

REGRESSÃO NÃO PARAMÉTRICA

TX 753- Métodos probabilísticos em Engenharia Ambiental

Cybelli Barbosa

Thiago Brandão

REGRESSÃO PARAMÉTRICA

- Relação funcional entre as variáveis explicativa e resposta é supostamente conhecida,
- Podem existir parâmetros com valores desconhecidos, mas estes podem ser estimados,
- Parâmetros livres geralmente possuem interpretação física,
- Objetivo principal é estimar o valor dos parâmetros,
- Os modelos que não são puramente paramétricos são denominados não-paramétricos ou semi-paramétricos.

REGRESSÃO PARAMÉTRICA

Exemplo geral: $y_i = f(\beta, x_i) + \varepsilon_i$

- onde β é o vetor de parâmetros a ser estimado,
- x é o vetor de predição,
- o erro, ε , é assumido normal e independentemente distribuído, com média zero e variância desconhecida.

REGRESSÃO NÃO PARAMÉTRICA

- Não há conhecimento *a priori* a respeito da forma da função, e esta pode adquirir um conjunto amplo de formas,
- Objetivo principal não é estimar o valor dos parâmetros, mas reduzir as possibilidades para a forma da função,
- A maioria dos métodos de regressão não paramétrica assume que f é uma função suave e contínua,
- Estimadores podem não possuir interpretação física.

REGRESSÃO NÃO PARAMÉTRICA

Exemplo geral: $y_i = f(x_i) + \varepsilon_i$

- A função f é menos especificada, pela falta de dados *a priori*,
- A aproximação não paramétrica é mais flexível pois determina f a partir de uma família de funções.
- Um caso especial importante do modelo geral é regressão não paramétrica simples, onde existe apenas um preditor: x_i ,
- Dificuldade: ajustar e exibir o modelo de regressão não paramétrica geral quando existem muitos preditores.

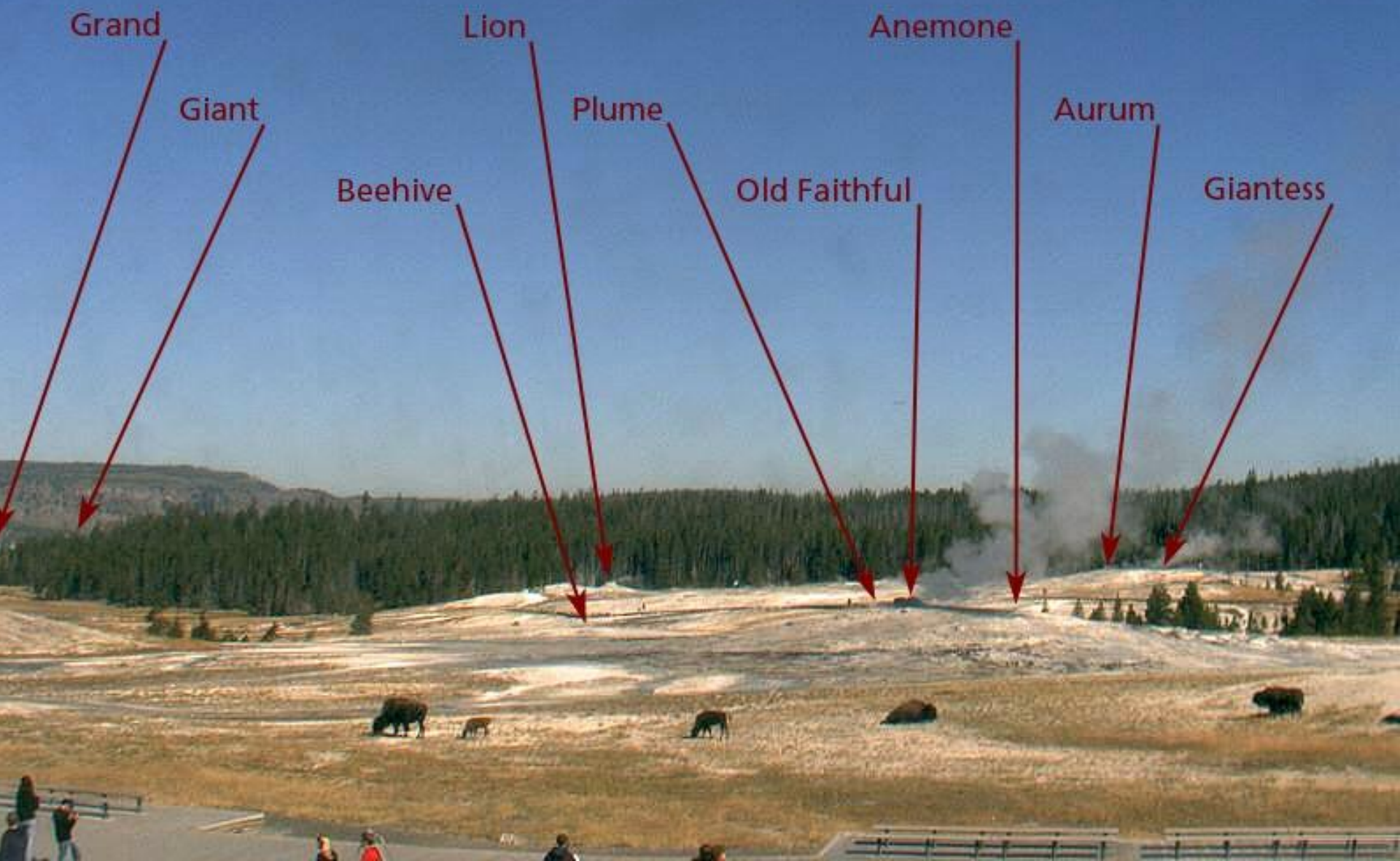
EXEMPLO: FARAWAY

Utilização de três conjuntos de dados:

- Exemplo A (simulado): $f(x)=\sin^3(2\pi x^3)$
- Exemplo B (simulado): $f(x)=0$
- Dados reais Old Faithful:
tempo de espera entre as erupções e a duração da erupção do gêiser Old Faithful em Yellowstone National Park, Wyoming, EUA.



Prominent Geysers in the Vicinity of Old Faithful



Grand

Giant

Beehive

Lion

Plume

Old Faithful

Anemone

Aurum

Giantess

Método do Núcleo (*Kernel*)


```
>require (faraway)
```

```
# dados do exemplo A (simulado):  $f(x)=\sin^3(2*\pi*x^3)$ 
```

```
>data(exa)
```

```
>plot (y ~ x, exa, main="Exemplo A", pch=".")
```

```
>lines (m ~ x, exa)
```

```
# dados do exemplo B (simulado):  $f(x)=0$ 
```

```
>data(exb)
```

```
>plot (y ~ x, exb, main="Exemplo B", pch=".")
```

```
>lines (m ~ x, exb)
```

```
# Dados reais Old Faithful:
```

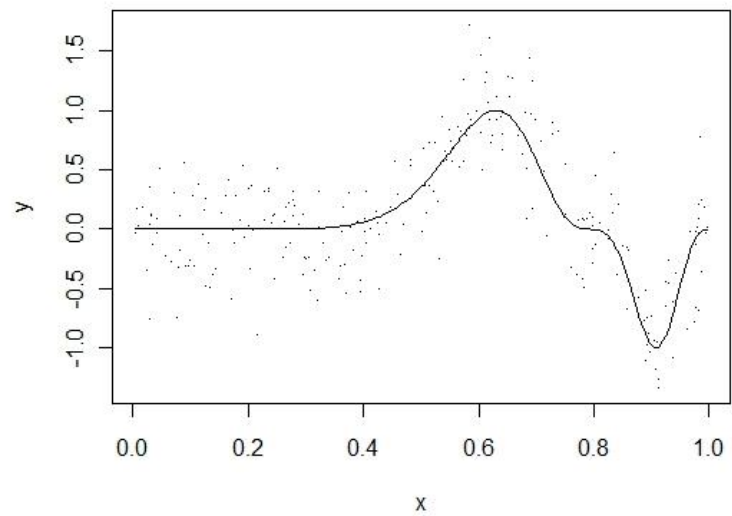
```
# eruptions: duração da erupção em minutos.
```

```
# waiting: tempo de espera até a próxima erupção, em minutos.
```

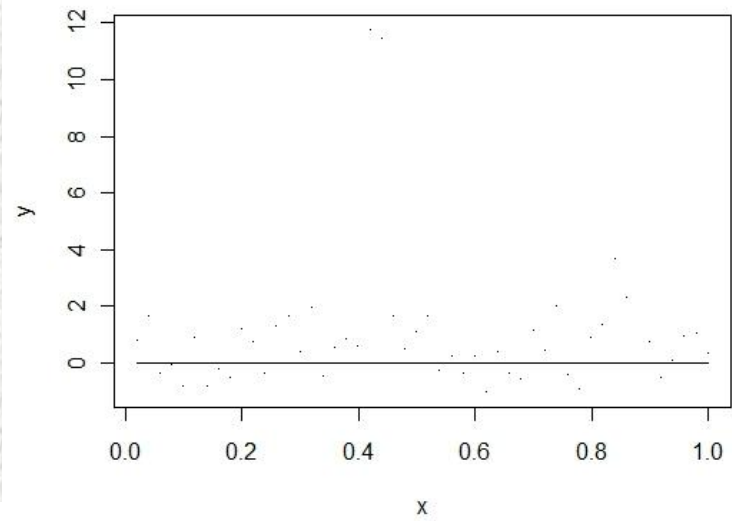
```
>data(faithful)
```

```
>plot (waiting ~ eruptions, faithful, main="Old Faithful", pch=".")
```

