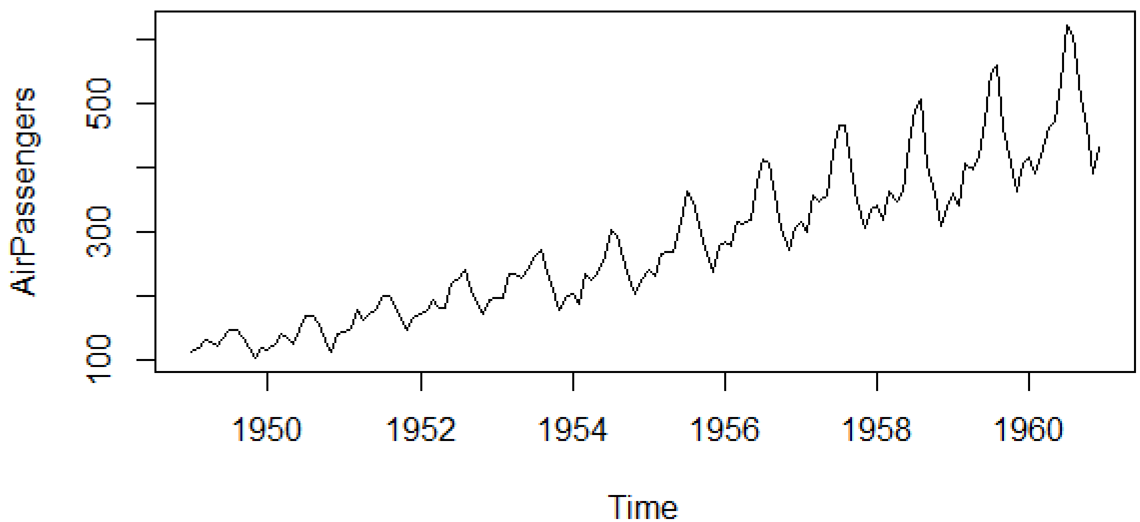
Tendência e Sazonalidade

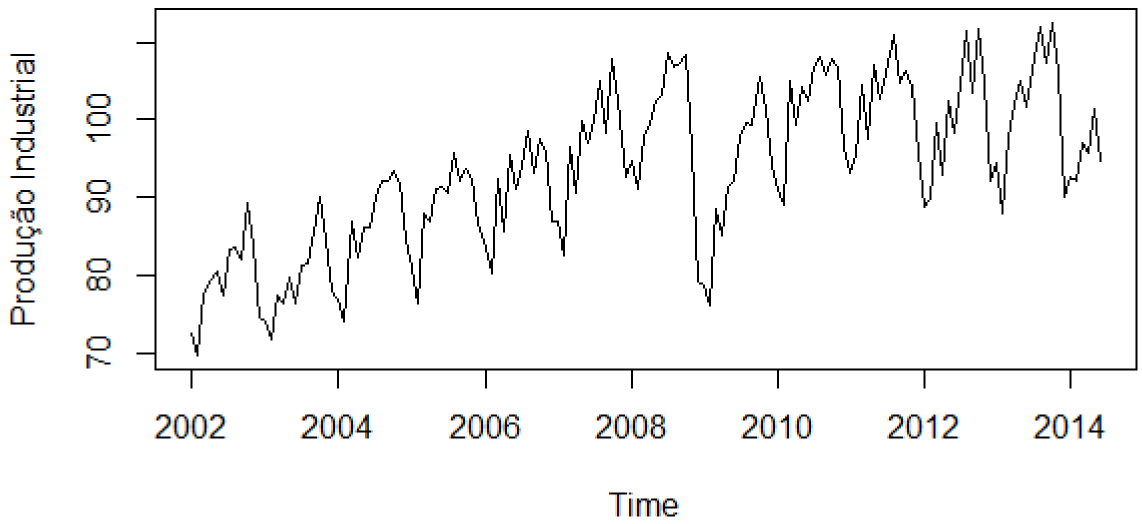
Lane Alencar
www.ime.usp.br/~lane

Número de passageiros de voos internacionais



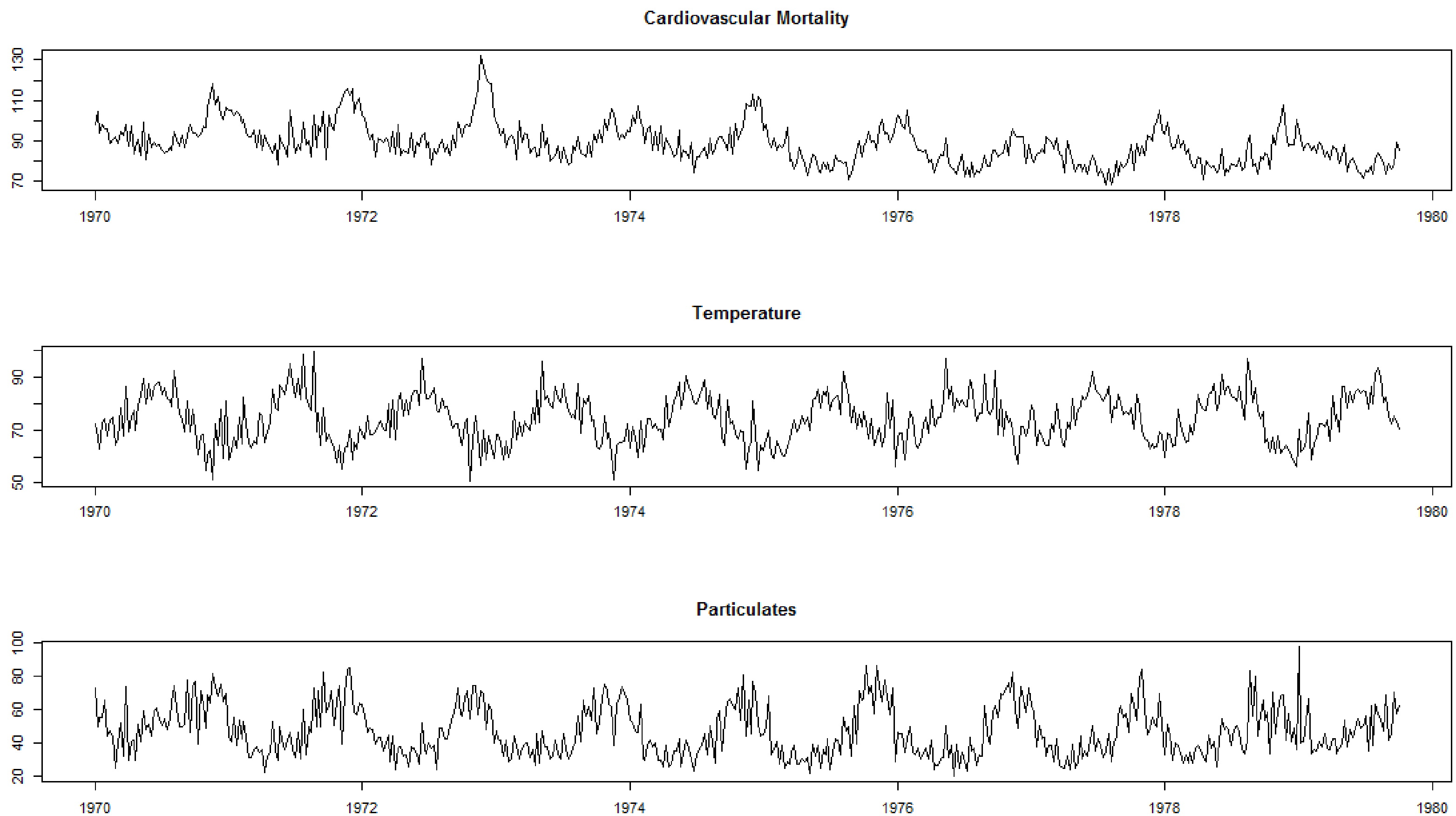
?AirPassengers
ts.plot(AirPassengers)

Produção Industrial – IBGE



- Fonte: IPEA
- www.ipeadata.gov.br
- Media de 2002=100

Mortalidade, Temperatura, Poluição



Fonte: Shumway e Stoffer (2006)

Componentes

- Podemos decompor o processo $Z = \{Z(t), t \in T\}$ em:

- $Z_t = T_t + S_t + a_t$ com

- T_t : tendência
- S_t : componente sazonal
- a_t : erro aleatório com $E(a_t)=0$ e $Var(a_t) = \sigma^2$.

