Análise de Regressao e Informaçao Mútua

David R. Brillinger
Departamento de Estatística
Universidade da California em Berkeley

www.stat.berkeley.edu/~brill

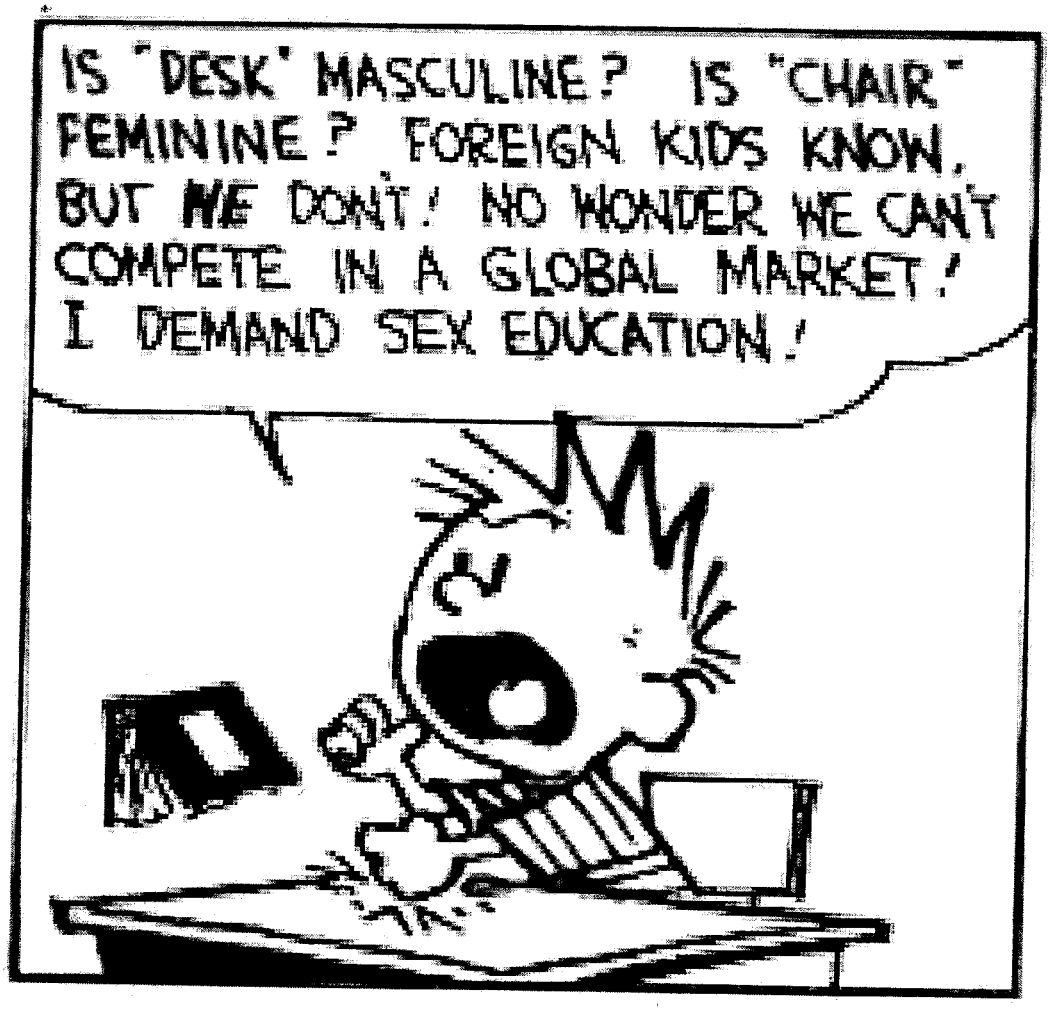
$2\pi \neq 1$

$2\pi \neq 1$

$2\pi \neq 1$

Estrutura da palestra.

1. Introdução
2. Regressão
3. Informação mútua
4. Exemplos e resultados
5. Extensões
6. Sumário
7. Agradecimentos
8. Referências
9. Provas


IS "DESK" MASCULINE? IS "CHAIR" FEMININE? FOREIGN KIDS KNOW. BUT ME DON'T! NO WONDER WE CAN'T COMPETE IN A GLOBAL MARKET! I DEMAND SEX EDUCATION!

1. Introdução.

Ciência estuda relações

Análise de regressão estuda relações entre variáveis Y e X

Pergunta: o que é a força de uma relação?

Uma resposta: o coeficiente de determinação

Outra resposta: o coeficiente de informação mútua

2. Regressão standard.

Coeficiente de determinação

$$\rho_{XY}^2 = corr\{X,Y\}^2$$

Um método simétrico e invariante por:

1) independência

2) variação explicada

3) força de dependência linear

4) incerteza de estimadores

A gente quer mais!

A mãe última de futebol

3. Informação mútua.

Variáveis discretas.

$$Prob\{X_j = j, Y_k = k\} = p_{jk}$$

$$I_{XY} = \sum_{j,k} p_{jk} \log \frac{p_{jk}}{p_j p_k}, \quad p_{jk} \neq 0$$

Variáveis contĭnuas.

Dados $x_j \varepsilon \delta_j$, $y_k \varepsilon \Delta_k$ temos

$$Prob\{X \varepsilon \delta_j, Y \varepsilon \Delta_k\} \approx p(x_j, y_k) |\delta| |\Delta|$$

de pois obtemos

$$\iint p(x,y) \log \frac{p(x,y)}{p_X(x) p_Y(y)} dxdy$$

(Variáveis mistas.)

Propriedades de I_{XY}.

1) Não negativo, $I_{XY} \geq 0$

2) Invariante, $I_{XY} = I_{UV}$ se 1-1 transformações

3) Medida da força de dependência,

i) $I_{XY} = 0 <=> X \ indep \ Y$

ii) $I_{XY} = \infty$ se $Y = g(X)$

iii) $I_{XZ} \leq I_{XY}$ se X indep $Z \mid Y$ $\qquad X - Y - Z$

Usos.

Perguntas - mudança?, tendência?, autocorrelação?, dimensão?, aderência ao modelo?, ...

Estimação - atraso, registro de imagens, seleção de variáveis, seleção do modelo, associação, ...

Modelo não correto -

$$\iint p(x,y)\log\frac{p(x,y)}{p_X(x)p_Y(y)}dxdy \geq \iint p(x,y)\log\frac{q(x,y)}{q_X(x)q_Y(y)}dxdy$$

Predição - nível mais baixo

$$E\{Y - g(X)\}^2 \geq \frac{1}{2\pi e}\exp\{2(I_{YY} - I_{XY})\}$$

Exemplos.

Normal bivariada

$$I_{XY} = -\frac{1}{2}\log(1-\rho_{XY}^2)$$

Caso de regressão, p. ex. modelo linear generalizado

$$I_{XY} = E \log \frac{p_{Y|X}(Y|X)}{p_Y(Y)}$$

O jogo primeiro do futebol do mundo

Estimação. Dados $(x_i, y_i,\ i=1,\ldots,n)$

Paramétrica.