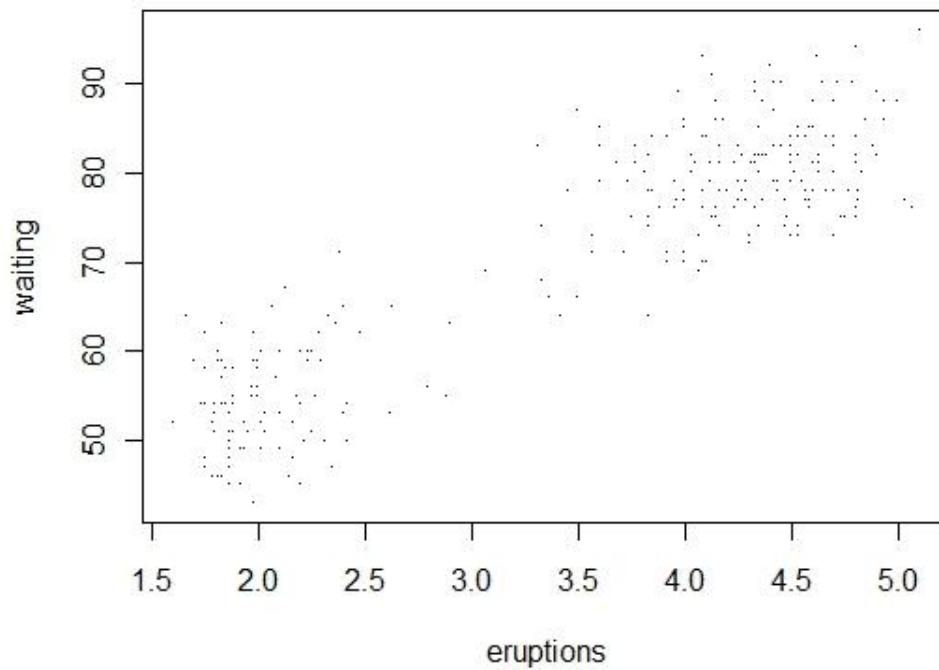
Exemplo A



Exemplo B



Old Faithful



ESTIMADORES *KERNEL* $\int K=1$

(janela móvel)

- Geral:

$$\hat{f}_\lambda(x) = \frac{1}{n\lambda} \sum_{j=1}^n K\left(\frac{x-x_j}{\lambda}\right) Y_j = \frac{1}{n} \sum_{j=1}^n w_j Y_j \quad w_j = K\left(\frac{x-x_j}{\lambda}\right) / \lambda$$

λ : largura da banda ou parâmetro de alisamento, controla a suavidade da curva ajustada.

- A melhor escolha de λ fornece:

MSE: *mean squared error*

$$\text{MSE}(x) = E(f(x) - \hat{f}_\lambda(x))^2 = O(n^{-4/5})$$

- Nadaraya-Watson:

utilizado quando o espaçamento da variável explicativa é bastante desigual (intervalos heterogêneos em x).

$$f_\lambda(x) = \frac{\sum_{j=1}^n w_j Y_j}{\sum_{j=1}^n w_j}$$

- Epanechnikov Kernel:

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2) & |x| < 1 \\ 0 & \text{otherwise} \end{cases}$$

Exemplo do estimador *Kernel* Nadaraya-Watson com 3 diferentes larguras de banda λ :

Estimador *Kernel*:

lambda = 0.1:

```
plot(waiting ~ eruptions, faithful, main="bandwidth=0.1", pch=".")  
lines(ksmooth(faithful$eruptions, faithful$waiting, "normal", 0.1))
```

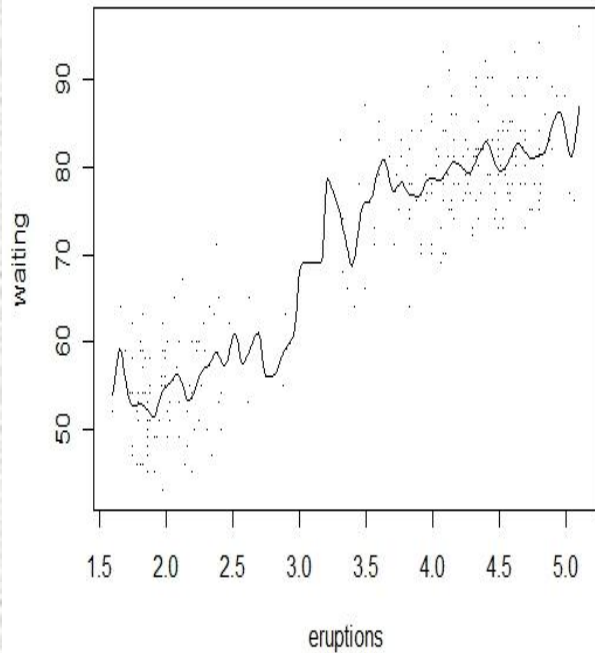
lambda = 0.5:

```
plot(waiting ~ eruptions, faithful, main="bandwidth=0.5", pch=".")  
lines(ksmooth(faithful$eruptions, faithful$waiting, "normal", 0.5))
```

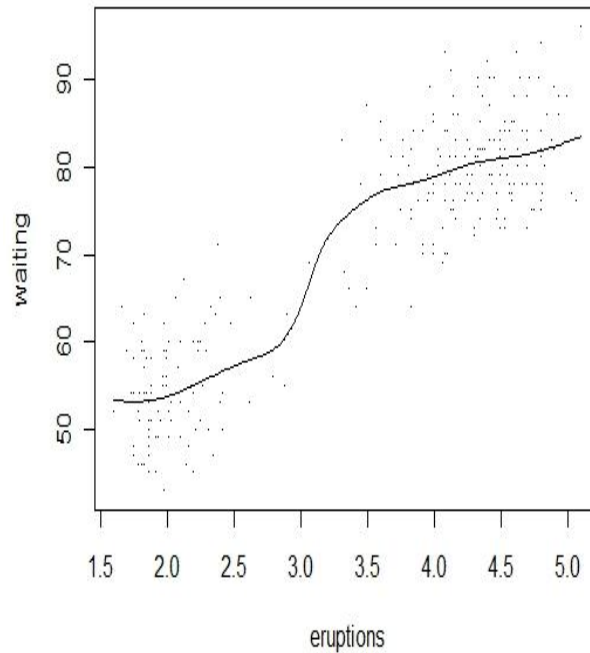
lambda = 2:

```
plot(waiting ~ eruptions, faithful, main="bandwidth=2", pch=".")  
lines(ksmooth(faithful$eruptions, faithful$waiting, "normal", 2))
```

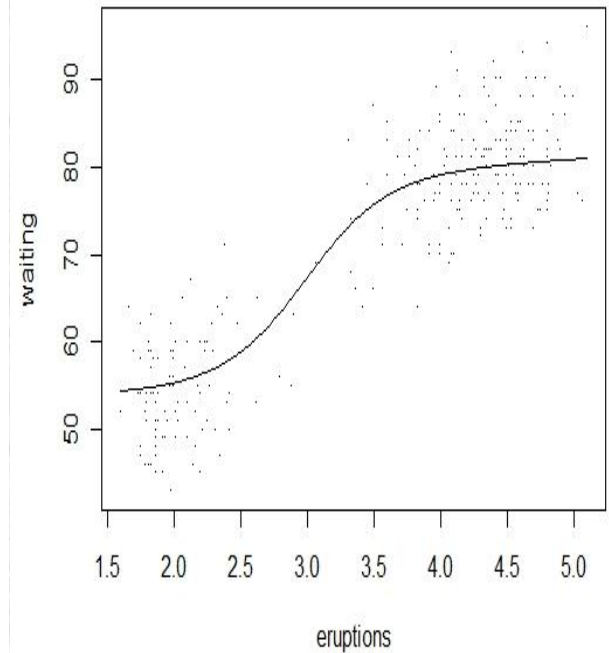
bandwidth=0.1



bandwidth=0.5



bandwidth=2



A figura central (bandwidth=0.5) é a melhor escolha entre as três opções.

- λ pode ser escolhido iterativamente utilizando este método subjetivo,
- Métodos automáticos de seleção da "quantidade de suavização" também são utilizados,
- Método de validação cruzada (CV) é amplamente utilizado, porém tem um alto custo computacional,

$$CV(\lambda) = \frac{1}{n} \sum_{j=1}^n (y_j - \hat{f}_{\lambda(j)}(x_j))^2$$

- Validação cruzada generalizada (GCV) é uma aproximação de CV,
- Utilização de métodos automáticos com cautela.

