Pesquisa Científica Utilizando Ciência de Dados

Ciclo de Seminários - Instituto do Cérebro

Marcus Nunes

29 de Novembro de 2019

Departamento de Estatística - UFRN

Quem sou Eu?

Quem sou Eu?

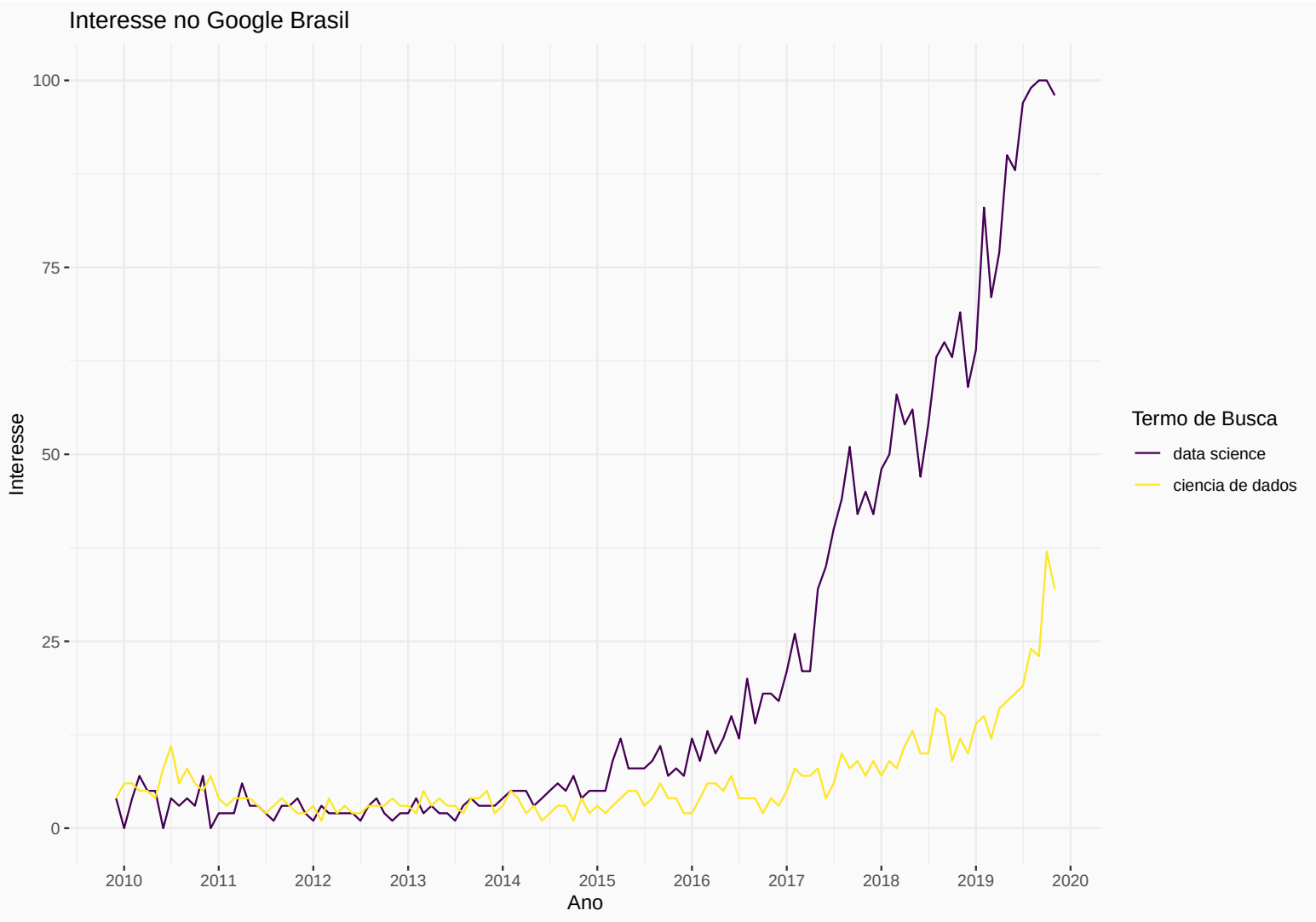
- Marcus Nunes, Professor Adjunto no Departamento de Estatística da UFRN
- PhD em Estatística pela Penn State University
- Ciência de dados, aprendizagem de máquina, aplicações da estatística, programação em r, educação estatística
- Diretor do Laboratório de Estatística Aplicada: lea.estatistica.ccet.ufrn.br/
- Site pessoal: marcusnunes.me

O Que é Ciência de Dados?