Modelo $p(x,y \mid \theta)$, com X e Y independentes quando $\theta = 0$

$$p(x,y \mid 0) = p_X(x)p_Y(y)$$

Teste de independência

$$\hat{I}_{XY} = \log(razao\ de\ verossimilhan\c{c}a/n)$$

(n tamanho da amostra)

Distribuição null approx

$$\chi^2_\nu \,/\, 2n,\ \nu = dim(\theta)$$

$$\iint p(x,y) \log p(x,y)dxdy - \iint p(x)p(y) \log p(x)p(y)\ dxdy$$

Estimação não-paramétrica.

$\hat{p}(x,y)$ estimador da $p(x,y)$, p.ex. histograma ou núcleo

$$I_{XY} \hat{=} \sum_{j,k} \hat{p}(x_j,y_k) \log \frac{\hat{p}(x_j,y_k)}{\hat{p}_X(x_j)\hat{p}_Y(y_k)}$$

Distribuição null approx

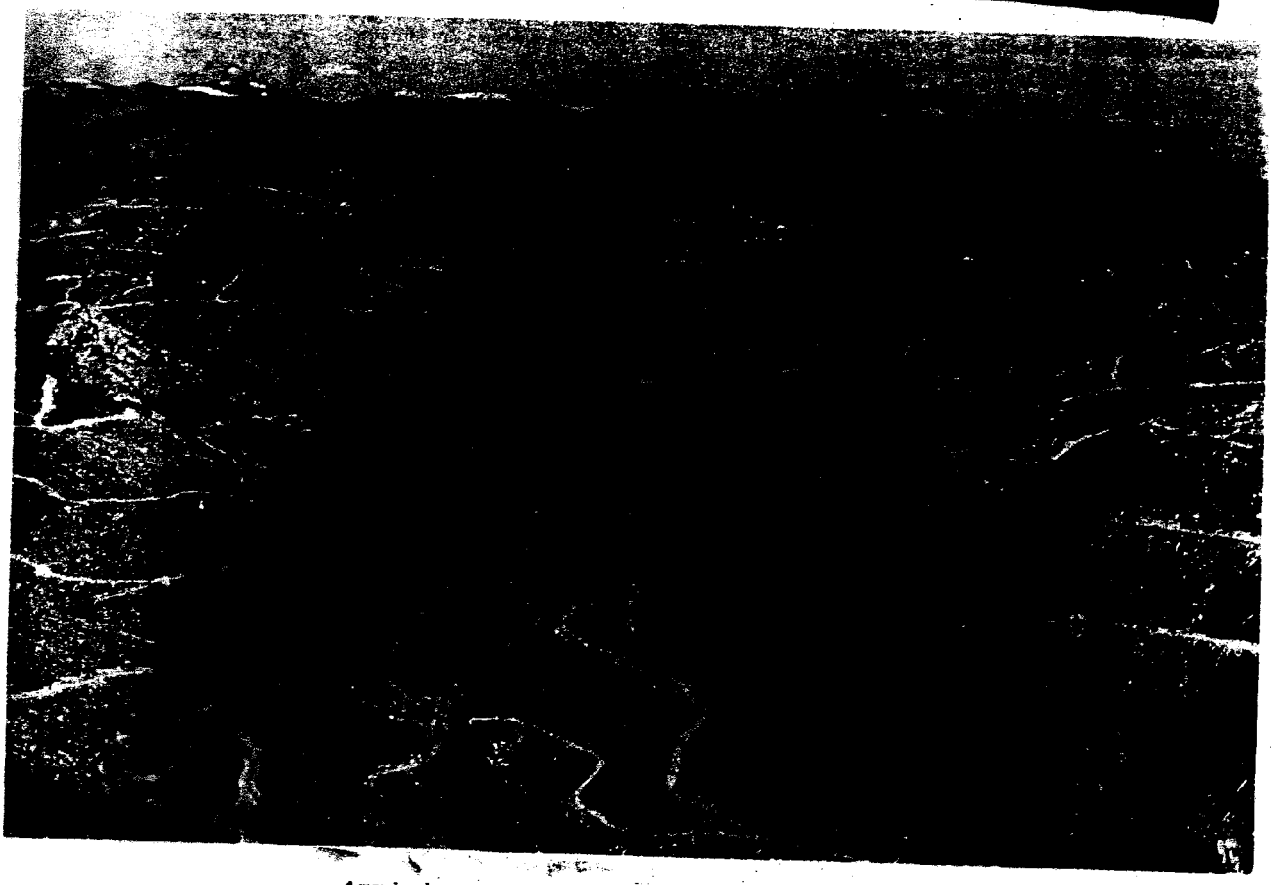
$$\chi^2_\nu \,/\, 2n,\ \nu = (J-1)*(K-1)$$

Não-null

$$\hat{I}_{XY} \rightarrow I_{XY} \textit{ em prob}$$

(Entropia - Bilmes, Fernandes, Hall & Morton, Joe, Kozachenko & Leovenko, Parzen, Robinson)

Brazil of the North



A massive clearcut on southern Vancouver Island, British Columbia

Garth Lenz

The National and Global Crisis in Canada's Forests

In Canada, one acre of forest is clearcut every 12 seconds. In Brazil, one acre is cut or burned every 9 seconds.

Controversy rages over whether Canada can appropriately be called the Brazil of the North. Politicians and the forest industry hotly contend that this title is the language of extremists. Several have travelled to Europe to assure consumers buyers of Canada's wood products that all is well in Canada's forests. Yet even as they speak, they are up to their eyebrows in a flood of information that exposes their claims as untrue. In these pages we present an overview of the crisis situation in every major forest-producing province.

Global warming, erosion, loss of biological diversity, shattering of native cultures, and dwindling economic support bases are some of the impacts which Brazil and Canada share because of deforestation. Some of these impacts will affect people around the world.

The greatest difference between the forest policies of Brazil and Canada is that, in Brazil, the desperation of population growth and poverty have been the driving factors, whereas in Canada the greed of multinational corporations and their ability to tyrannize over the public by virtue of their wealth and political influence have been the central cause. Millions of Canadian citizens do not think this difference is flattering and do not want this situation to continue. Polls have shown they are willing to pay for increased environmental protection, but the federal and provincial governments ignore them.

We don't ask that the cutting of Canada's forest stop. But we do ask that the rate of cutting be reduced to a level which is sustainable over the long term and will allow our depleted forest to be restored. This must be based on an accurate inventory of the forest across the country. Clearcutting must be replaced by ecologically sensitive methods of logging, such as selection logging. We also ask that Canada meet its stated goal of preserving 12% of the country, including an adequate and proportional amount of old-growth forest.

What's happening to Canada's forest is a crisis of national and international proportions. Please help us bring about the necessary changes.