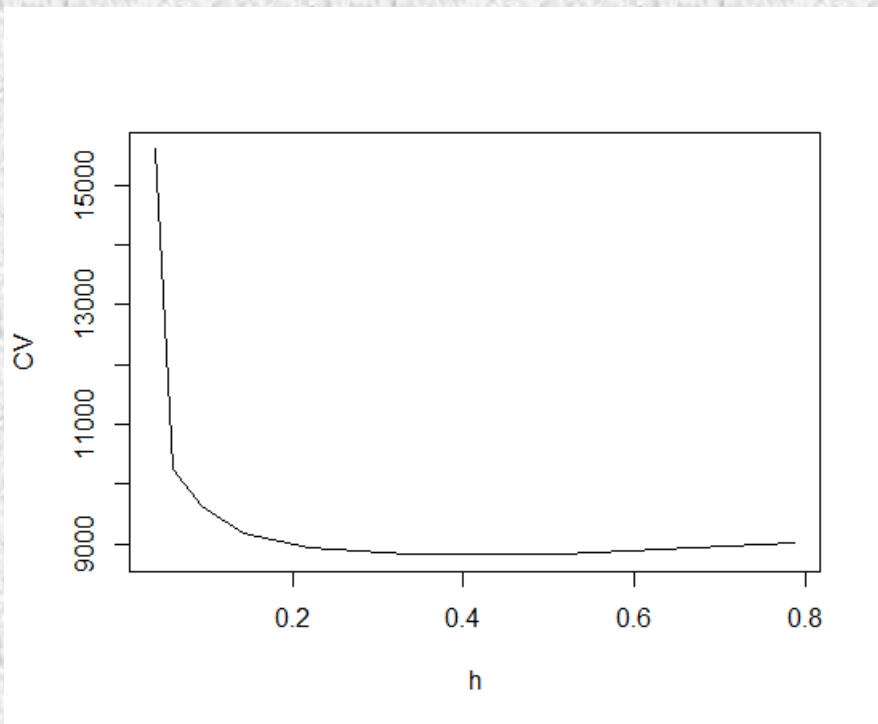
Validação cruzada:

Old Faithful

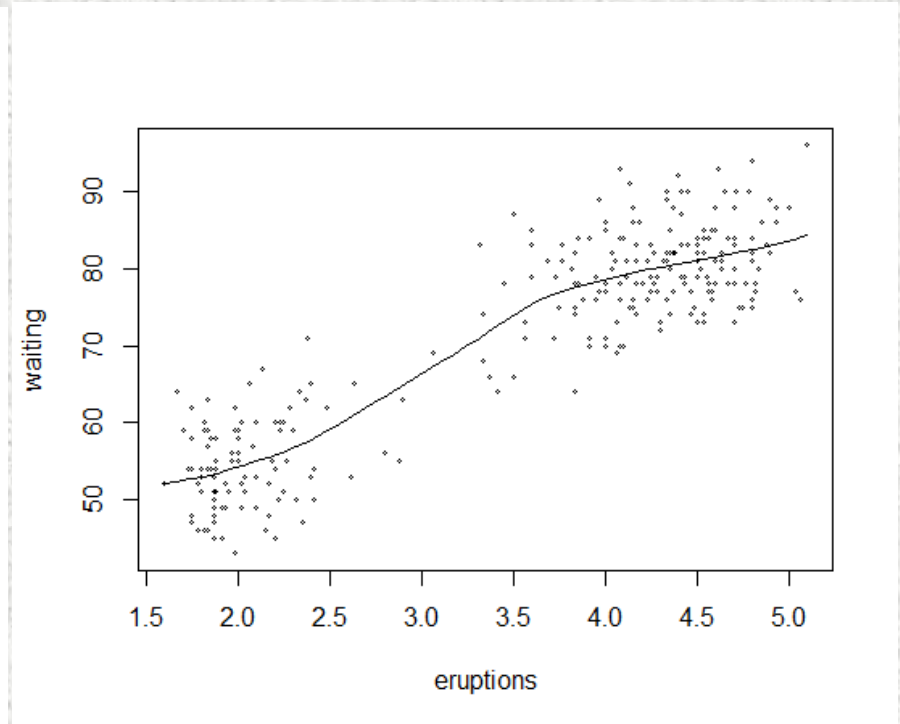
```
>library(sm) # utiliza o Kernel Gaussiano onde a suavização é o desvio padrão do Kernel
```

```
>hm <- hcv(faithful$eruptions, faithful$waiting, display="lines")
```

```
>sm.regression(faithful$eruptions, faithful$waiting, h=hm, xlab="eruptions", ylab="waiting")
```



Critério de validação cruzada como uma função da suavidade, o mínimo ocorre em 0.424.

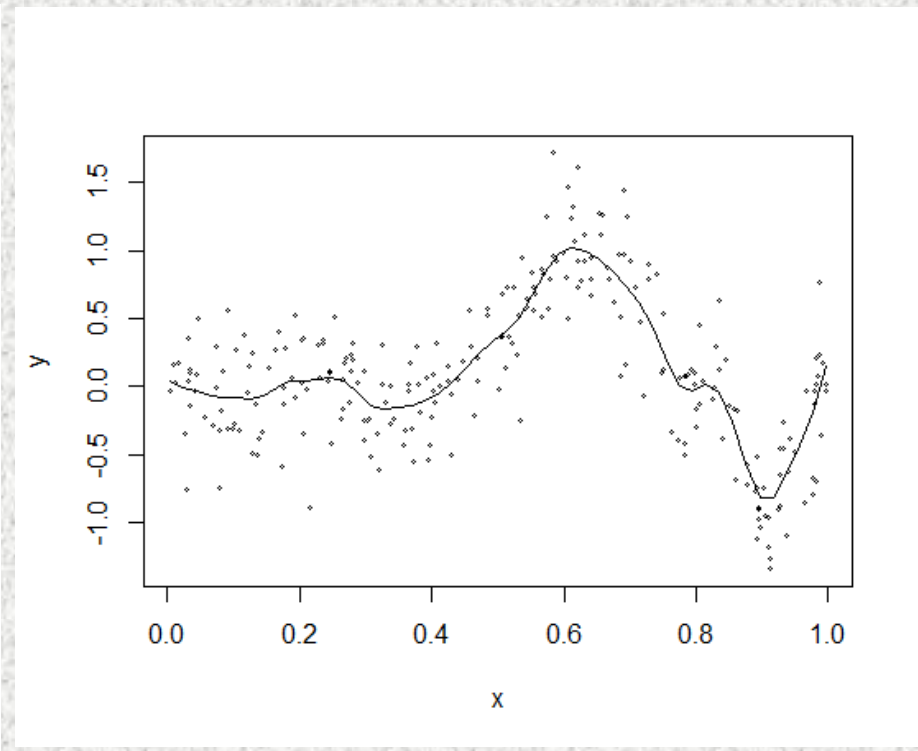
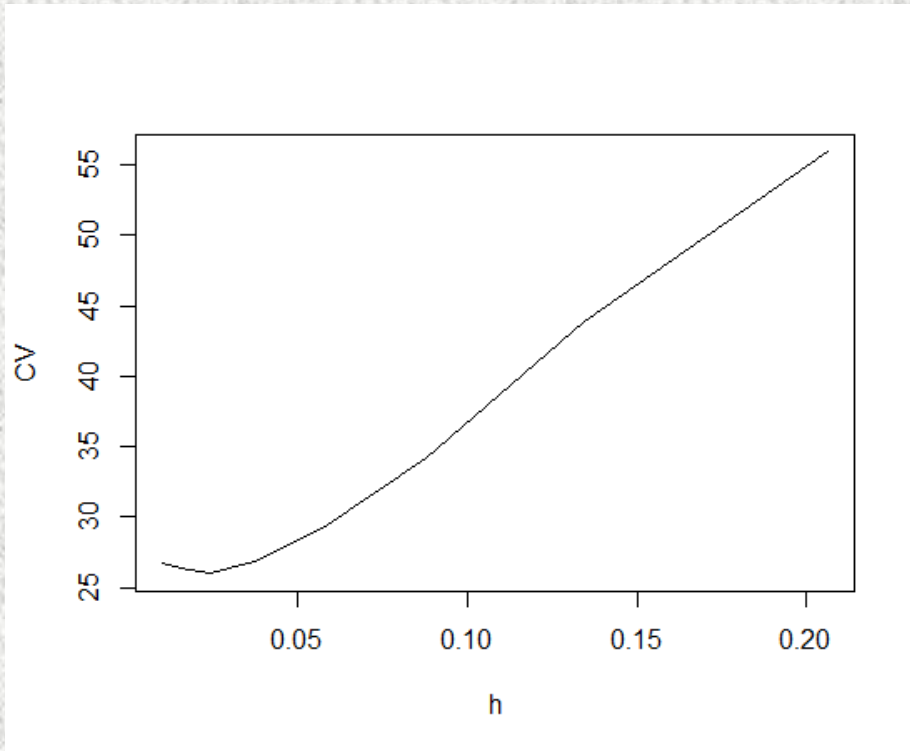


Estimador kernel com o valor estimado de suavização.

Ex A

```
>hm <- hcv(exa$x, exa$y, display="lines")
```

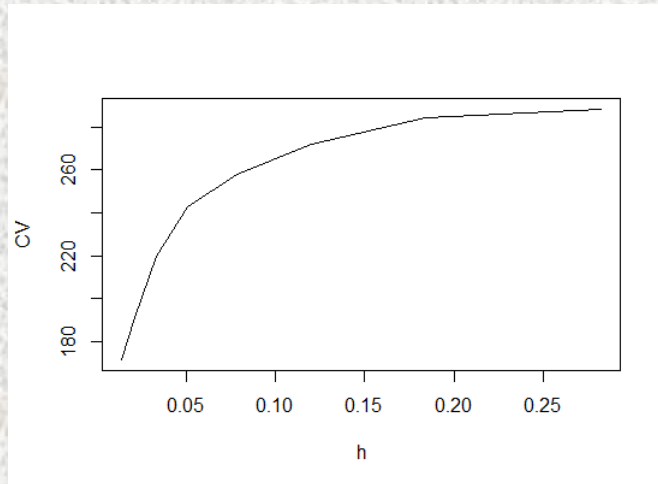
```
>sm.regression(exa$x, exa$y, h=hm, xlab="x", ylab="y")
```



Validação cruzada para exemplo A, mínimo ocorre em $h=0.022$.