O Que é Ciência de Dados?

- *Buzzword* muito utilizada atualmente
- Juntamente com *big data* e *data science*, o termo tem ganhado muita força nos últimos anos

O Que é Ciência de Dados?



O Que é Ciência de Dados?

- Alguém tem alguma definição?

Multidisciplinaridade

- Multidisciplinaridade
- Competências de um profissional 100% capacitado para trabalhar com Ciência de Dados:
    - Estatística
    - Programação
    - Negócios
    - Conhecer bem a área de atuação (internet, varejo, finanças etc)

Multidisciplinaridade

- Que tipo de profissionais temos no momento?
  - Bons estatísticos e matemáticos que escrevem códigos sem otimização
  - Bons cientistas da computação que entendem um pouco de estatística e matemática
  - Bons cientistas da computação que entendem um pouco de negócios, depois de muita experiência na área
  - Especialistas em alguma área de atuação
  - Gerentes que sabem fazer estas pessoas trabalharem juntas

Quem Trabalha com Ciência de Dados?

- Estatísticos
- Programadores
- Físicos
- Cientistas de Dados

Quem Trabalha com Ciência de Dados?

Figure 1: Como eu me sinto

O que é um Cientista de Dados?

- Cientista de Dados (*Data Scientist*) é o novo nome para Estatístico
- No fundo, ambos são a mesma coisa, embora uma destas profissões trabalhe melhor seu marketing pessoal
- Para mim, é alguém que entende mais de programação do que um Estatístico tradicional
- Também entende mais de estatística do que um Cientista da Computação tradicional
- E, principalmente, é alguém que consegue encontrar soluções para problemas juntando estas duas áreas do conhecimento com multidisciplinaridade

Aplicações

Psicolinguística

- Godoy et al., (2017). O papel do conhecimento de eventos no processamento de sentenças isoladas. Letrônica, 10 (2), pp 538-554.
- O conhecimento de eventos faz parte de uma coleção de pistas pragmáticas que impactam o processo de compreensão da linguagem
- Experimento de leitura autocadenciada

Psicolinguística

- O jornalista checou a *ortografia* do seu último relatório. (Argumento previsível)
- O mecânico checou os *freios* do carro. (Argumento previsível)
- O jornalista checou os *freios* do carro. (Argumento imprevisível)
- O mecânico checou a *ortografia* do seu último relatório. (Argumento imprevisível)

Psicolinguística

$$Y_{ijklmn} = \mu + I_i + E_j + (IE)_{ij} + S_k + L_l + P_m(l) + \varepsilon_{ijklmn}$$

- Y_{ijklmn}: tempo de resposta (ms)
- μ : média geral
- I_i: argumento interno
- E_j: argumento externo
- $(IE)_{ij}$: interação entre os argumentos
- $S_k \sim N(0, \sigma_S^2)$: sujeito
- $L_l \sim N(0, \sigma_L^2)$: lista de palavras
- $P_m(l) \sim N(0, \sigma_P^2)$: palavra m dentro da lista l
- $\varepsilon_{ijklmn} \sim N(0, \sigma_\varepsilon^2)$: erro aleatório

Psicolinguística

- 4 listas, 32 itens experimentais, 24 sujeitos
- Foi ajustado um modelo de regressão linear misto
- Não foram detectados efeitos dos argumentos