Industry and government critics of the "Brazil of the North" campaign argue that Canadian deforestation is different because the wood is utilized and the forests are replanted. There is also massive wood waste in the clearcutting of Canada's forest, but there are many other similarities. Consider the following:

Size of Canada: 9.9 million square kilometres
Size of Brazil: 8.5 million square kilometres

Percent of Canada covered by forest:	45%
Percent of Brazil covered by Amazon rainforest:	41%
Hectares of forest cleared in Canada in 1988 (latest figures available; 1990 will be similar or higher):	1,021,619
Hectares of Brazilian Amazon cleared or burned in 1990:	1,582,000
Amount of productive Canadian forest that is now either barren or "not sufficiently restocked" after clearcutting:	30.3%
Amount of Brazilian rainforest that has disappeared:	12%
Estimated number of Indians & Metis in Canada's boreal forest:	100,000
Estimated number of Indians in the Amazonian forest:	170,000
Amount of forest officially protected in Canada:	2.4%
In Brazil:	9.2%

Source: Equinox Magazine, Forestry Statistics Canada, 1992 State of the World Report.

January 1993 - $2.00

CANADA'S FUTURE FOREST ALLIANCE

Exemplos.

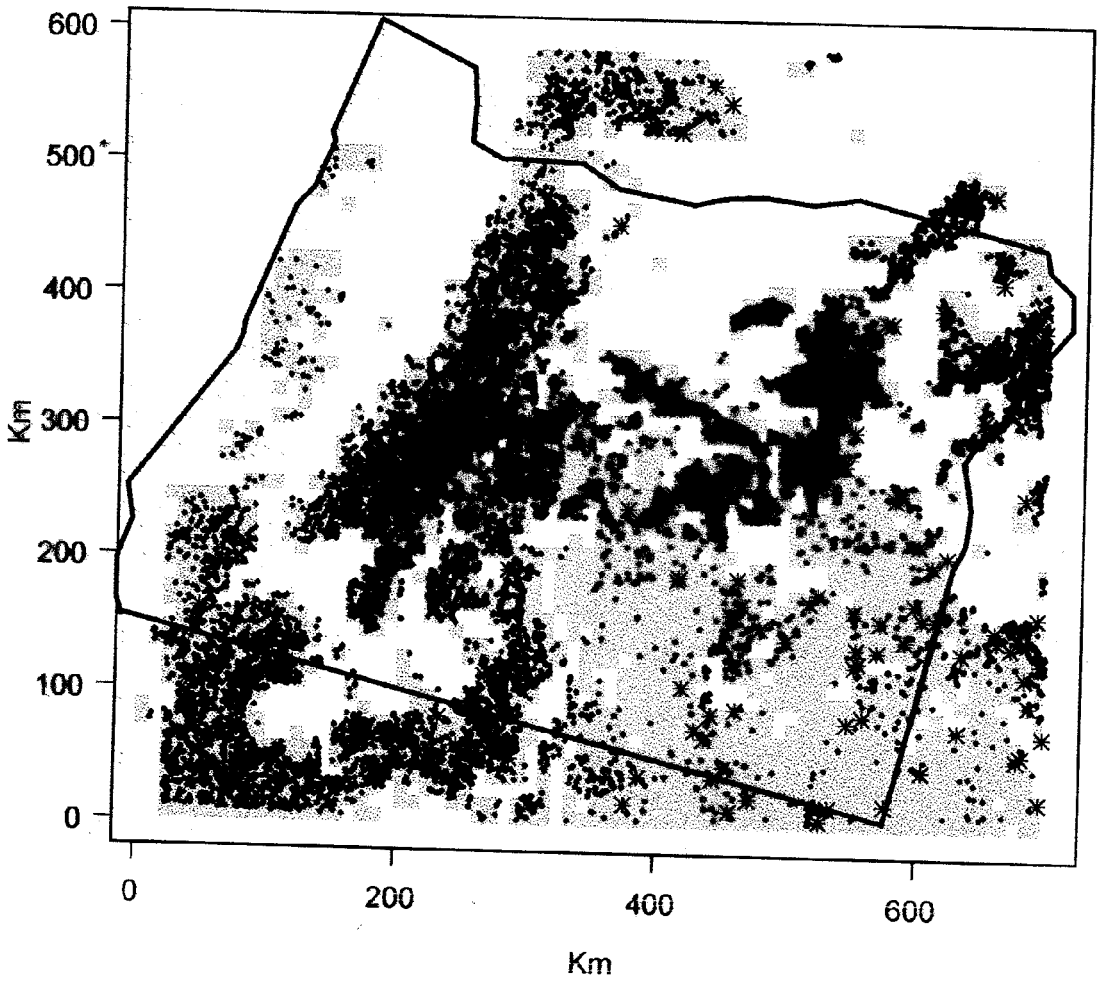
a. Discreta-contínua - incêndios florestais

Dados (espaciais-temporais) de incêndios no Estado de Oregon nos anos 1989-1996.

Pergunta - Quais variáveis são mais associados com a intensidade do incêndio?

Variáveis explicativas, X_j - índices de queima, altura, ...

Resposta, Y, binária: incêndio grande = 1, incêndio = 0


600
500
400
300
200
100
0
Km
0
200
400
600
Km

n = 13834 incêndios, m = 425 grandes

Estimador histograma

Número das celulas $2(log_2 n + 1)$

Variáveis explicativas.