Métodos descritivos e de suavização

- Modelo Polinomial:
- Médias Móveis: Suavizar os valores da série para estimar a tendência em cada instante;
- Lowess: Suavizar a série com ajustes sucessivos de retas de mínimos quadrados (lowess);

Polinômio

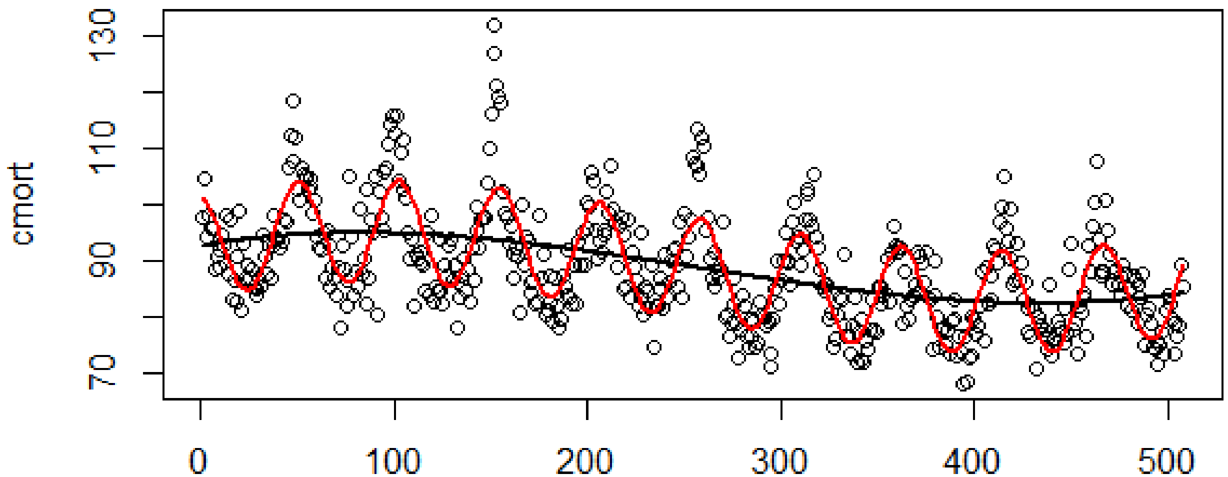
- Supomos $T_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_p t^p$. Ajustando o polinômio pelo método de mínimos quadrados temos que minimizar

$$\sum_{t=1}^{T} \left(Z_t - \beta_0 - \beta_1 t - \beta_2 t^2 - \ldots - \beta_p t^p \right)^2$$

Polinômio

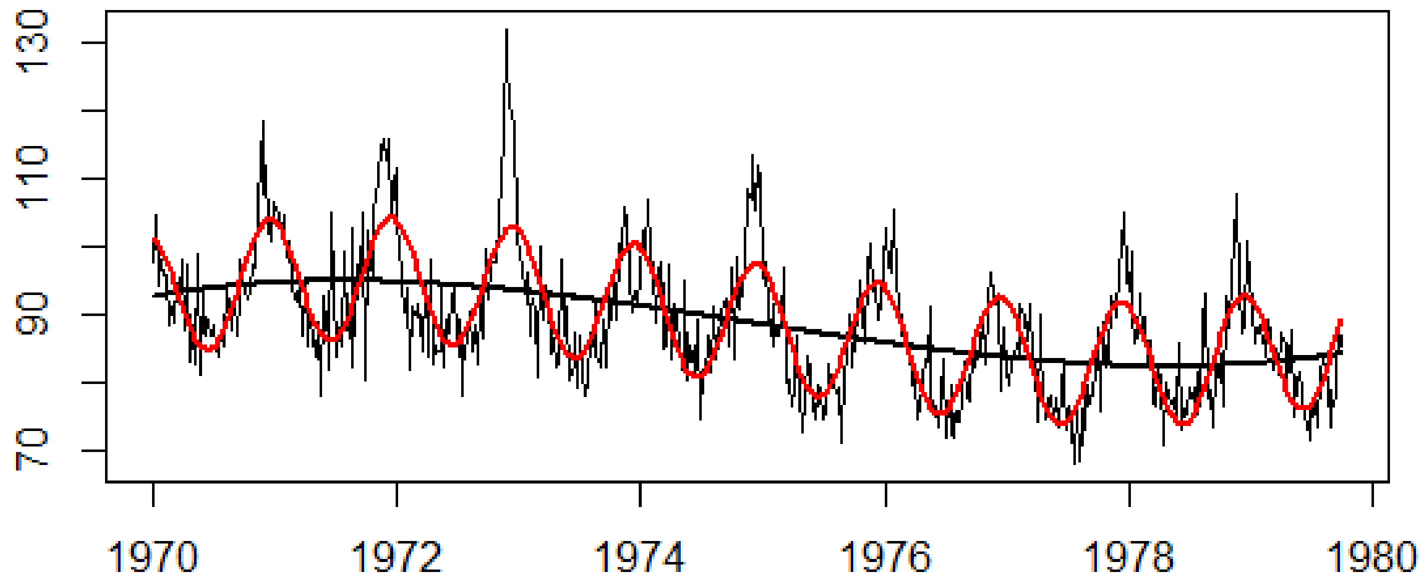
- # Mortalidade cardiovascular semanal de 1970 a 1979
- # Dados e programas em Shumway e Stoffer (2006)
- library(astsa)
- cmort=cmort$V1
- t = 1:length(cmort)
- t2 = t^2
- t3 = t^3
- c = cos(2*pi*t/52)
- s = sin(2*pi*t/52)
- fit1 = lm(mort ~ t + t2 + t3)
- fit2 = lm(mort ~ t + t2 + t3 + c + s)
- plot(t, mort)
- lines(fit1$fit)
- lines(fit2$fit)

Mortalidade Cardiovascular



- Average weekly cardiovascular mortality in Los Angeles County; 508 six-day smoothed averages obtained by filtering daily values over the 10 year period 1970-1979.