Ex B

```
>hm <- hcv(exb$x, exb$y, display="lines")
```



hcv: boundary of search area reached.
Try readjusting hstart and hend.

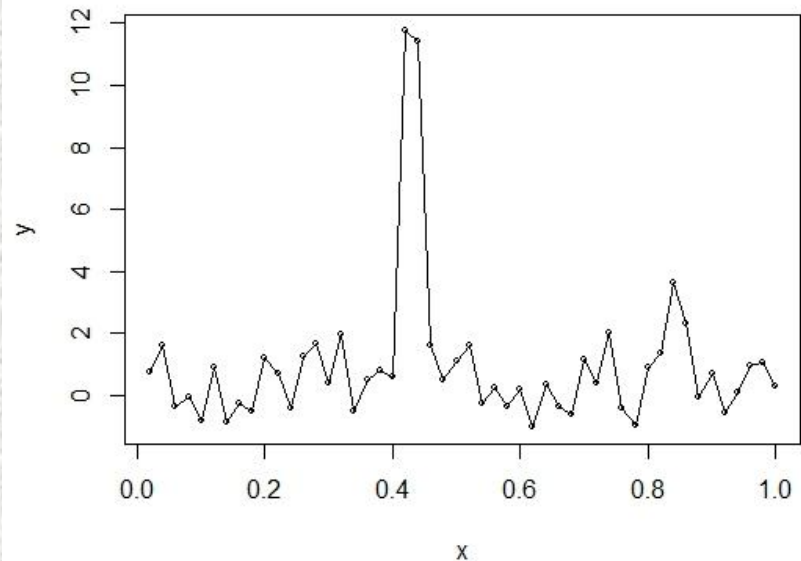
hstart: 0.01412223

hend : 0.2824446

	h	cv
[1,]	0.01412223	171.4676
[2,]	0.02166530	190.9929
[3,]	0.03323733	219.8658
[4,]	0.05099029	242.9889
[5,]	0.07822560	258.1320
[6,]	0.12000804	272.2397
[7,]	0.18410763	284.3874
[8,]	0.28244455	288.4751

Erro em hcv(exb\$x, exb\$y, display = "lines") :

```
>sm.regression(exb$x, exb$y, h=0.005)
```

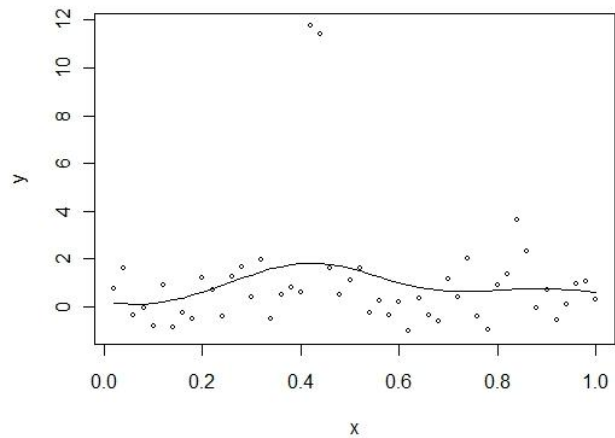


Ex B: largura da banda

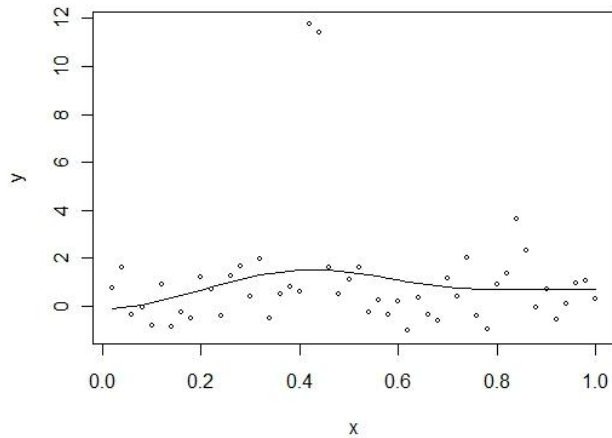
hstart: 0.01412223

hend : 0.2824446

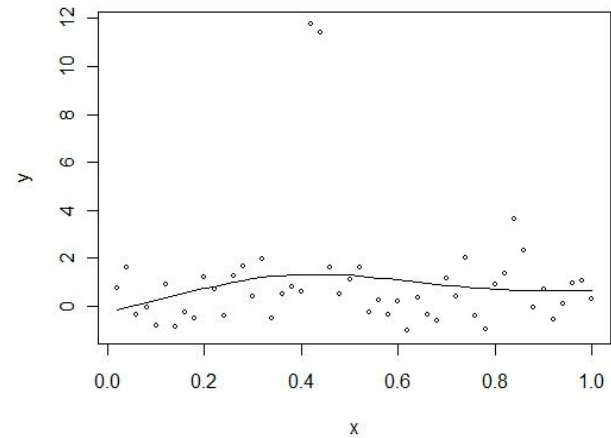
h=0.141



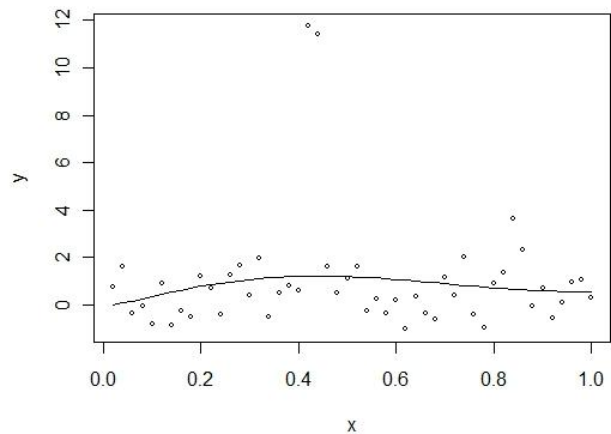
h=0.180



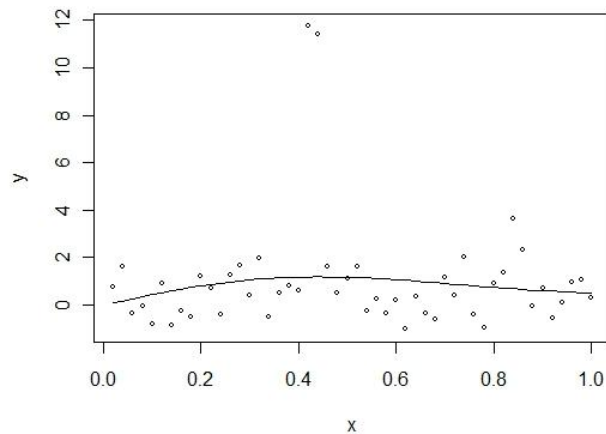
h=0.220



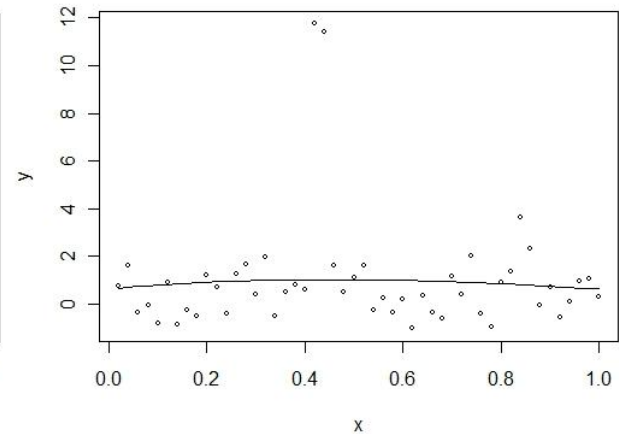
h=0.260



h=0.282



h=0.500



Método das *splines*

Suavização das splines

```
>plot (waiting ~ eruptions, faithful, pch=".")
```

```
>lines(smooth.spline(faithful$eruptions, faithful$waiting))
```

```
>plot (y ~ x, exa, pch=".")
```

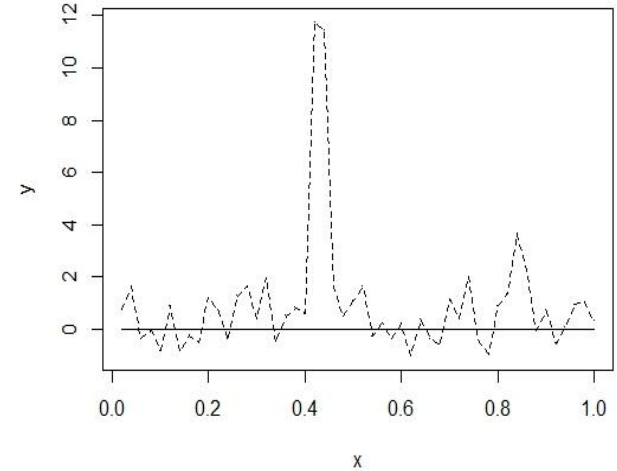
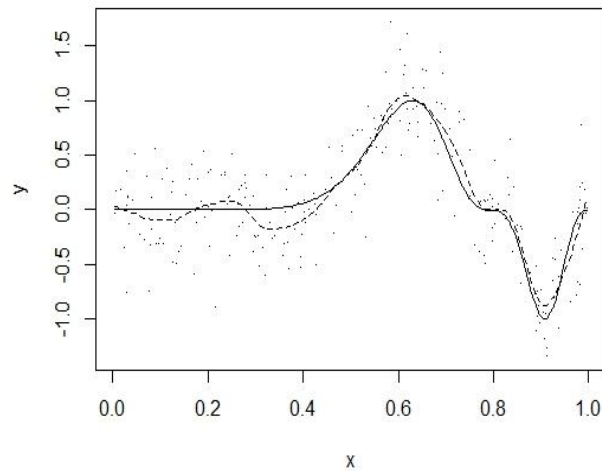
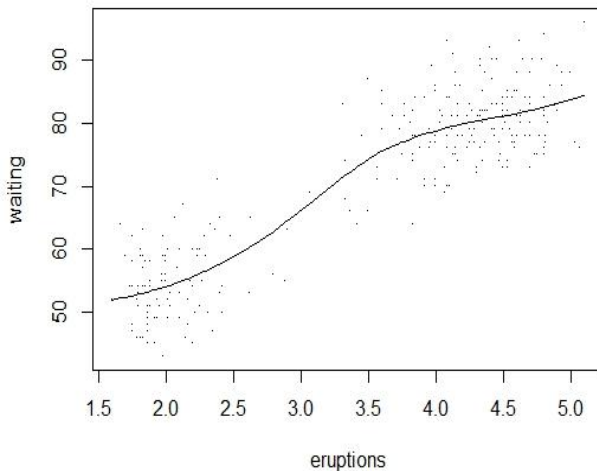
```
>lines(exa$x, exa$m)
```

```
>lines(smooth.spline(exa$x, exa$y), lty=2)
```

```
>plot (y ~ x, exb, pch=".")
```

```
>lines(exb$x, exb$m)
```