2. Dia do ano

3. Elevação

4. Longitude

5. Latitude

6. Temperatura de bulbo seco

7. Umidade relativa (%)

8. Velocidade do vento (mph)

9. Indice de umidade combustível em 10 horas

10. Indice de umidade combustível em 1000 horas

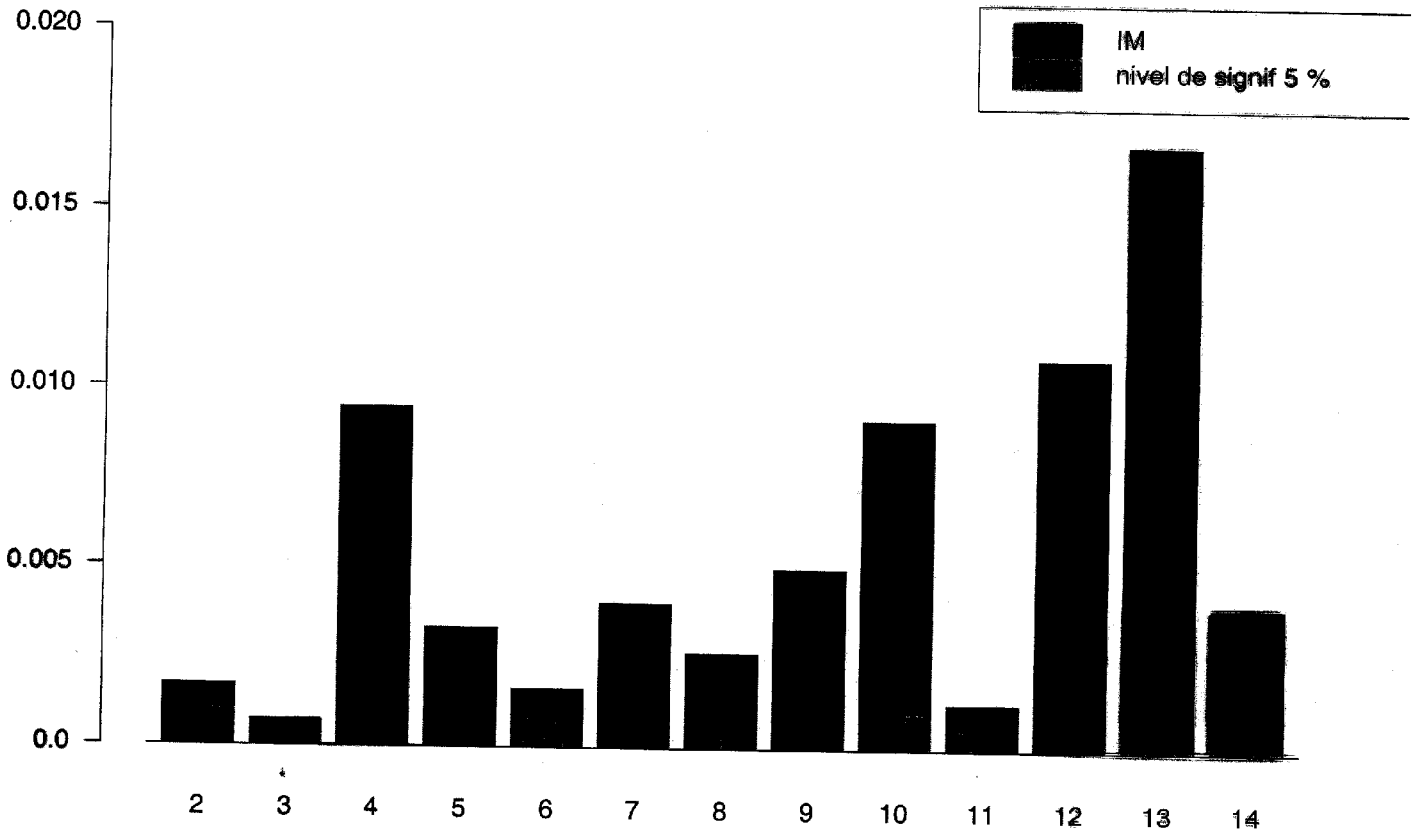
11. Indice de seca Keetch-Byram

12. Indice de potencial de incêndio

13. Indice de expansão

14. Liberação da energia

IM estimando - incendios



Canada shocks Brazil in tie

Canada 1, Brazil 1

By Reg Curren
The Canadian Press

EDMONTON — Eddy Berdusco had the most modest of goals in mind when he found himself alone in front of the Brazilian soccer net.

"I just wanted to get (the ball) on net," said Berdusco, the scoring hero in Canada's shocking 1-1 draw with Brazil on Sunday.

Berdusco broke in on Brazilian goalie Claudio Taffarel and hammered a shot into the World Cup favorite's net to give underdog Canada the draw in the friendly.

"Luckily, it went in for me," said the modest Berdusco. "It was unbelievable, even just playing against Brazil is a great honor.

"To score against them is even better, especially playing in front of 51,000 people. I just wanted to run toward the crowd."

The result will be dimly viewed by the soccer-mad Brazilian fans, who consider Canada a third-rate soccer country. After the game, the Brazilian players stomped off the field, refusing to shake hands with the Canadians.

"We didn't score when we had to," said coach Carlos Parreira.

Berdusco's stunning coup for Canadian soccer came in the 71st minute after he broke in behind two Brazilian defenders. Canada had been given little chance against the Brazilians, and most observers had predicted a soccer slaughter.

The game was the second of five the Canadian team is playing against teams headed for the World Cup, which begins in the United States later this month.

Brazil had scored late in the first half and then stormed out in the second, pinning Canada in its own end.

But Berdusco, who was a substitution in the 62nd minute, sent almost 52,000 Canadian fans into a

— CP photo

b. Duas variáveis discretas - futebol

Pergunta - Em que países existe forte associação entre o número de gols marcados e o fato de o time jogar em casa?