Mortalidade Cardiovascular



Médias Móveis

- Médias móveis em torno de Z_t

$$Z_t^* = \frac{1}{2n+1} \sum_{j=-n}^{n} Z_{t+j}.$$

- De modo mais geral:

$$Z_t^* = \sum_{j=-n}^{n} c_j Z_{t+j}.$$

Médias Móveis

- Médias móveis em torno de Zt

$$
\begin{aligned}
Z_t^* &= \sum_{j=-n}^{n} c_j (T_{t+j} + a_{t+j}) \\
&= \sum_{j=-n}^{n} c_j T_{t+j} + \sum_{j=-n}^{n} c_j a_{t+j},
\end{aligned}
$$

- vamos denotar

$$a_t^* = \sum_{j=-n}^{n} c_j a_{t+j}.$$

Médias Móveis

Supondo

- $E(a_t)=0$
- Tendência suave:

$$E(Z_t^*) = \sum_{j=-n}^{n} c_j T_{t+j} \approx \sum_{j=-n}^{n} c_j T_t = T_t = E(Z_t).$$

- $Var(a_t^*) = \sigma^2 \sum_{j=-n}^{n} c_j^2$ e para $\sum c_j^2 < 1$, temos $Var(a_t^*) < Var(a_t)$ e consequentemente $Var(Z_t^*) < Var(Z_t)$.

Médias Móveis

- Introduzimos correlação nos erros.
- Supondo Cov(a_t , a_s) = 0, s ≠ t , obtemos:

  Cov$(a_t^*, a_s^*) = \sigma^2 \sum_{j=-n+h}^{n} c_j c_{j-h}$, $h = 0,1,\dots,2n$

- e zero caso contrário.

Médias Móveis

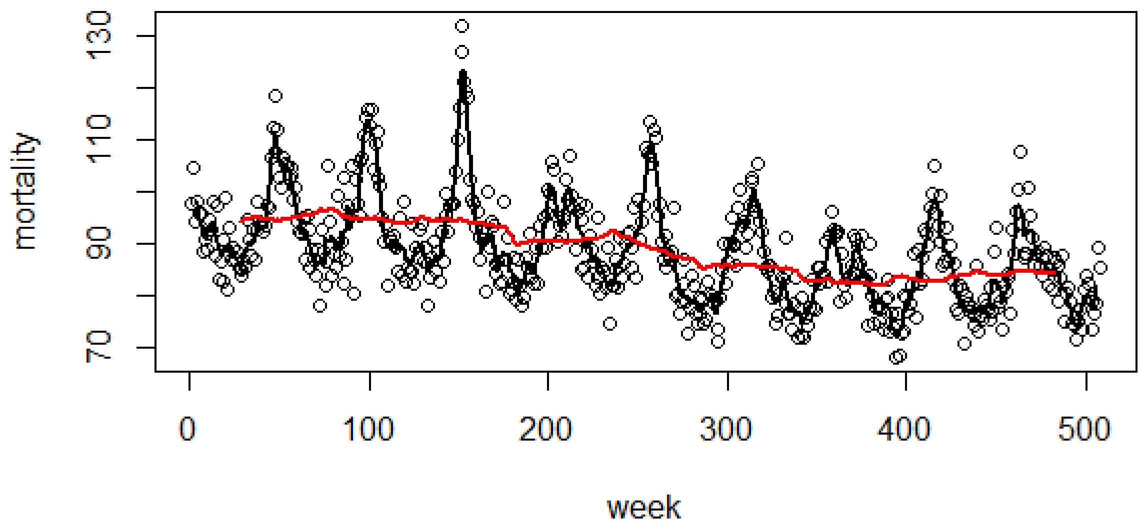
- Inferências são limitadas pois o método não é baseado em modelo estatístico.
- Não podemos obter previsões para $t=T+1,...$

- É bastante usado para análise descritiva.
- Podemos utilizar medianas móveis.

Médias Móveis

- mort <- read.table("http://www.stat.pitt.edu/stoffer/tsa2/data/cmort.dat", header=FALSE, sep=",", na.strings="NA", dec=".", strip.white=TRUE)
- mort=mort$V1
- t = 1:length(mort)
- ma5 = filter(mort, sides=2, rep(1,5)/5)
- ma53 = filter(mort, sides=2, rep(1,53)/53)
- plot(t, mort, xlab="week", ylab="mortality")
- lines(ma5)
- lines(ma53)

Médias móveis – Mortalidade



Kernel Smoothing

- Média ponderada

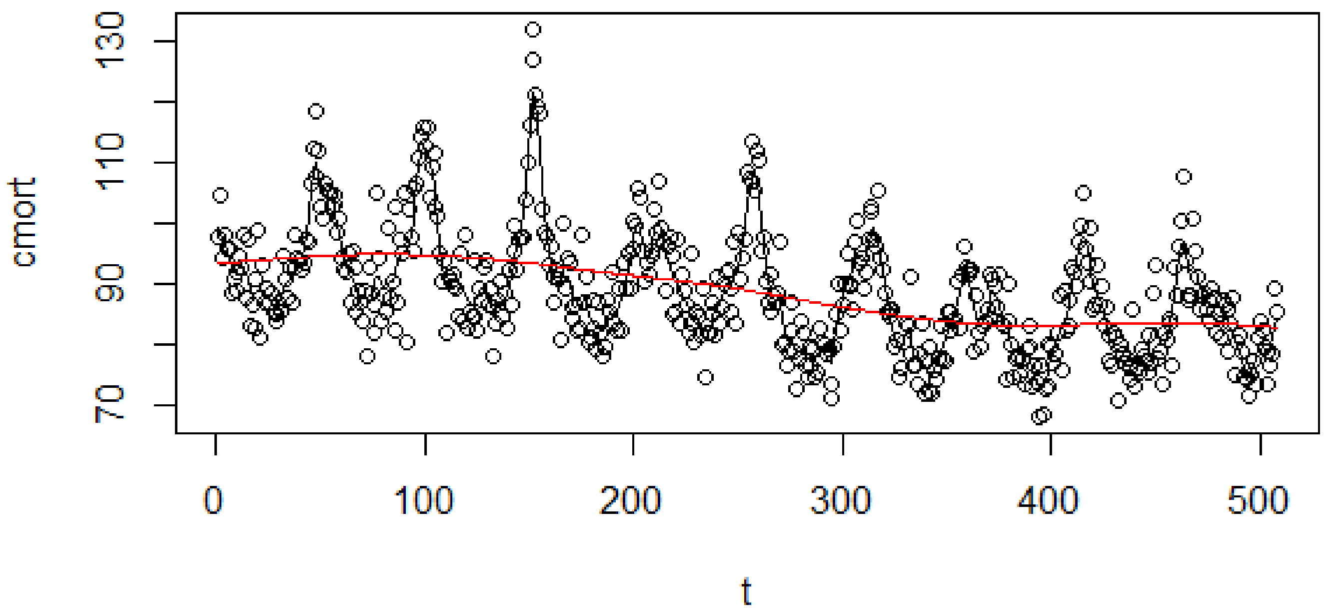
$$\hat{f}_t = \sum_{i=1}^{n} w_t(i) x_i$$

$$w_t(i) = \frac{k\left(\frac{t-i}{b}\right)}{\sum_{j=1}^{n} k\left(\frac{t-j}{b}\right)}$$

- O estimador $\hat{f}_t$ é chamado estimador Naradaya-Watson (Watson, 1966).
- Por exemplo, o kernel normal é $k(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$
- Quanto maior b, mais suave é $\hat{f}_t$.