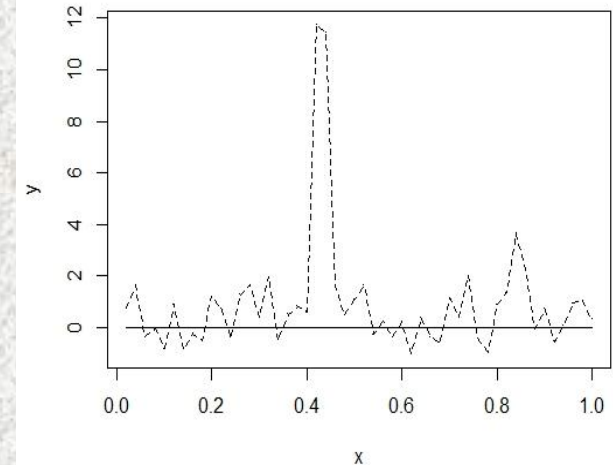
```
>lines(smooth.spline(exb$x, exb$y), lty=2)
```



Função cúbica atende ao requisito de continuidade e suavidade.

SPLINES DE SUAVIZAÇÃO:

- Ajuste melhorado da função,
- Escolha automática pode ser perigosa (exemplo B: interpolação dos dados).

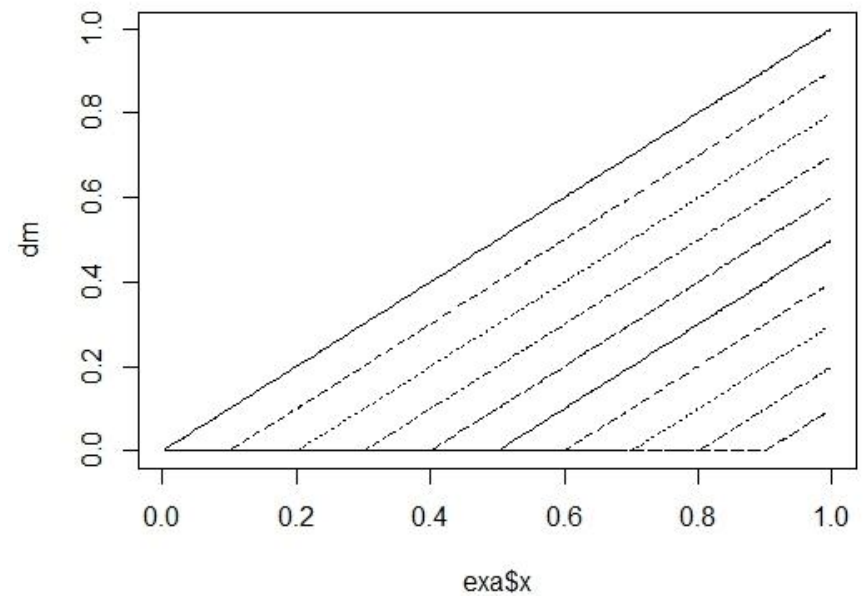
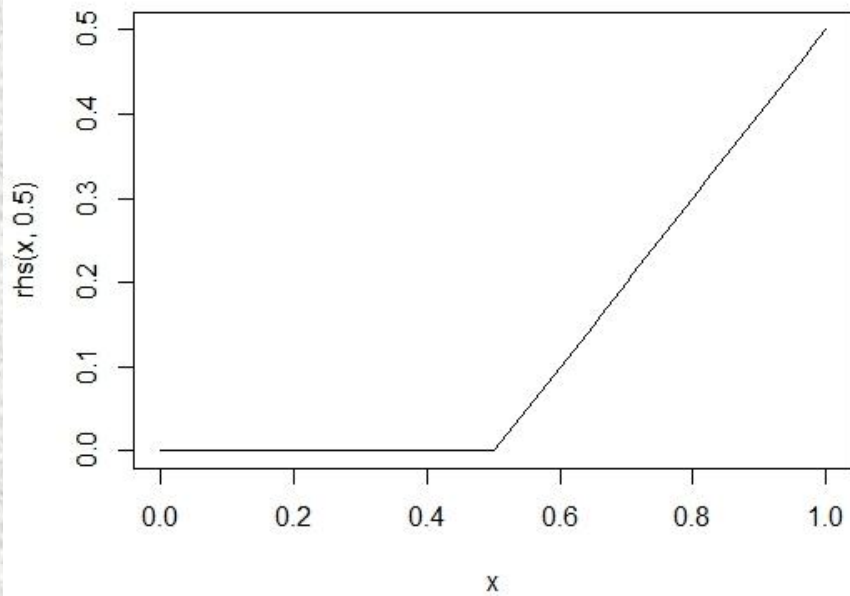


SPLINES DE REGRESSÃO:

- Difere da anterior pois utiliza uma quantidade de nós menor que o tamanho da amostra,
- Número de nós (ou ‘janelas’) controla a “quantidade de suavização”, anteriormente controlada pelo λ .
- Método não-paramétrico fornece a liberdade de escolha do número de nós (‘janelas’).

Splines de regressão

```
>rhs <- function (x,c) ifelse (x > c, x - c, 0)
>plot (rhs)
>curve (rhs(x,0.5), 0, 1)
>knots <- 0:9/10; knots
>dm <- outer (exa$x, knots, rhs)
>matplot (exa$x, dm, type="l", col=1)
```

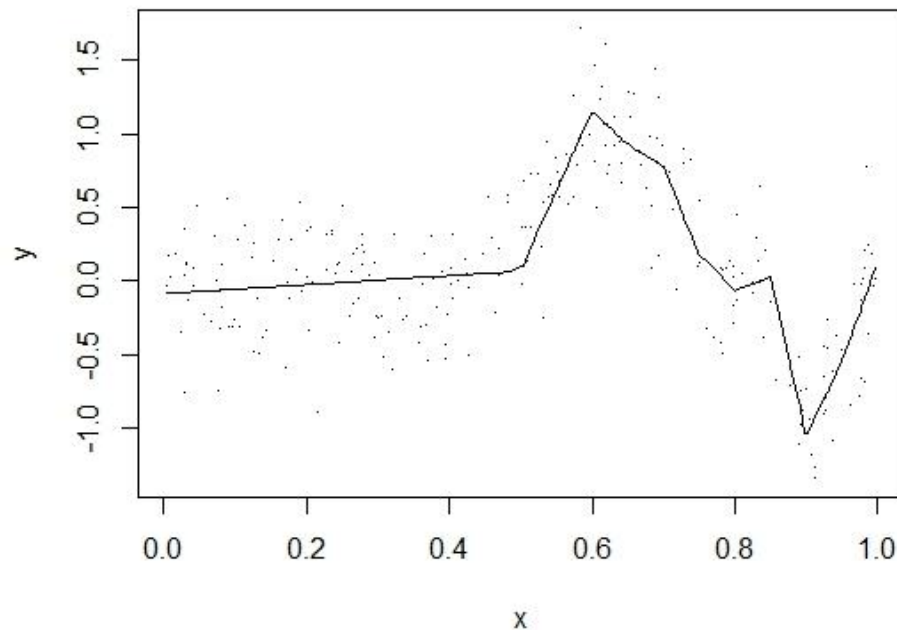
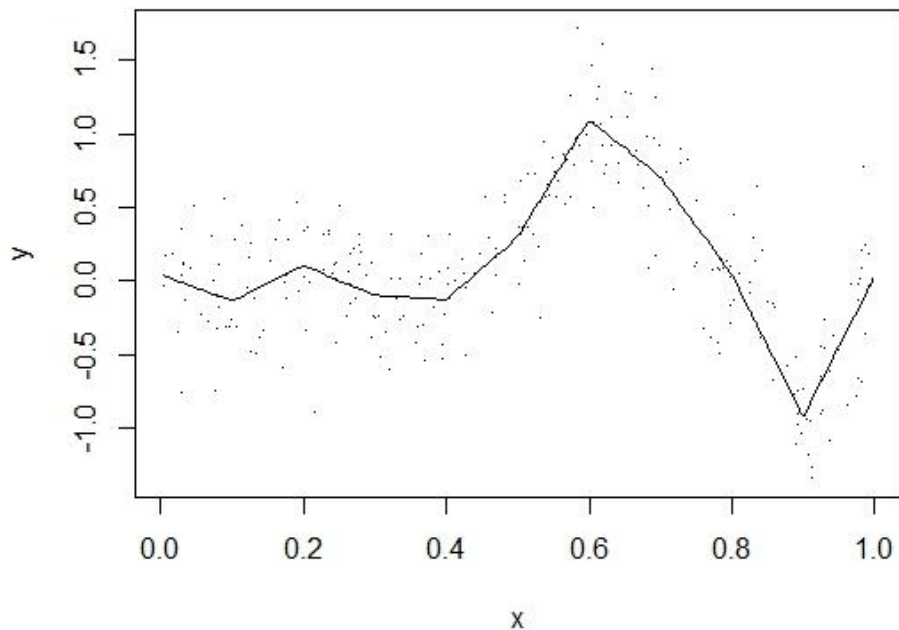