Y = 0,1,2,3,4+ gols de uma equipe, X - jogo em casa ou fora da casa

Dados das ligas principais, 2001-2002

http://sunsite.tut.fi/rec/riku/soccer2.html

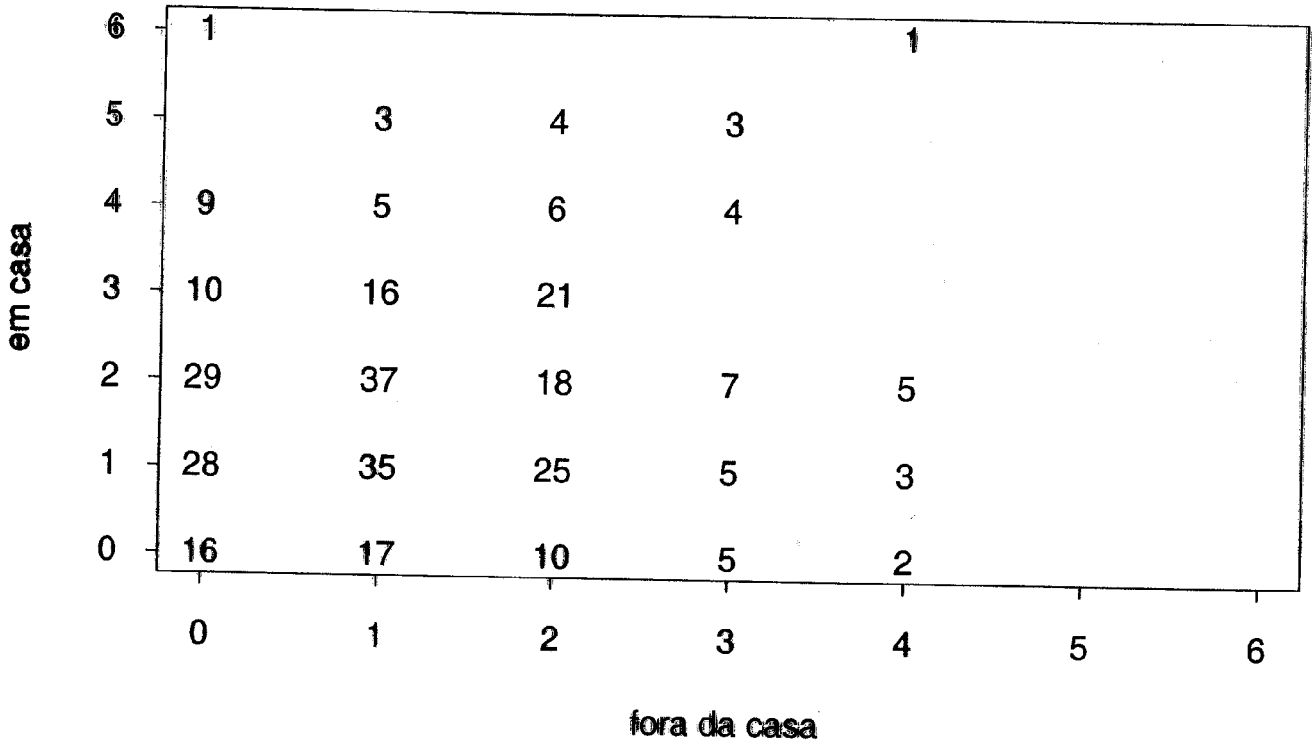
Estimador não paramétrico

Brazilian Serie A - Games played so far

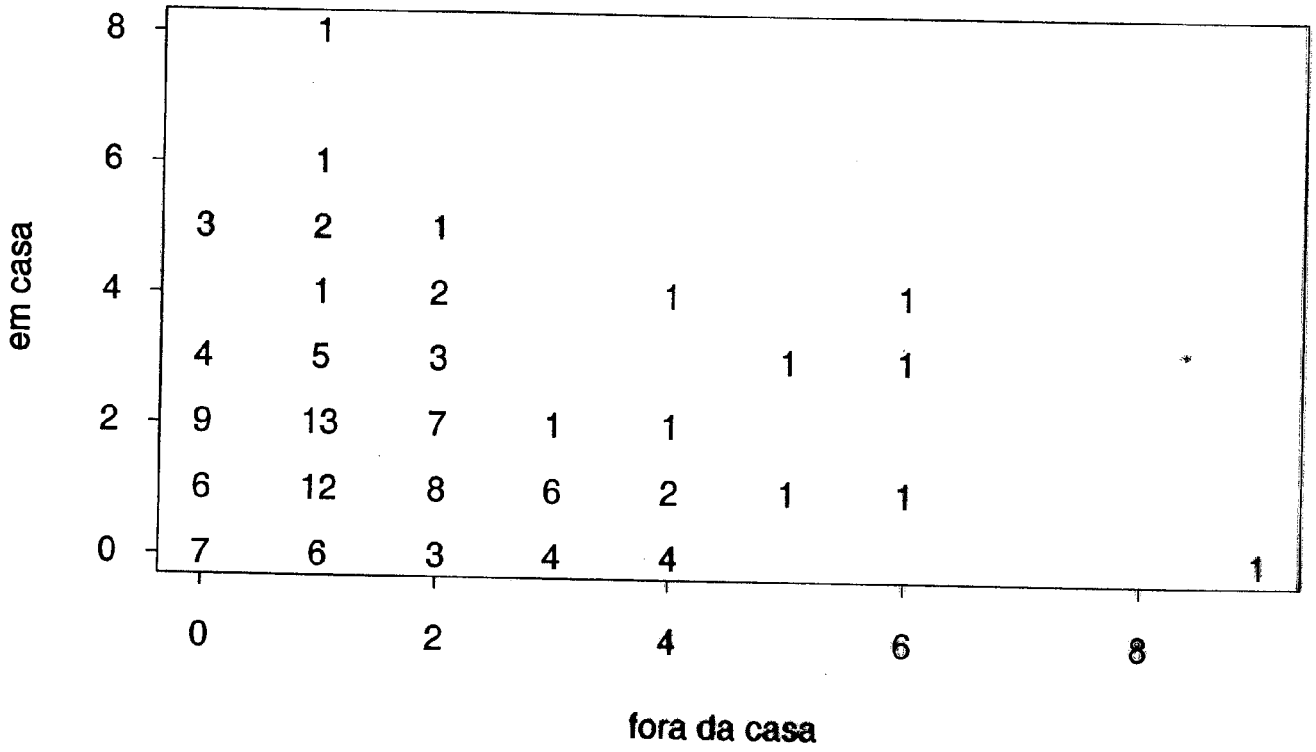
Aug 10, 2002	Vasco	- Figueirense	2 - 0
	Sao_Paulo	- Paysandu	4 - 2
	Parana	- Sao_Caetano	2 - 1
	Santos	- Botafogo-RJ	2 - 1
	Goias	- Portuguesa	3 - 1
Aug 11, 2002	Fluminense	- Cruzeiro	5 - 1
	Palmeiras	- Gremio	1 - 1
	At.Mineiro	- Corinthians	1 - 2
	Internacional	- Flamengo	1 - 3
	Guarani	- Atletico-PR	2 - 1
	Coritiba	- Vitoria	1 - 0
	Bahia	- Gama	1 - 0
	Juventude	- Ponte_Preta	1 - 0

Aug 14, 2002	Sao_Caetano	- Fluminense	2 - 0
	Botafogo-RJ	- At.Mineiro	1 - 1
	Cruzeiro	- Palmeiras	1 - 1
	Gremio	- Vasco	3 - 2