Kernel Smoothing

- plot(t, mort)
- lines(ksmooth(t, mort, "normal", bandwidth=5))
- lines(ksmooth(t, mort, "normal", bandwidth=104))



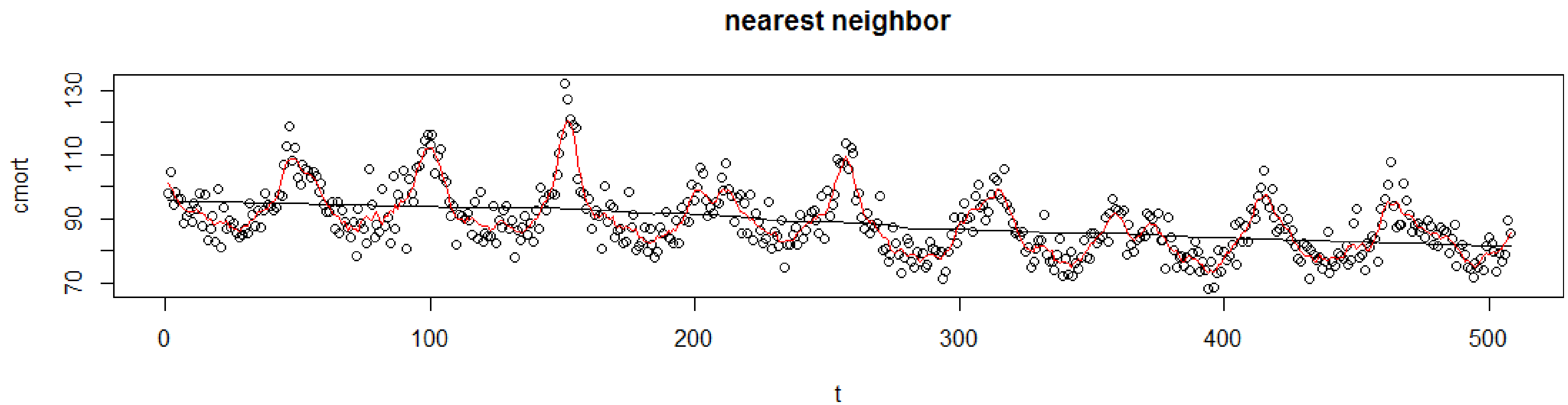
Regressão vizinho mais próximo

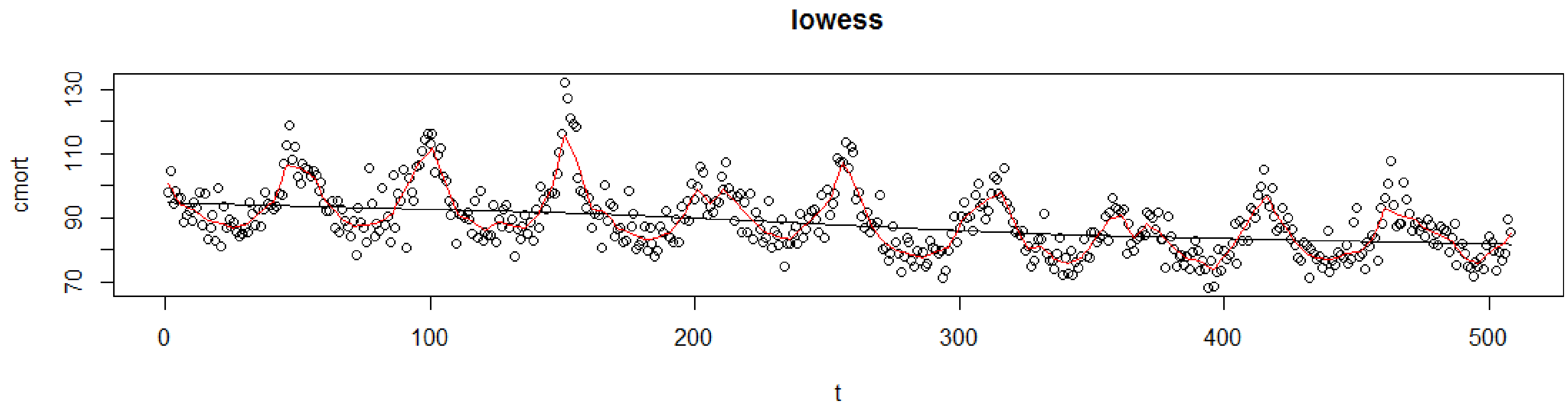
- Regressão com o método vizinho mais próximo é simplesmente uma regressão linear aplicada aos k vizinhos mais próximos $\{xt-k/2, \ldots, xt, \ldots, xt+k/2\}$ para predizer xt.
- Locally weighted regression scatter plot smoothing (Lowess) é um método mais complexo que usa uma certa proporção de vizinhos em uma regressão ponderada robusta (após regressão ponderada usa os resíduos para reponderar.
- Leiam Cleveland (1979)

Regressão vizinho mais próximo

- #mortalidade cardiovascular semanal
- n=508
- k=n/2 #tendência,
- k=n/100 #componente sazonal
- par(mfrow=c(2,1))
- plot(t, mort, main="nearest neighbor")
- lines(supsmu(t, mort, span=.5))
- lines(supsmu(t, mort, span=.01))
- plot(t, mort, main="lowess")
- lines(lowess(t, mort, .02))
- lines(lowess(t, mort, 2/3))

Vizinho mais próximo e Lowess





Splines

- Primeiro os tempos $t = 1, \ldots, n$, são separados em k intervalos $[t_0 = 1, t_1], [t_1 + 1, t_2], \ldots, [t_{k-1} + 1, t_k = n]$.
- Os tempos $t_1, \ldots, t_k$ são denominados nós. Em cada intervalo é ajustado o modelo de regressão
- $f_t = T_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$ (em geral de grau 3 = splines cúbicos).
- A suavização usando splines é obtida minimizando

$$\sum_{t=1}^{n} (z_t - f_t)^2 + \lambda \int (f_t'')^2 dt.$$

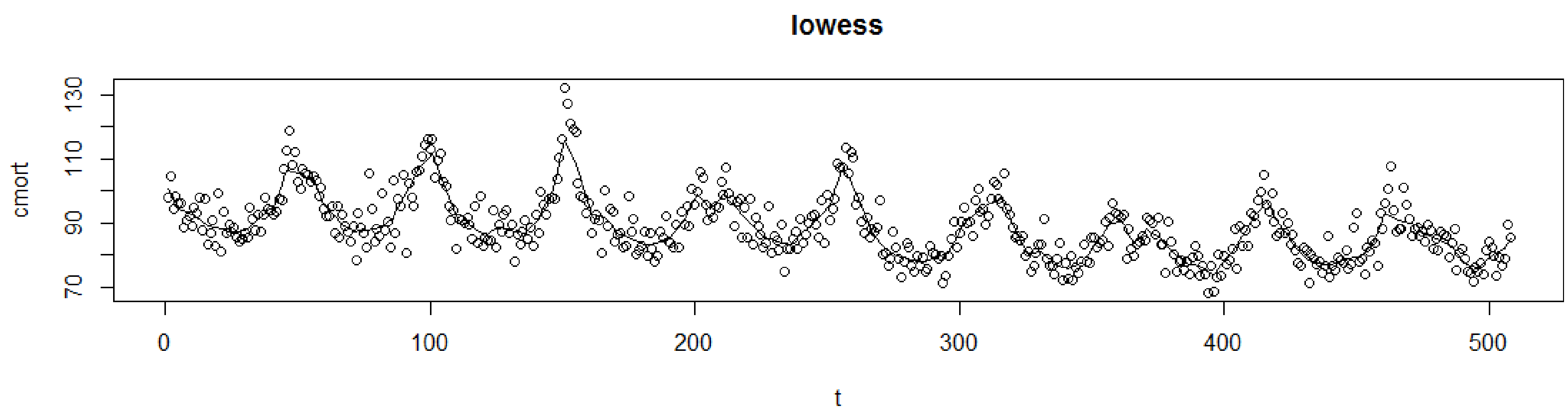
- A primeira parcela se refere ao ajuste do polınômio e a segunda ao grau de suavidade controlado por λ.