ajuste de regressão

```
>g <- lm(exa$y ~ dm)
>plot (y ~ x, exa, pch=".", xlab="x", ylab="y")
>lines (exa$x, predict(g))
```

adensamento dos nós

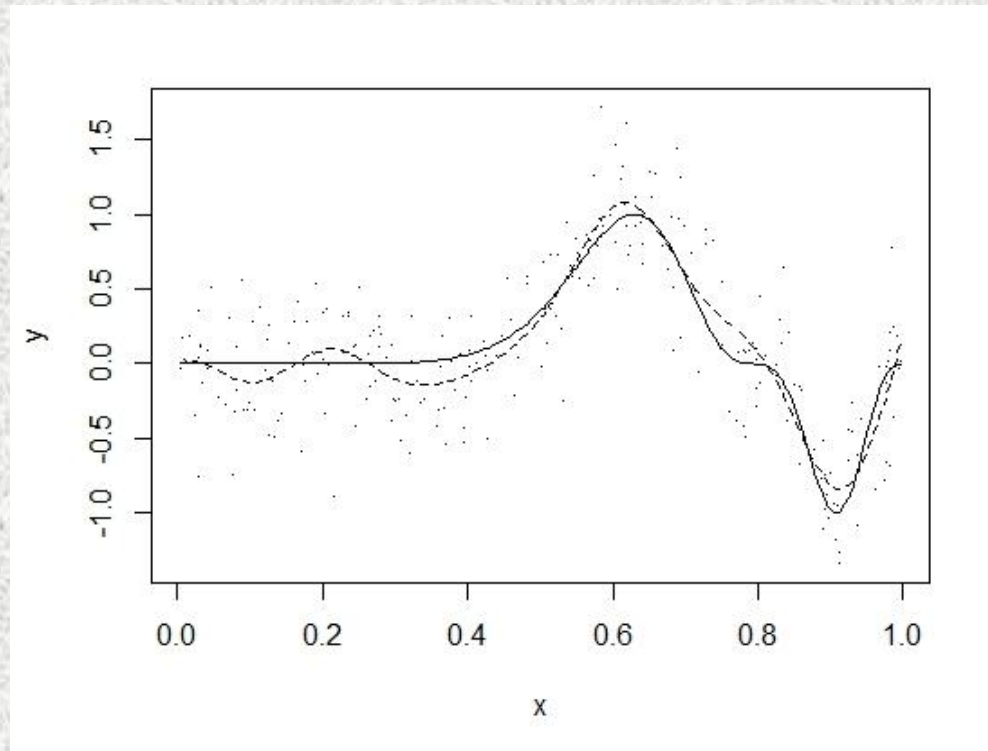
```
>newknots <- c(0, 0.5, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95)
>dmn <- outer(exa$x, newknots, rhs)
>gn <- lm(exa$y ~ dmn)
>plot (y ~ x, exa, pch=".", xlab="x", ylab="y")
>lines (exa$x, predict(gn))
```



Só é aplicado quando a curvatura original é conhecida!

#Ajuste mais fino:

```
>library (splines)
>sm1 <- lm(y ~ bs(x,12), exa)
>plot (y ~ x, exa, pch=".")
>lines(m ~ x, exa)
>lines(predict(sm1) ~ x, exa, lty=2)
```



Poderia ser melhorado incluindo nós na região de maior curvatura e menos na região mais plana, ou seja, alterando a 'largura da janela' em função da distribuição dos dados.

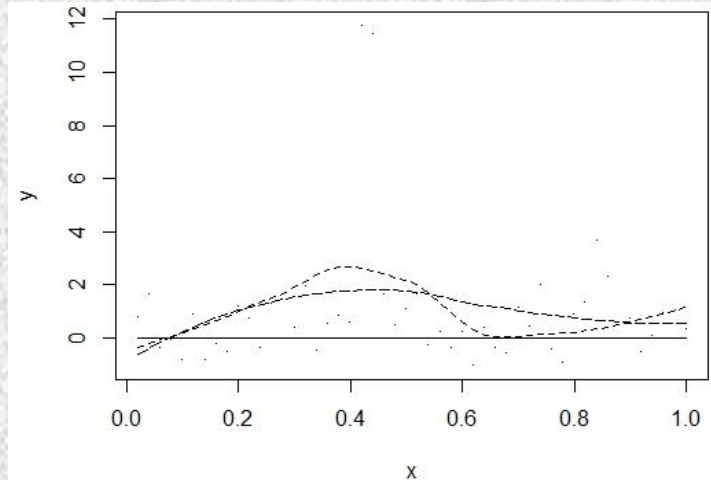
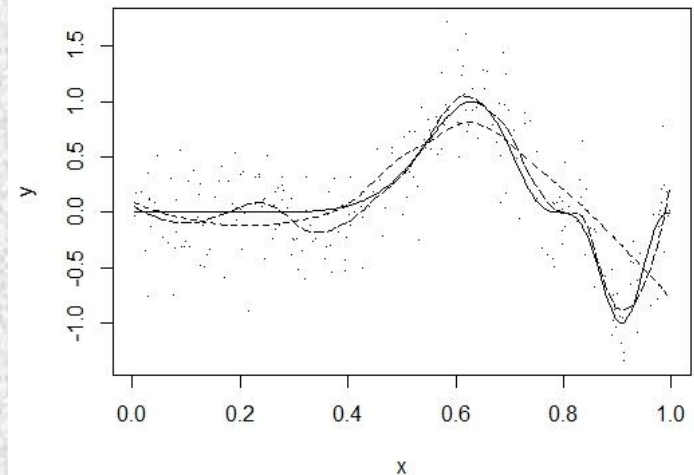
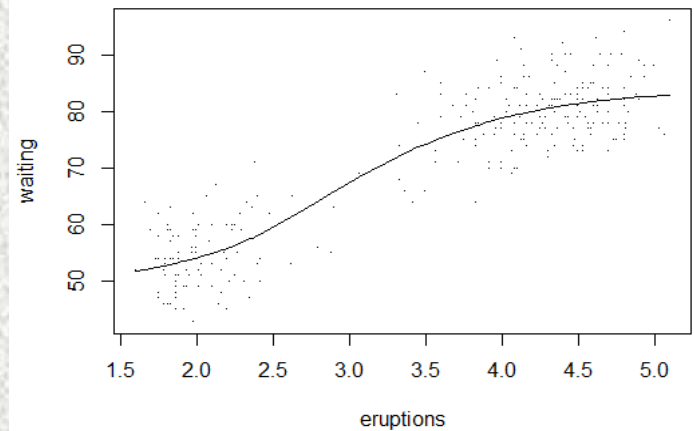
Método dos polinômios locais (*local polynomials*)

#Polinômios locais:

```
>plot (waiting ~ eruptions, faithful, pch=".")  
>f <- loess(waiting ~ eruptions, faithful)  
>i <- order (faithful$eruptions)  
>lines(f$x[i], f$fitted[i])
```

```
>plot(y ~ x, exa, pch=".")  
>lines(exa$x, exa$m, lty=1)  
>f <- loess(y ~ x, exa)  
>lines(f$x, f$fitted, lty=2)  
>f <- loess(y ~ x, exa, span=0.22)  
>lines(f$x, f$fitted, lty=5)
```