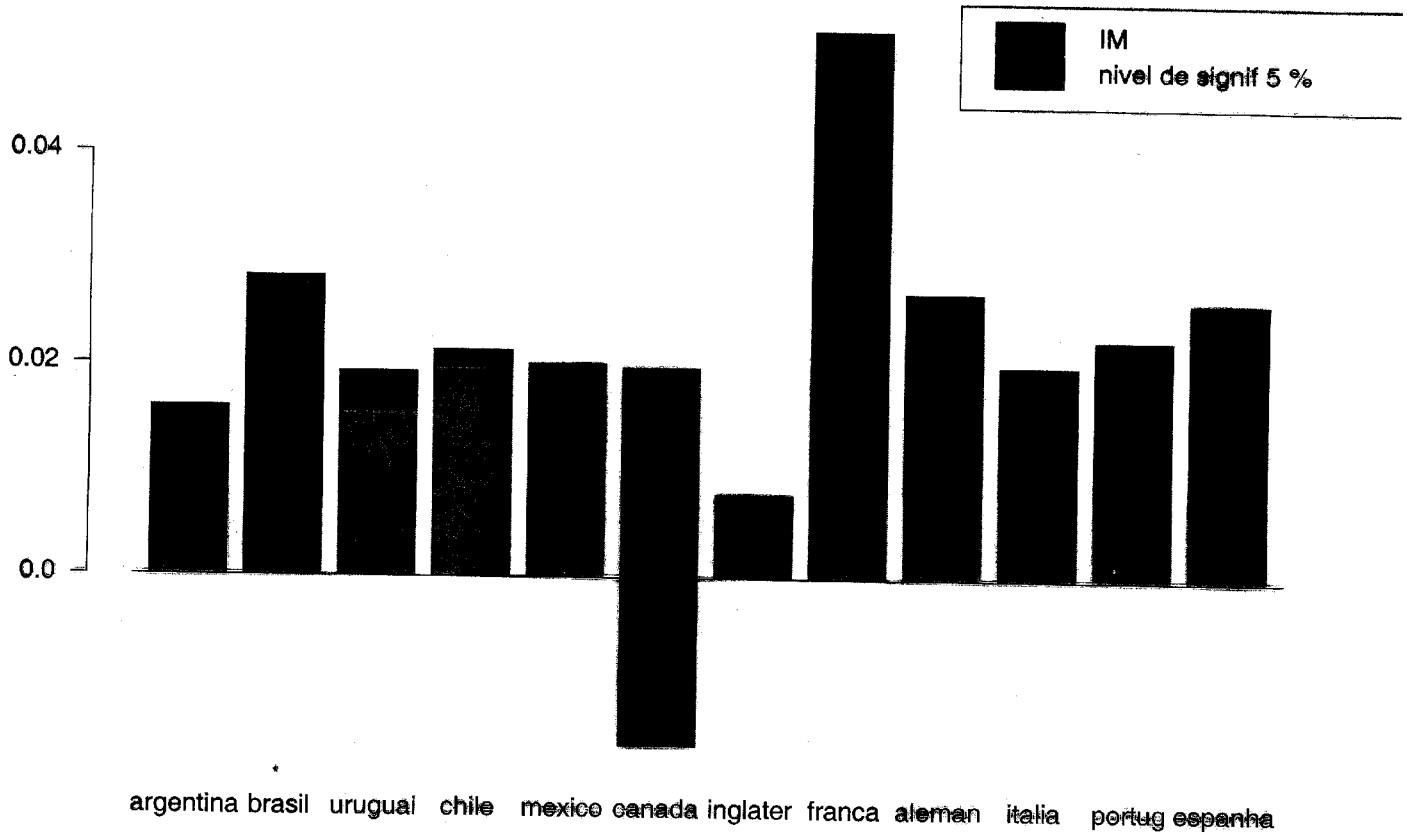
gols das equipes brasileiras

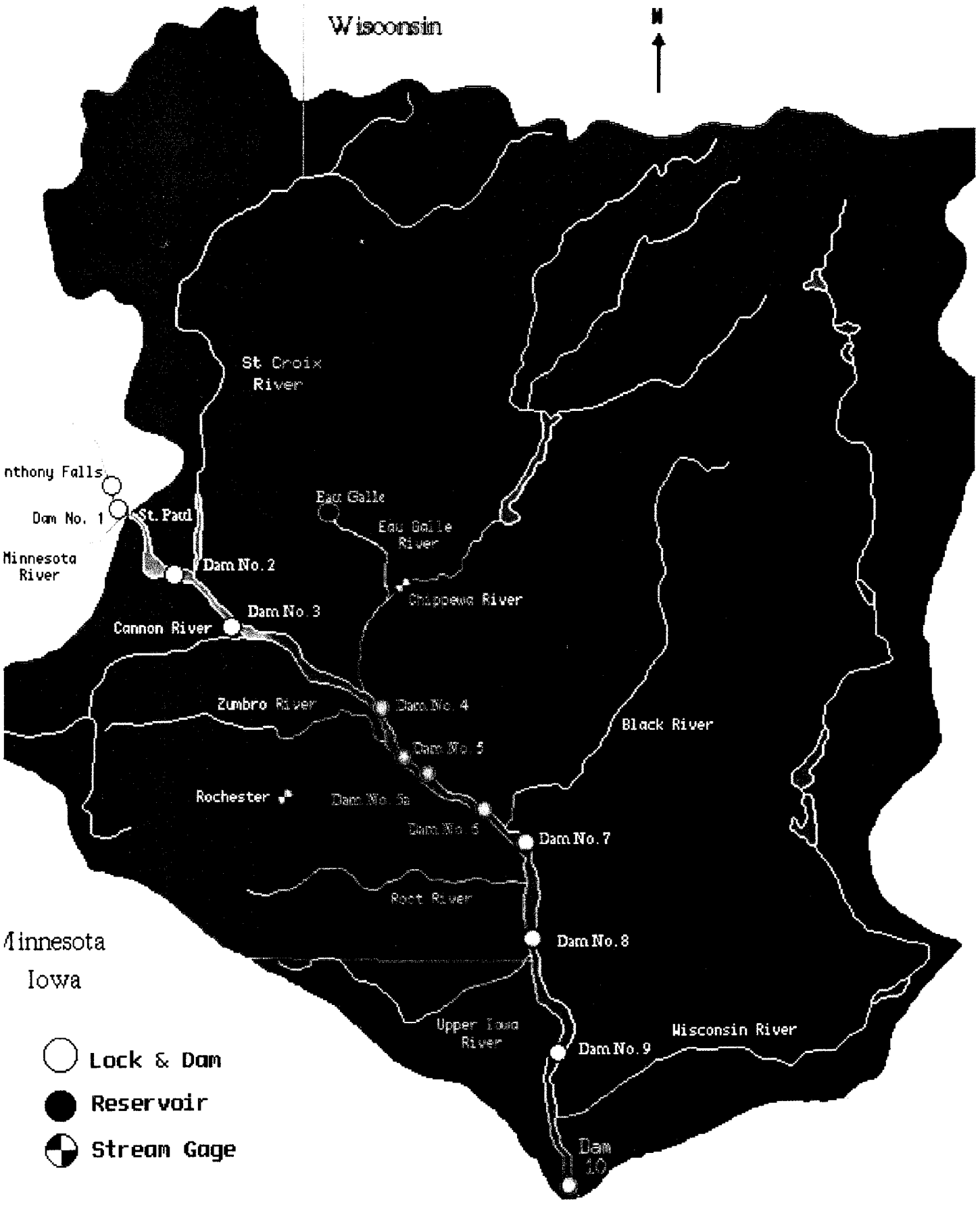


gols das equipes canadienses



IM estimando - futebol




Wisconsin
N
St Croix River
nthony Falls
Dam No. 1
St. Paul
Minnesota River
Dam No. 2
Eau Galle
Eau Galle River
Chippewa River
Dam No. 3
Cannon River
Zumbro River
Dam No. 4
Dam No. 5
Black River
Rochester
Dam No. 5a
Dam No. 6
Dam No. 7
Root River
Dam No. 8
Minnesota
Iowa
Upper Iowa River
Wisconsin River
Dam No. 9
Dam 10
Lock & Dam
Reservoir
Stream Gage

c. Caso contĭnuo-contĭnuo - Rio Mississippi.

Pergunta - Qual é a dependência entre vazões em dois pontos diferentes e distância?