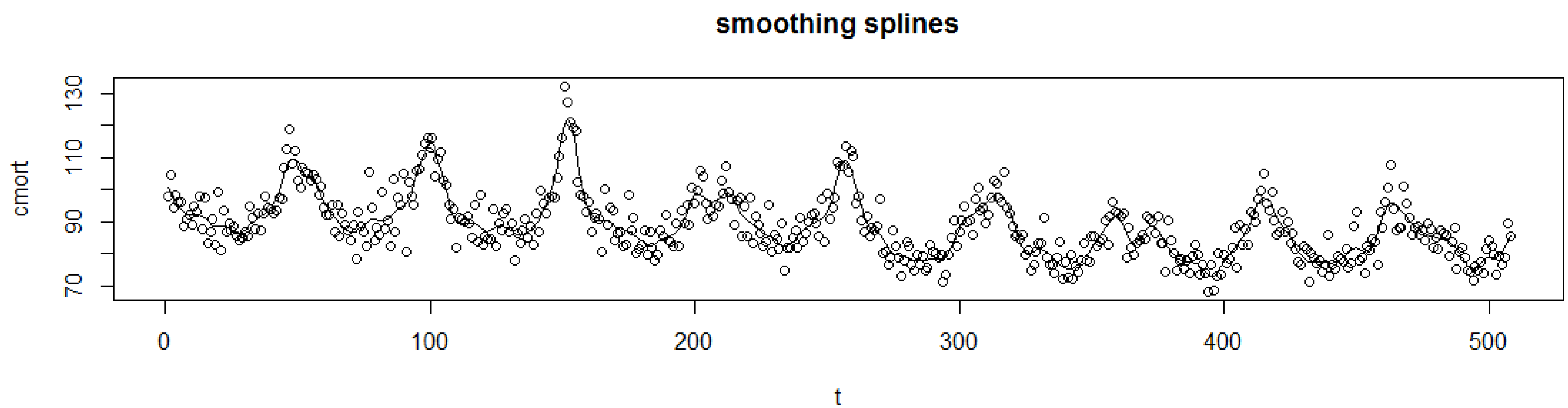
Splines

- plot(t, cmort)
- lines(smooth.spline(t, cmort, spar=1))
- lines(smooth.spline(t, cmort, spar=.1), col=2)

- par(mfrow=c(2,1))
- plot(t, cmort, main="lowess")
- lines(lowess(t,cmort))
- plot(t, cmort, main="smoothing splines")
- lines(smooth.spline(t,cmort))

Suavização


lowess
cmort
70
90
110
130
0
100
200
300
400
500
t


smoothing splines
cmort
70
90
110
130
0
100
200
300
400
500
t

Sazonalidade

- $Z_t = T_t + S_t + a_t$, $t = 1, \ldots, T$.
- Para o modelo aditivo acima, podemos:
  a. obter estimativas $\hat{S}_t$ de S_t;
  b. calcular série dessazonalizada $Z_t^{SA} = Z_t - \hat{S}_t$.

  Para modelo multiplicativo, podemos ter $Z_t = T_t S_t + a_t$ e $Y_t = \frac{Z_t}{\hat{S}_t}$.
- Sazonalidade determinística

  Por exemplo, para dados mensais:
  $Z_t = \beta_0 + \sum_{j=1}^{m} \beta_j t^j + \sum_{j=1}^{11} \alpha_j D_{jt} + a_t$,
  com D_{jt} variável indicadora que vale 1 se o mês t é igual a j.

Tendência e Sazonalidade

- Primeiro retira-se a tendência:

$$
\begin{aligned}
Y_t &= Z_t - \hat{T}_t \\
\hat{T}_t &= \sum_{j=-n}^{n} c_j Z_{t+j}.
\end{aligned}
$$

Para dados mensais (pode estender), temos que para a observação do ano i no mês j, Y_{ij}, temos:
$\bar{Y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$ é a média do mês j.
$\hat{S}_j = \bar{Y}_{.j} - \bar{Y}$ são estimativas das constantes sazonais para
$\bar{Y} = \frac{1}{12} \sum_{j=1}^{12} Y_{.j}$.
Calcula-se $Z_t^{SA} = Z_t - \hat{S}_t$.

Suavização Exponencial Simples

- Pode ser escrita como

$$\bar{Z}_t = \alpha Z_t + (1-\alpha)\bar{Z}_{t-1}, \bar{Z}_0 = Z_1, t = 1, \ldots, n.$$

- ou

$$\bar{Z}_t = \alpha \sum_{k=0}^{t-1} (1-\alpha)^k Z_{t-k} + (1-\alpha)^t \bar{Z}_0, \bar{Z}_0 = Z_1, t = 1, \ldots, n.$$

- onde $\bar{Z}_t$ é o valor suavizado e $0 \le \alpha \le 1$
- Os pesos são menores para observações mais antigas.

Suavização Exponencial Simples

- Previsão

$\widehat{Z}_t(h) = \bar{Z}_t, h = 1, 2, \ldots;$

A atualização da previsão é $\widehat{Z}_t(h) = \alpha Z_t + (1-\alpha)\widehat{Z}_{t-1}(h+1)$, sendo $\widehat{Z}_{t-1}(h+1)$ a previsão para o tempo t quando temos os dados até $t-1$.

$E(\widehat{Z}_t(h)) = \alpha \sum_{k=0}^{t-1}(1-\alpha)^k \mu_{t-k}$.