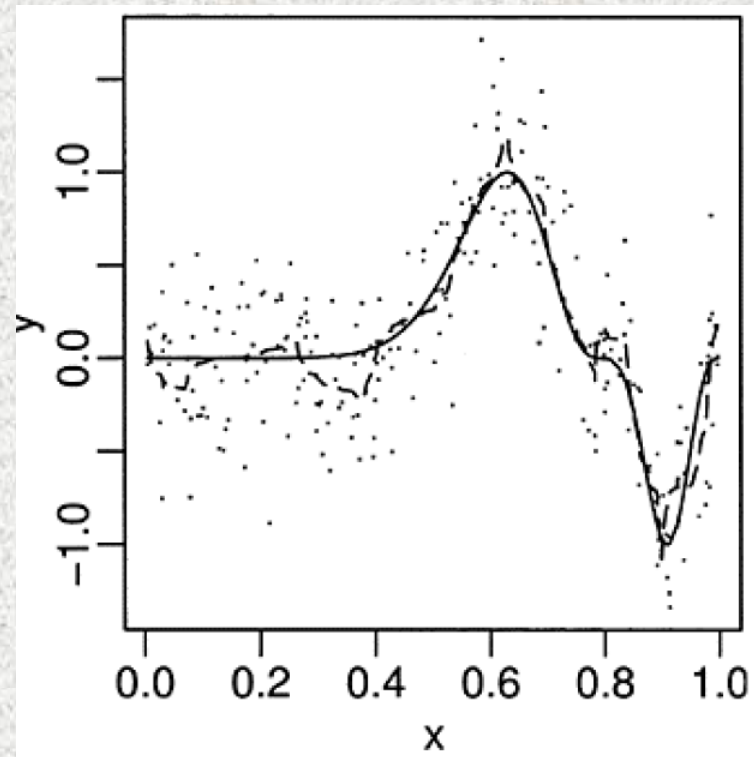
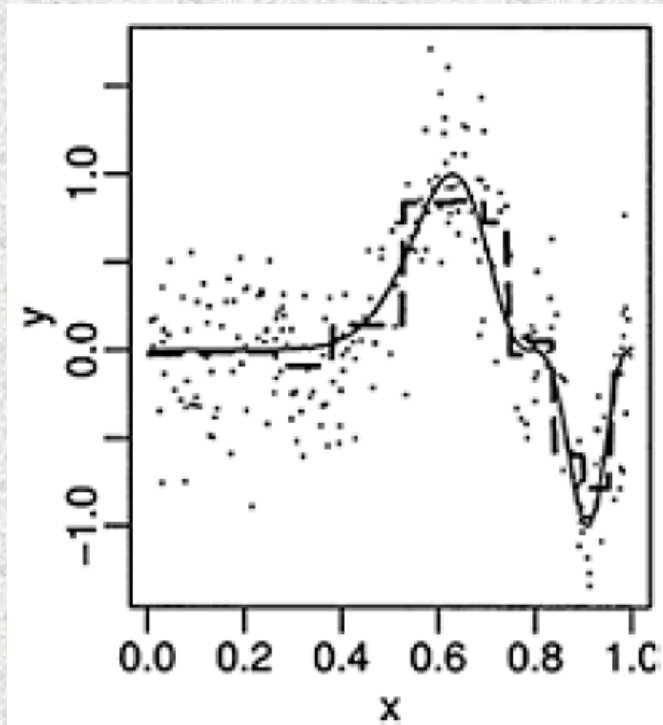
```
>plot (y ~ x, exb, pch=".")  
>f <- loess(y ~ x, exb)  
>lines(f$x, f$fitted, lty=2)  
>f <- loess(y ~ x, exb, span=1)  
>lines(f$x, f$fitted, lty=5)  
>lines(exb$x, exb$m)
```



OUTROS MÉTODOS

WAVELETS:

- Utilização de polinômios ortogonais e bases de Fourier,
- Propriedade de multirresolução,
- Alteração da largura da janela de acordo com a oscilação dos dados.



OUTROS MÉTODOS

NEAREST NEIGHBOR:

- O comprimento da janela varia de forma a comportar a mesma quantidade de pontos.

VARIABLE BANDWIDTH:

- Bandas menores nas regiões de alta variabilidade e bandas mais largas quando a função é mais suave,
- Requer conhecimento a priori da suavidade da função relativa a todo o conjunto de dados.

RUNNING MEDIANS:

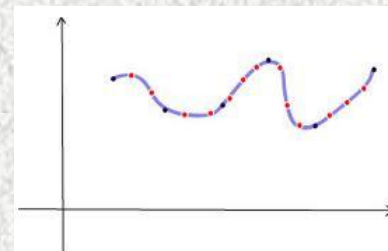
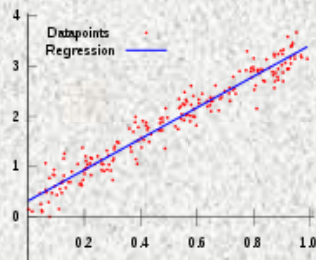
- Utilizado para conjunto de dados com muitos outliers,
- Métodos baseados em médias locais são muito afetados por dados discrepantes, nesse contexto a utilização de medianas é uma boa alternativa,
- Produz ajuste visual rústico.

COMPARAÇÃO DE MÉTODOS – CASO UNIVARIADO

- **Dados homogêneos e pouco ruído:**

interpolação

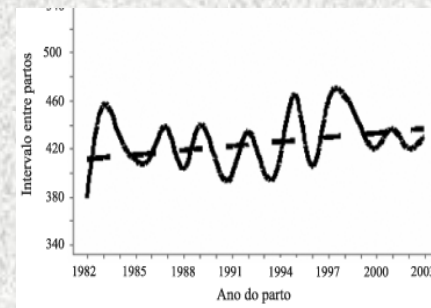
regressão linear simples



- **Ruído moderado:**

métodos não paramétricos

(sinal suficiente para justificar um ajuste flexível)



- **Muito ruído:**

preferência para métodos paramétricos

(não há sinal suficiente para justificar

modelo mais complexo)

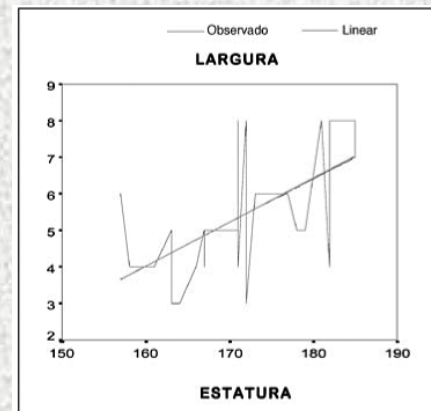
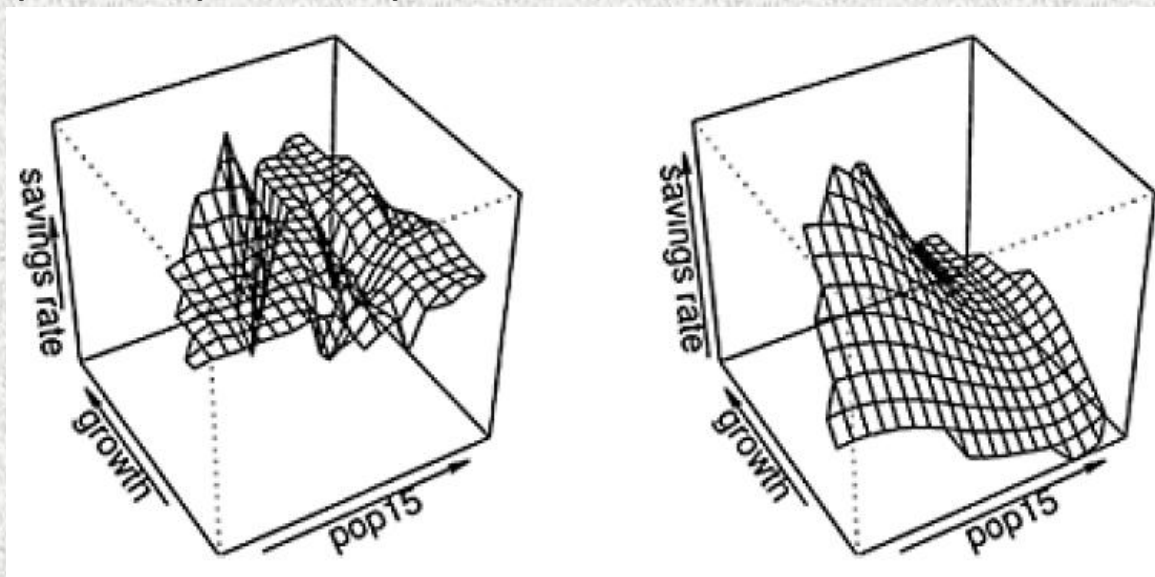


Gráfico 3: Diagrama de dispersão entre a largura do semitendão e a estatura do cadáver em centímetros.

PREDITORES MULTIVARIADOS

$$y_i = f(\mathbf{x}) + \varepsilon_i \quad i=1, \dots, n$$

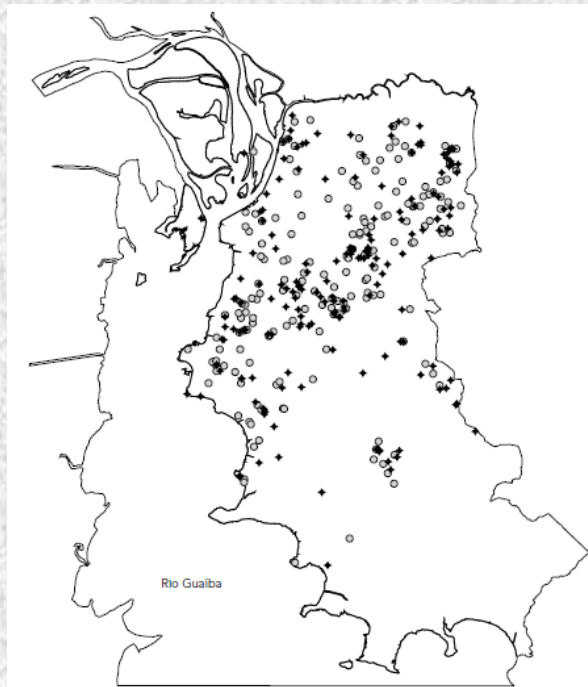
- Muitos dos métodos são estendidos para dimensões maiores,
- Ajustes não-paramétricos são mais complexos,
- Não costuma ser aplicado para mais de dois preditores (visualização não é possível),
- Maldição da dimensionalidade: tamanho da janela suficiente para capturar pontos para a média local.