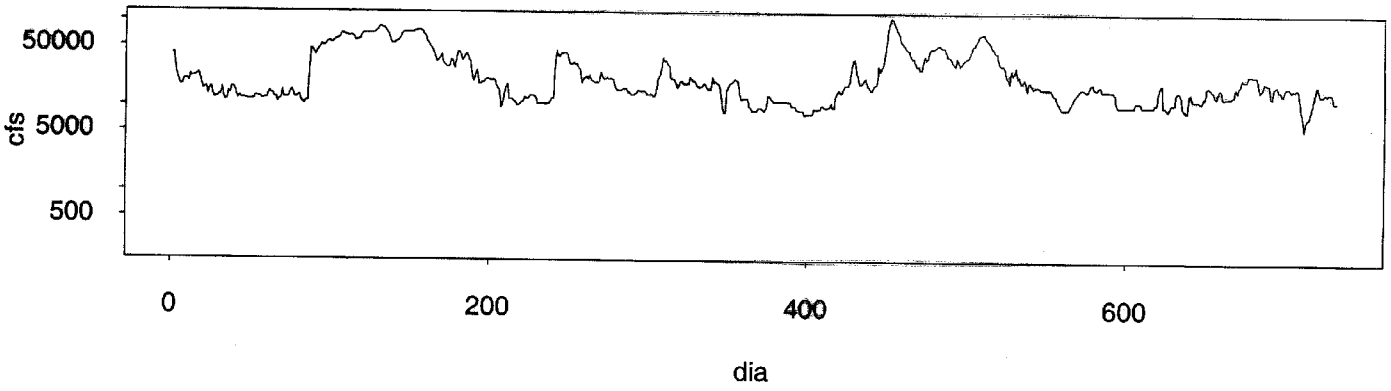
10 barragens do St. Paul, Minnesota a Iowa

Vazão diária 1 janeiro 1960 até 31 dezembro 1997, $\{Y_i(t), t=0,\pm1,\pm2,\ldots, i=1,\ldots,10\}$

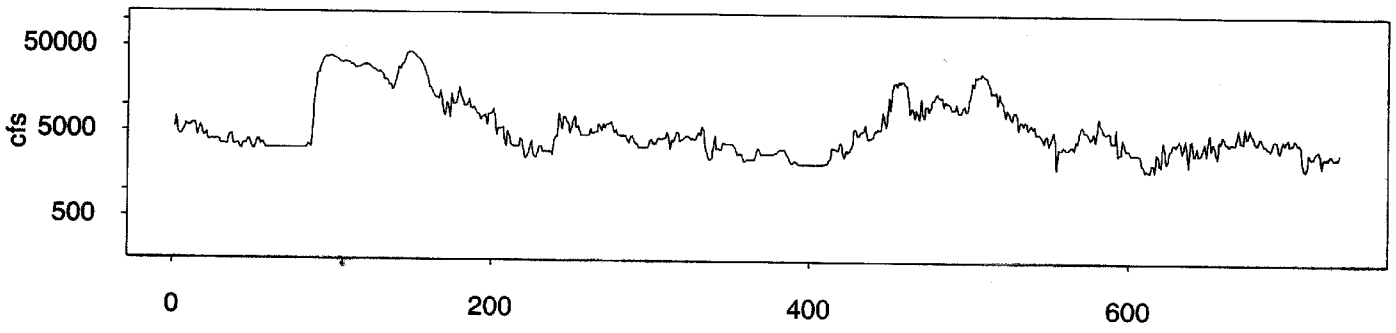
Temos as distâncias entre as barragens.

Estimador histograma

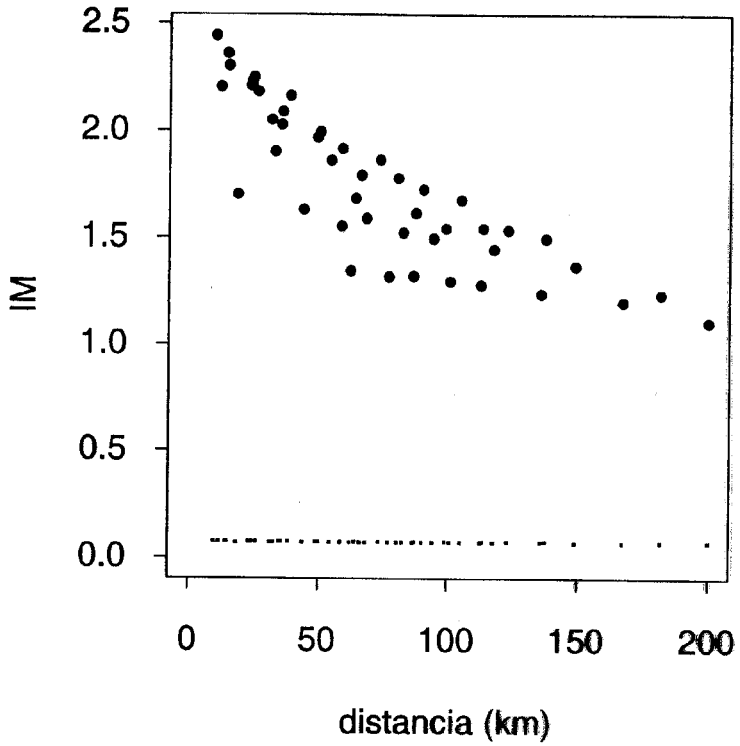
Barragem 8 vazao em 1960-1



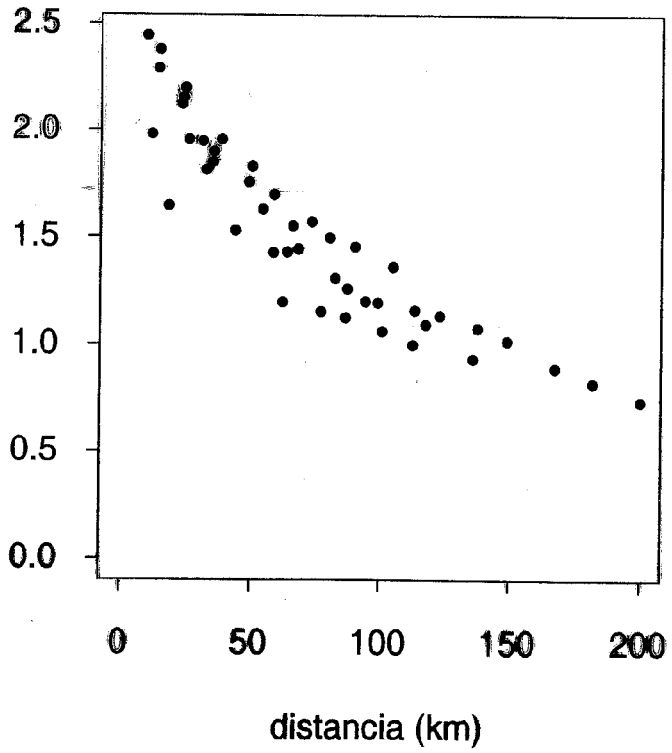
Barragem 2 vazao em 1960-1



Informacao mutua



-.5 * log(1 - rho^2)