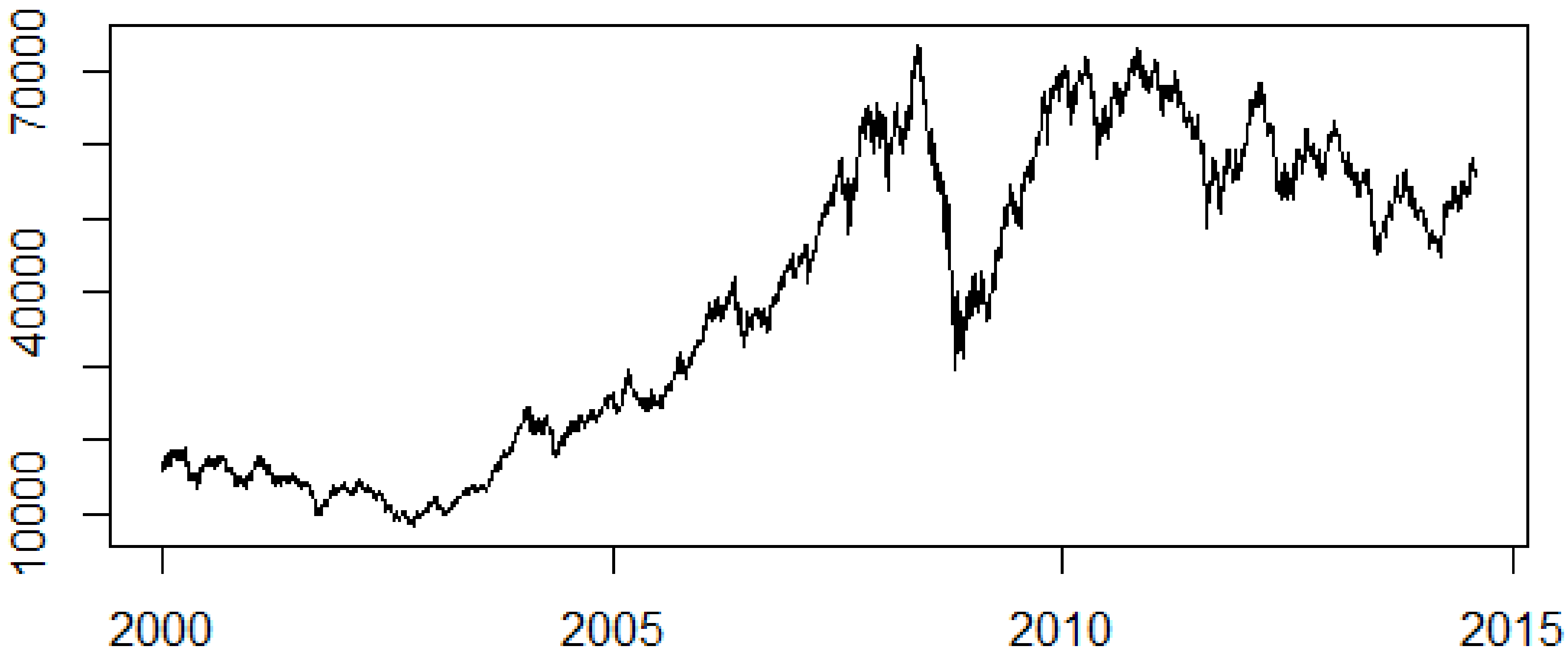
A expressão do $EQM(\widehat{Z}_t(h))$ encontra-se em MT(2006).

Assumindo estacionariedade $Z_t = \mu + a_t$, obtem-se expressões mais simples: $E(\widehat{Z}_t(h)) = \mu$ e $Var(\widehat{Z}_t(h)) = \frac{\alpha\sigma^2[1-(1-\alpha)^{2t}]}{2-\alpha} \rightarrow \frac{\alpha\sigma^2}{2-\alpha}$ e para $t \rightarrow \infty$.

Suavização Exponencial Simples

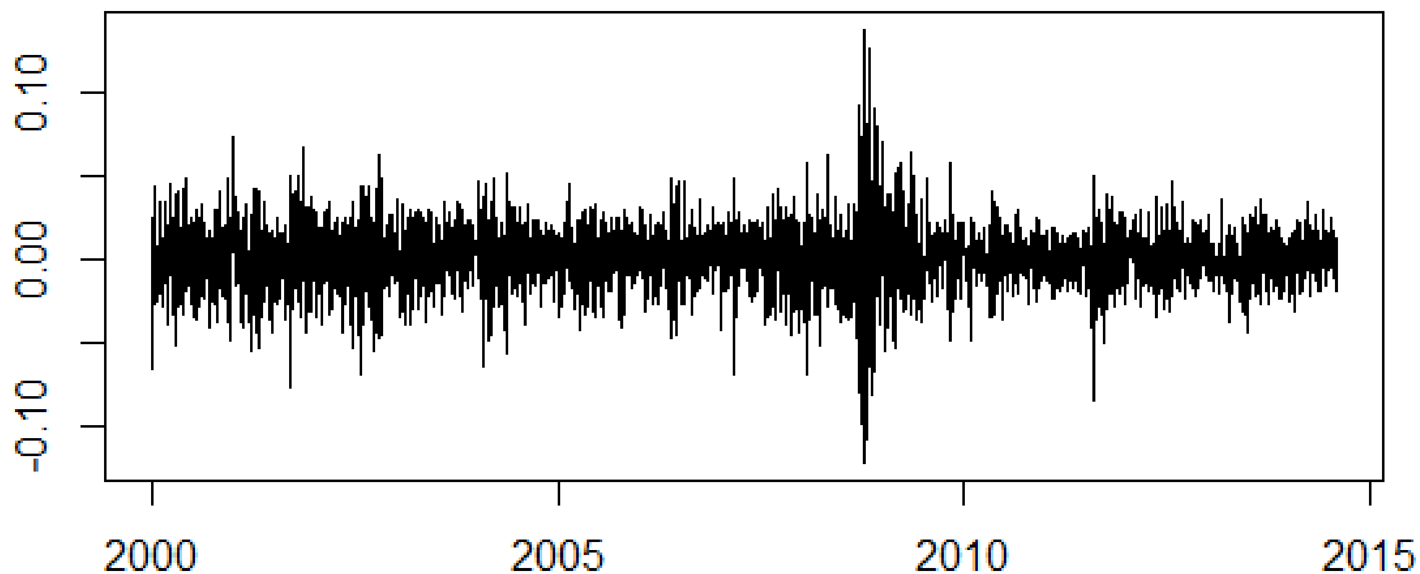
- Esse método é conhecido como EWMA (Exponential Weighting Moving Averages);
- O método é ótimo se Zt for ARIMA(0,1,1).
- Há vários métodos de escolha de α.
- Esse método é muito utilizado para estimar a volatilidade usando-se os log-retornos ao quadrado.