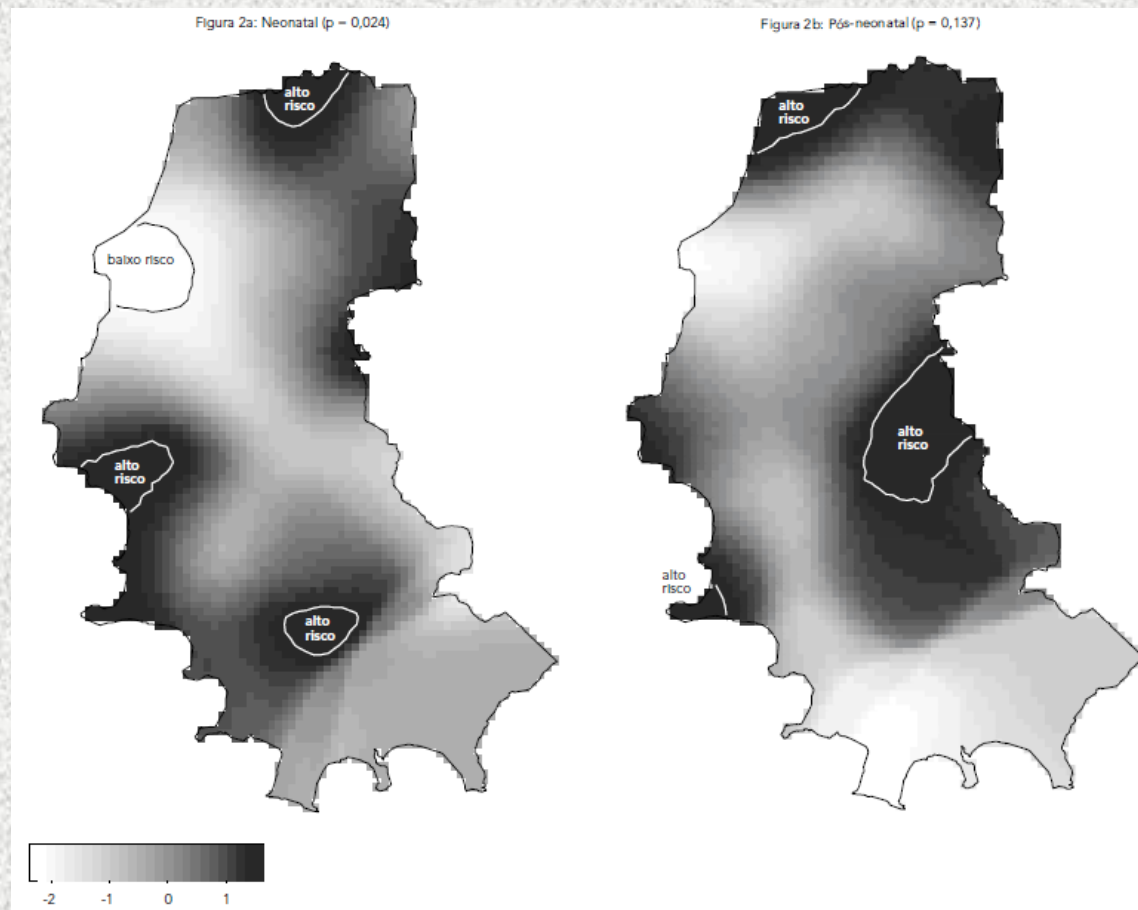
UTILIZAÇÕES DE MODELO DE REGRESSÃO NÃO-PARAMÉTRICA E SEMI-PARAMÉTRICA

MORTALIDADE INFANTIL EM PORTO ALEGRE

- Variáveis: idade e grau de instrução da mãe, tipo de gravidez e parto, localização da residência, peso ao nascer, semanas gestacionais e sexo de crianças que nasceram durante o ano de 1998, no Município de Porto Alegre, e que morreram antes de completarem o primeiro ano de vida,
- Abordagem semiparamétrica, ou abordagem de Modelos Aditivos Generalizados (GAM),
- Estimação espacial do risco.



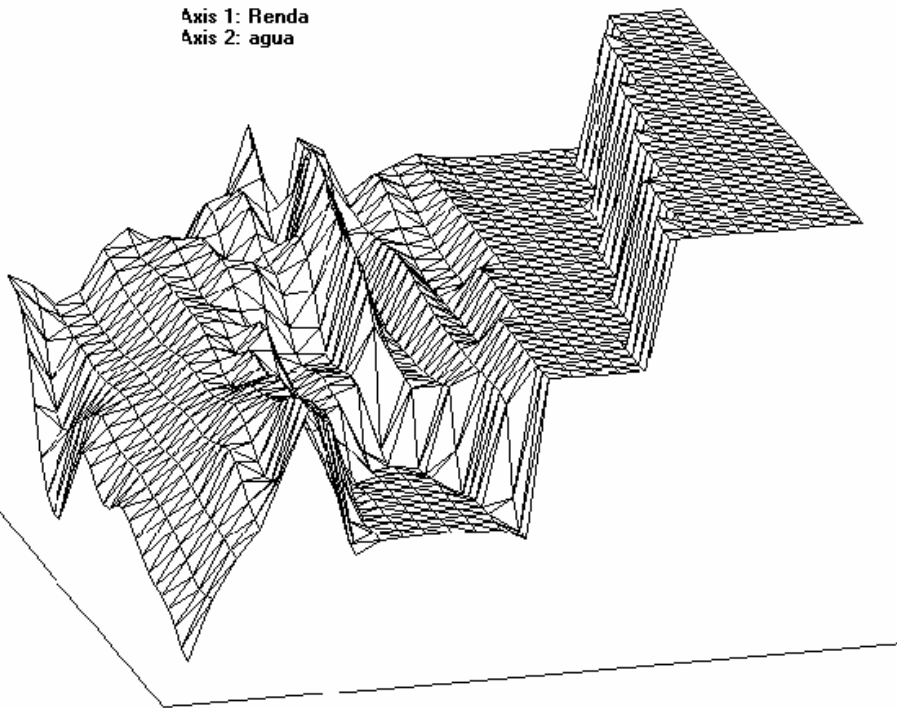
óbitos
+ neonatal
o posneonatal



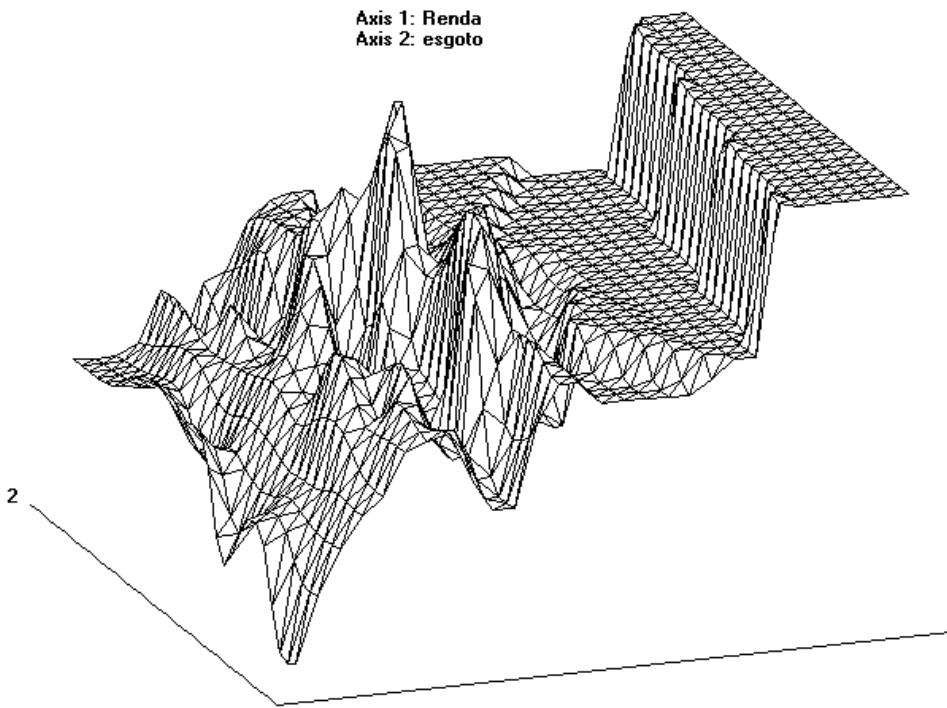
ECONOMETRIA

- Métodos de Econometria Não-Paramétrica.
- Relação entre a esperança de vida ao nascer e as condições sócio-econômicas dos municípios nordestinos a partir de variáveis como renda per capita, proporção de domicílios com água canalizada e proporção de domicílios com acesso a rede de esgotos.

Axis 1: Renda
Axis 2: agua



Axis 1: Renda
Axis 2: esgoto



REFERENCIAS

- Faraway, J. J. ***Extending the linear model with R***: Generalized linear, mixed effects and nonparametric regression models. Taylor & Francis Group, LLC: 2006.
- Fox, J. ***Nonparametric Regression*** - Appendix to An R and S-PLUS Companion to Applied Regression. 2002.
- Shimakura, S. E.; *et al.* Distribuição espacial do risco: modelagem da mortalidade infantil em Porto Alegre, Rio Grande do Sul, Brasil **in Cad. Saude Publica**, 1251-1261, Rio de Janeiro, set-out, 2001.
- Simonassi, A. G. **Econometria Não Paramétrica e Expectativa e de Vida nos Municípios do Nordeste**: Uma Aplicação do Estimador de Nadaraya-Watson. FGV. Disponível em: <http://www.bnb.gov.br/content/aplicacao/ETENE/Anais/docs/mesa9_texto3.pdf>
- Von Zuben, F. ***Regressão Paramétrica e Não-Paramétrica***. DCA/FEEC/Unicamp. Notas de aula. Disponível em: <<ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia353/aula13.pdf>>