Análise de Fourier *IM* função da freqüéncia

Séries bivariada $\{U(t), V(t)\}$

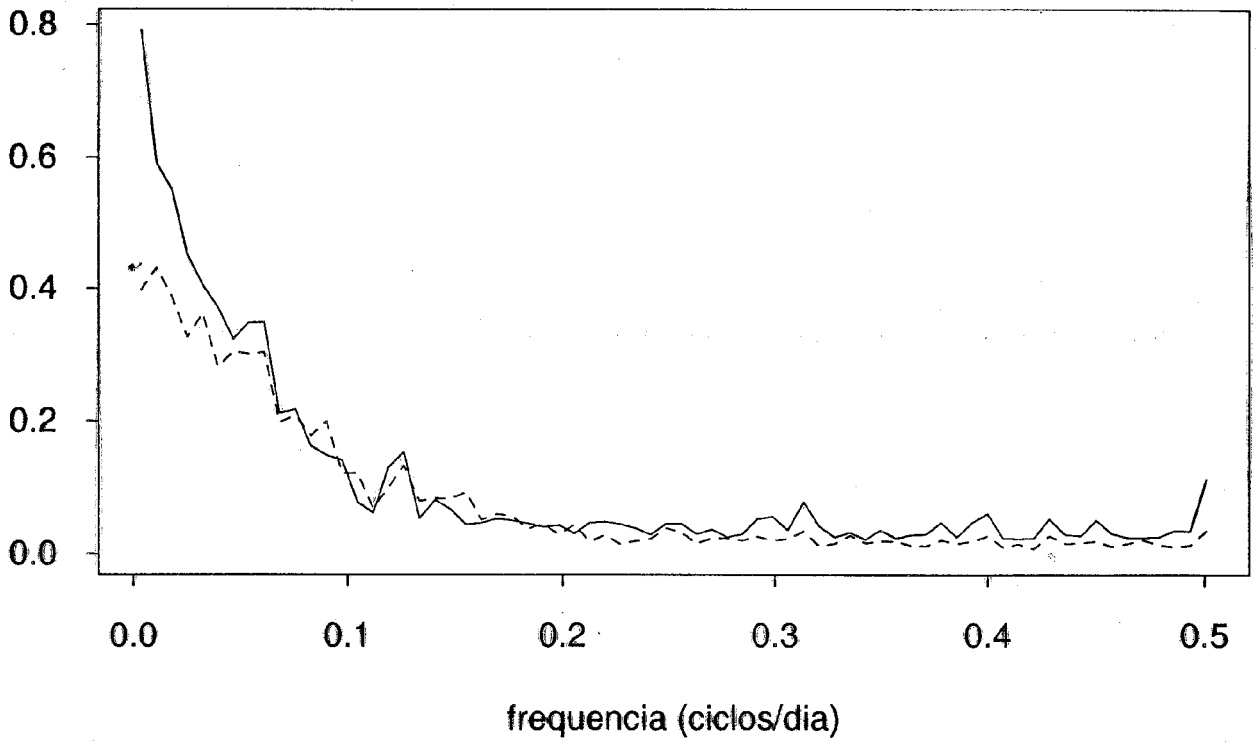
$$d_U{}^T(\lambda) = \sum_{t=0}^{T-1} e^{-i\lambda t} U(t), \quad -\infty < \lambda < \infty$$

Por λ_j proxima de λ, "dados":

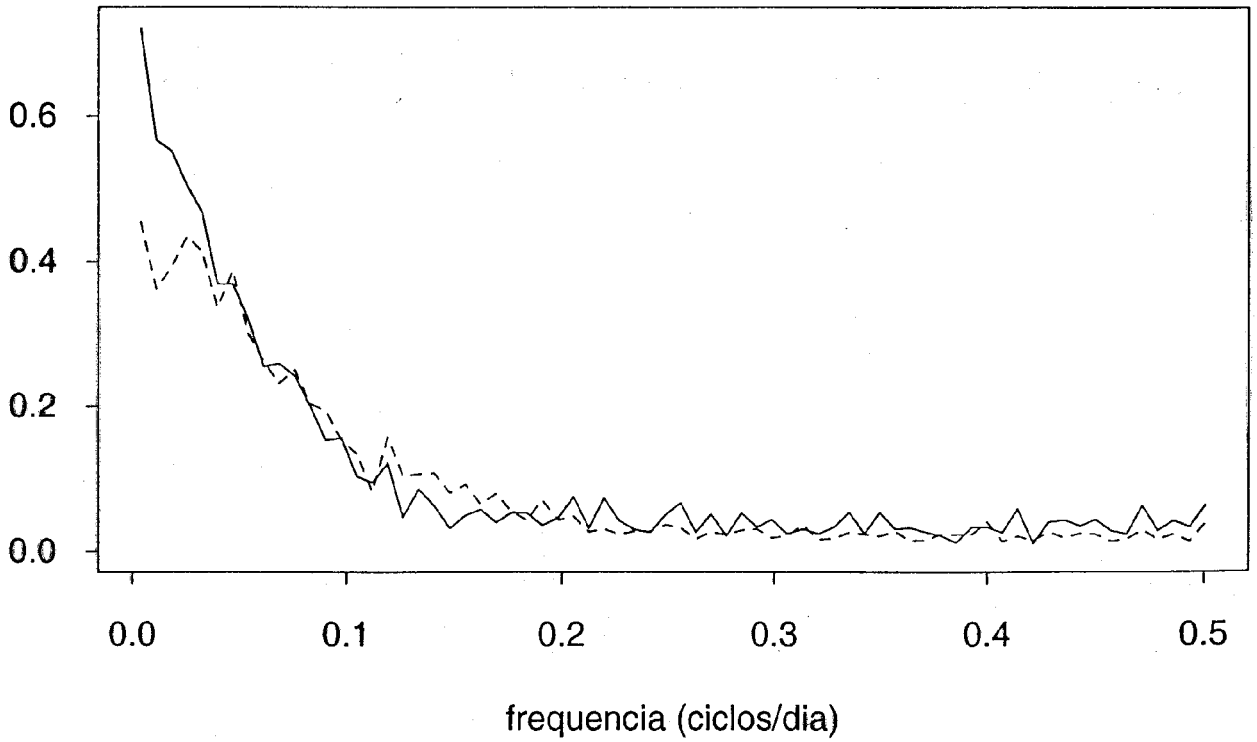
$$X_j = \text{Re } d_U{}^T(\lambda_j), \quad Y_j = \text{Re } d_V{}^T(\lambda_j)$$

Também Im

Barragens 7 e 9 - Informacao Mutua - Partes reais



Partes complexas



5. Extensões.

Outros estimadores, p. ex. tempos transcorridos

A *IM* decompõe-se, $I_{YX} = I_{YX_1} + I_{YX_2|X_1}$

Incerteza

Dificuldades com valores pequenos de p

Suavizar $\hat{p}_{jk}$?

Casos multivariados

6. Sumário.

Abordagem interessante

IM um conceito que extende a correlação

IM substitui r^2 e R^2

"A hipótese de independência é rejeitada."
torna-se

"A força de conexão é I."

As formas funcionais são úteis

Google: *corr* 2,730,000 registros, *IM* 29,800

7. Agradecimentos.

Fernando Gonçalves, Tiago Ribeiro, Daniel Sousa

Serviço Florestal dos Estados Unidos

Silvia Lopes, Ronaldo Dias