Ibovespa – Fechamento diário



Fonte: Yahoo Finance

Ibovespa – Log retornos

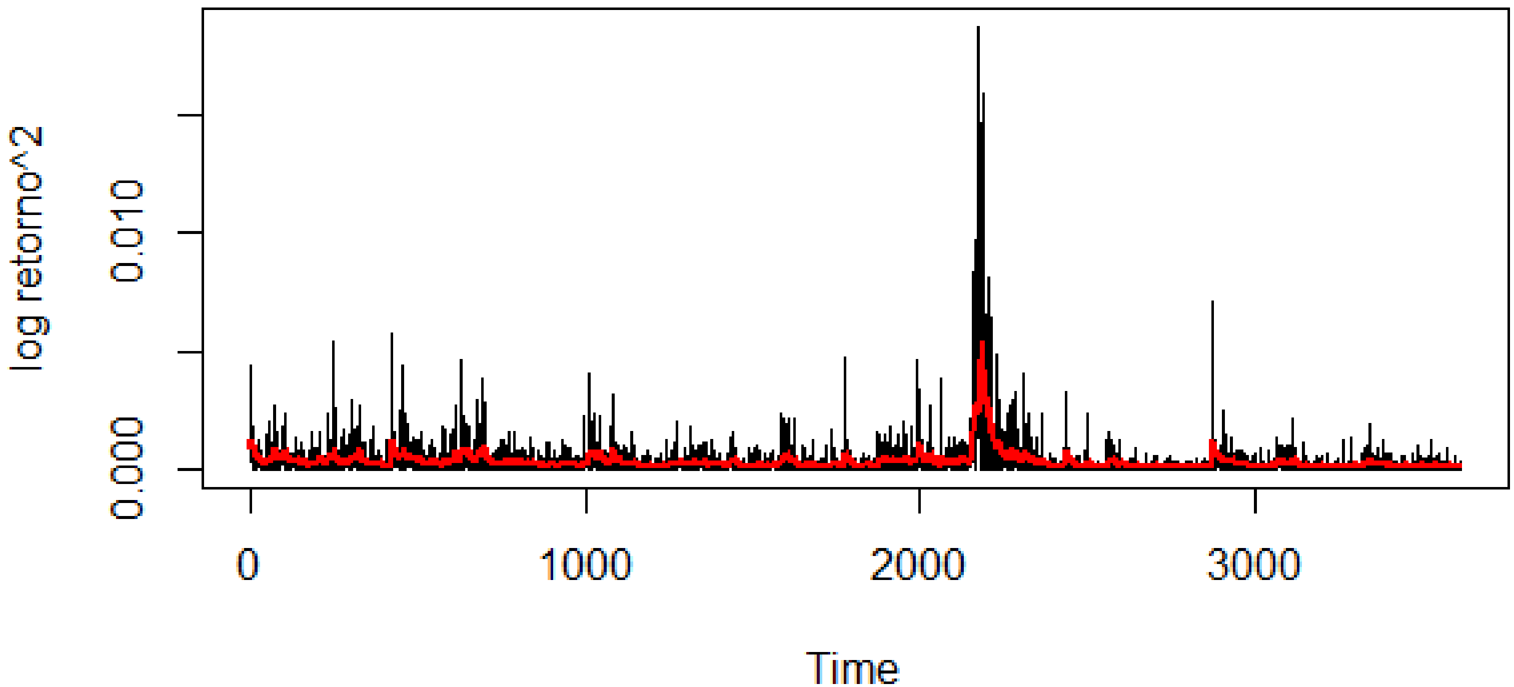


Log retorno= $\ln(P_t)-\ln(P_{t-1})$, sendo P_t o preço de fechamento

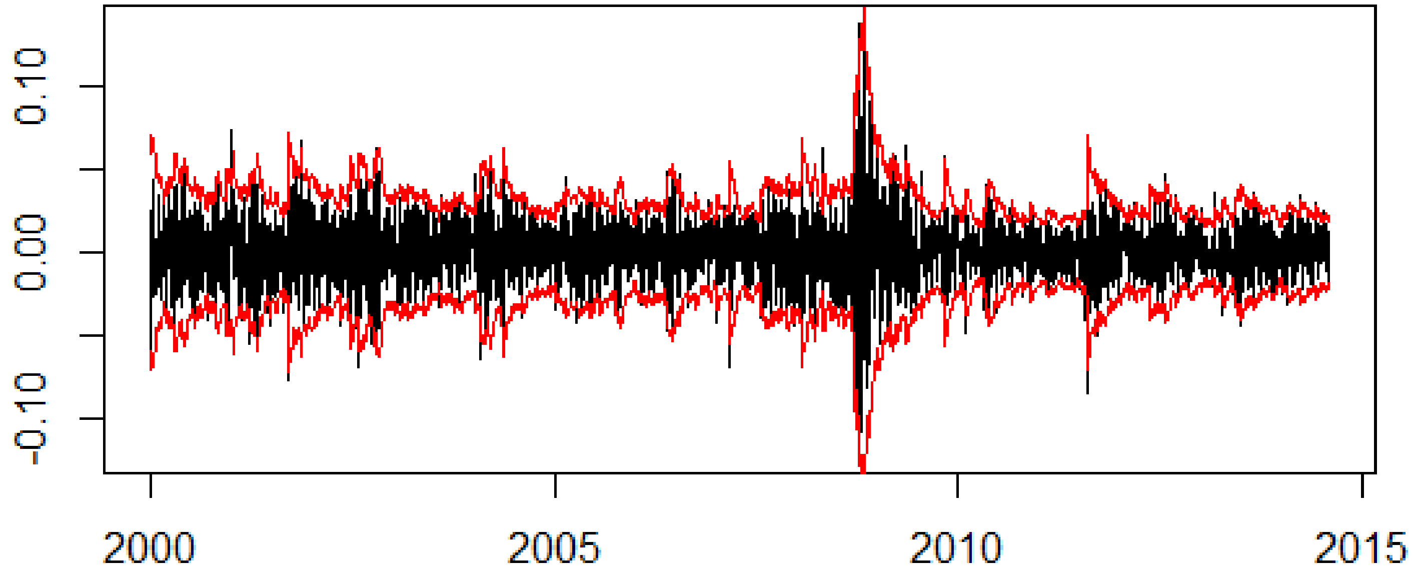
Suavização Exponencial Simples – Ibovespa

- library(forecast)
- ? ses
- d <- read.table("/dados/lnretibov.csv", header=TRUE, sep=";", na.strings="NA", dec=".", strip.white=TRUE)
- ts.plot(d$lnretibov)
- f=ses(d$lnretibov^2)
- ts.plot(d$lnretibov)
- lines(2*sqrt(f$fitted),col=2)
- lines(-2*sqrt(f$fitted),col=2)

Suavização exponencial simples do log retorno



Intervalo para log retorno



SES de $logret^2$ para estimar variâncias
IC=[0 ± raiz(variância)]

Suavização de Holt

- Para $Z_t = \mu + T_t + a_t$
- Os valores de nível e tendência são estimados por

$$\bar{Z}_t = AZ_t + (1-A)(\bar{Z}_{t-1} + \widehat{T}_{t-1}), 0 < A < 1;$$
$$\widehat{T}_t = C(\bar{Z}_t - \bar{Z}_{t-1}) + (1-C)\widehat{T}_{t-1}, 0 < C < 1.$$

- Observações

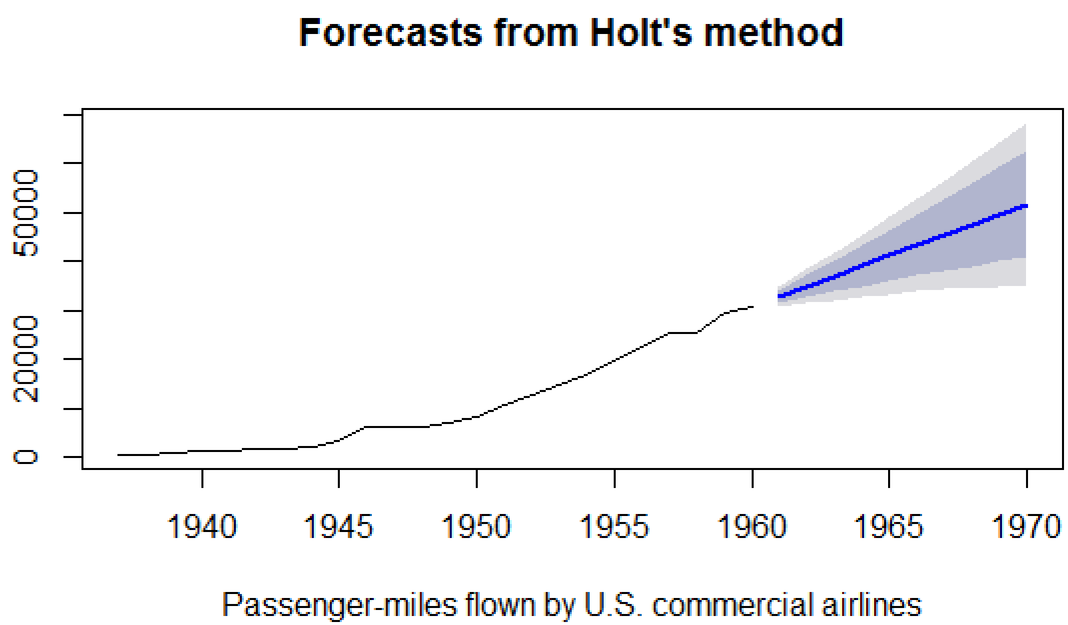
  Previsão: $\widehat{Z}_t(h) = \bar{Z}_t + h\widehat{T}_t, h > 0.$

  A e C são estimados por exemplo de modo a minimizar $\sum a_t^2$.

  Fornece previsões ótimas se Z_t for gerado a partir de processo ARIMA(0,2,2).

- fcast <- holt(airmiles)
- plot(fcast)

Suavização Holt



- fcast <- holt(airmiles)
- plot(fcast)

Holt–Winters

- A série com período s, fator sazonal multiplicativo (F_t) e tendência aditiva é
- $Zt = \mu_t F_t + T_t + a_t$, $t = 1, . . . , T$.
- Equações de suavização para $t = s+1, . . . , T$:

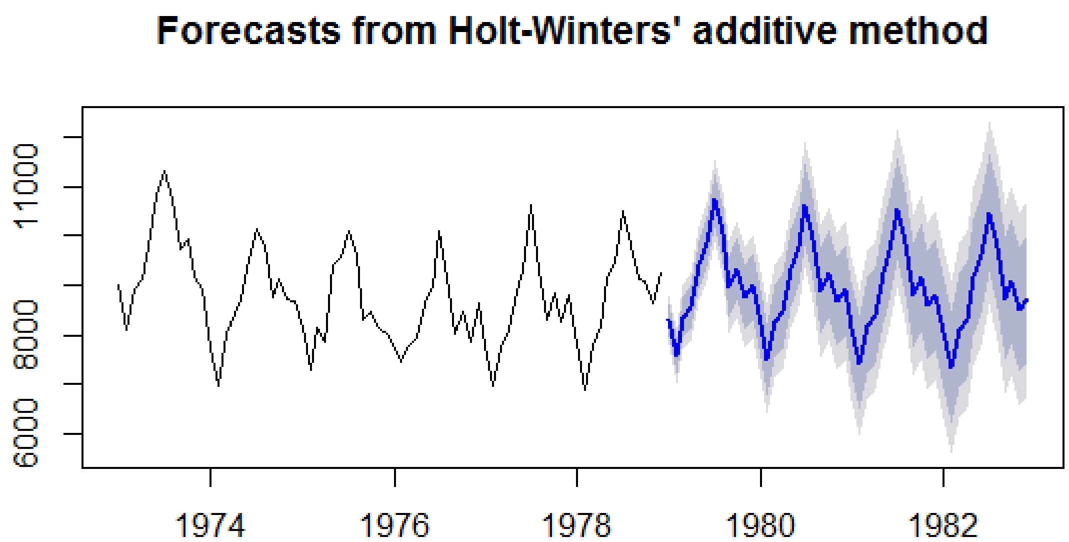
$$\widehat{F}_t = D\frac{Z_t}{\bar{Z}_t} + (1 - D)\widehat{F}_{t-s}, 0 < D < 1.$$

$$\bar{Z}_t = A\frac{Z_t}{\widehat{F}_{t-s}} + (1 - A)(\bar{Z}_{t-1} - \widehat{T}_{t-1}), 0 < A < 1;$$

$$\widehat{T}_t = C(\bar{Z}_t - \bar{Z}_{t-1}) + (1 - C)\widehat{T}_{t-1}, 0 < C < 1.$$

- A, C e D estimados por mínimos quadrados.

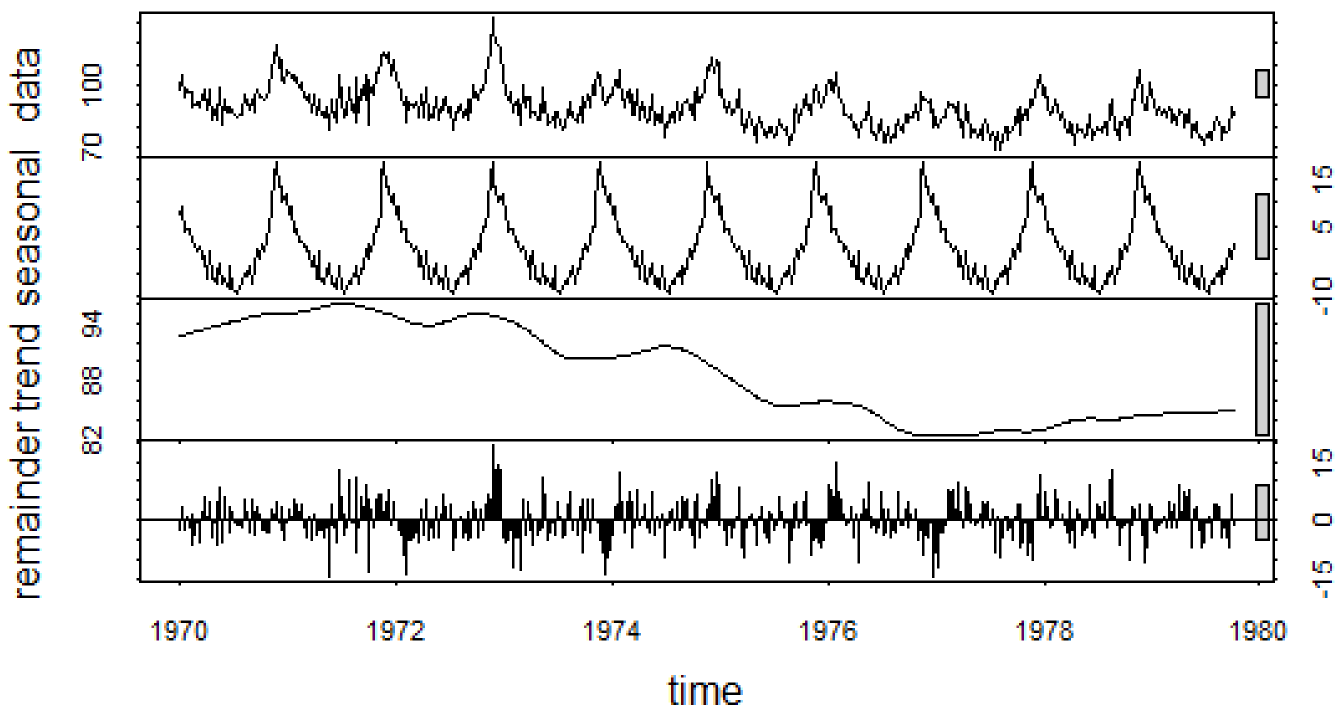
Holt Winters



- deaths.fcast <- hw(USAccDeaths,h=48)
- plot(deaths.fcast)

Decomposição Tendência Sazonalidade usando lowess

- Cleveland(1990) – STL



Mortalidade Cardiovascular – Shumway e Stoffer

- plot(stl(log(co2), s.window="per", t.window=1000))

- Ler artigo Cleveland et al. (1990)

Referências

- Shumway and Stoffer (2006)
- Morettin e Toloi (2006)
- Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. American Statistical Association* **74**, 829–836.
- Cleveland, W. S. (1981) LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician* **35**, 54
- R. B. Cleveland, W. S. Cleveland, J.E. McRae, and I. Terpenning (1990) STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, **6**, 3–73.
- www.r-project.org