Captura de Peixes

- Lima et al. (2020). Declining fisheries and increasing prices: The economic cost of tropical rivers impoundment. Fisheries Research, 221.
- Com a construção de barragens no leito do rio, a reprodução dos peixes ficou comprometida
- A pesca diminuiu 58% em 25 anos, enquanto o preço aumentou 49% durante o mesmo período

Captura de Peixes

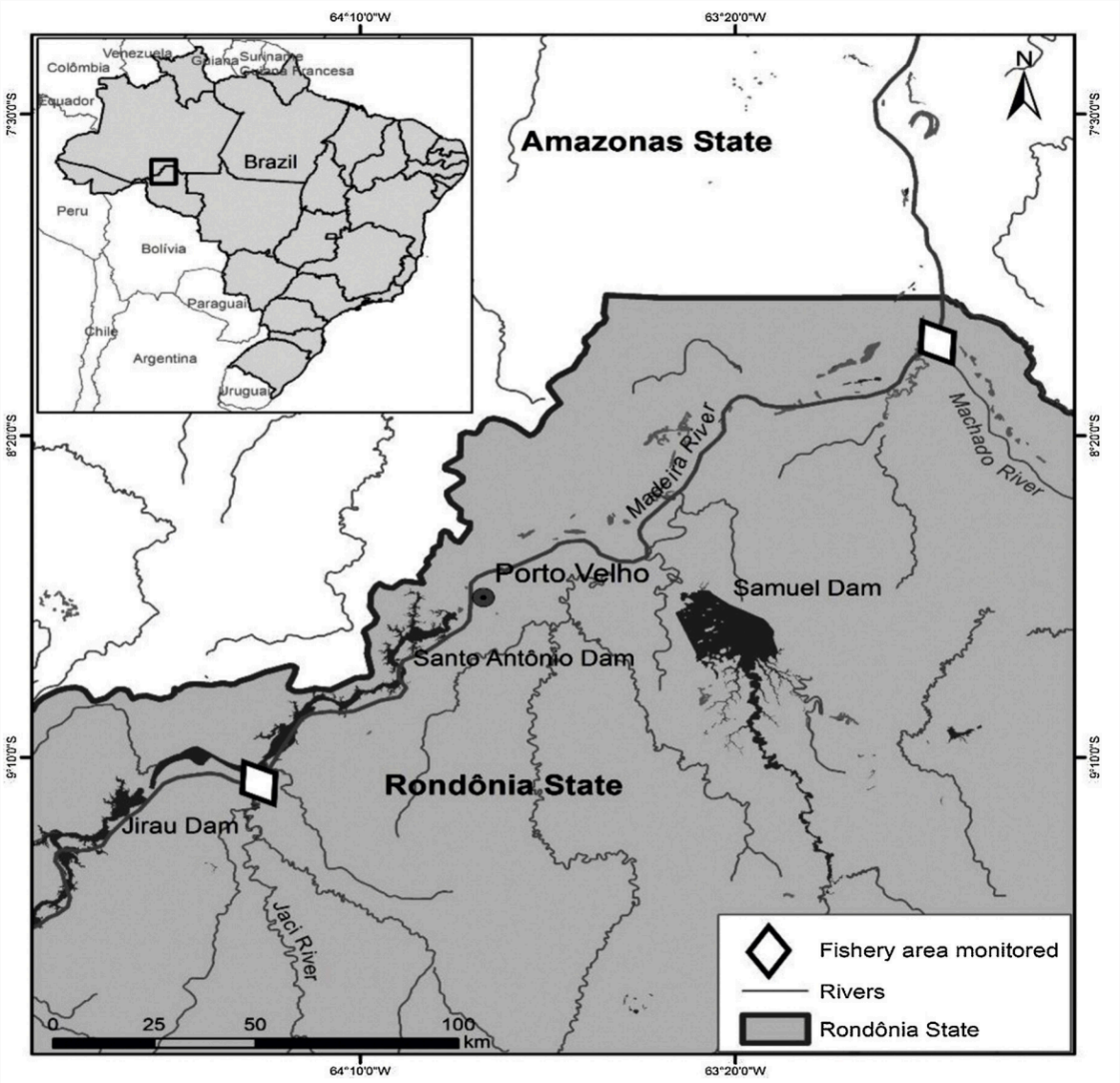
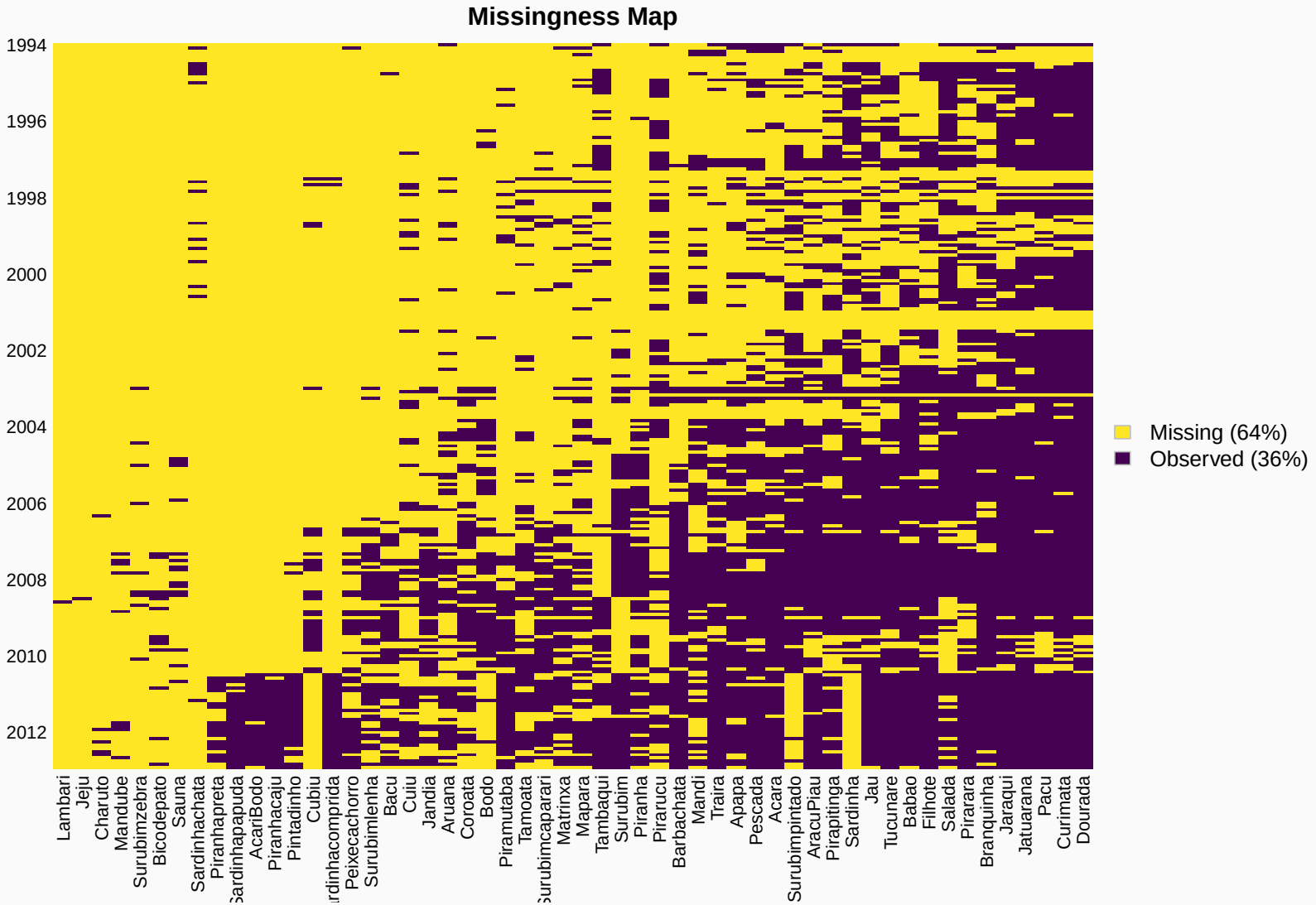
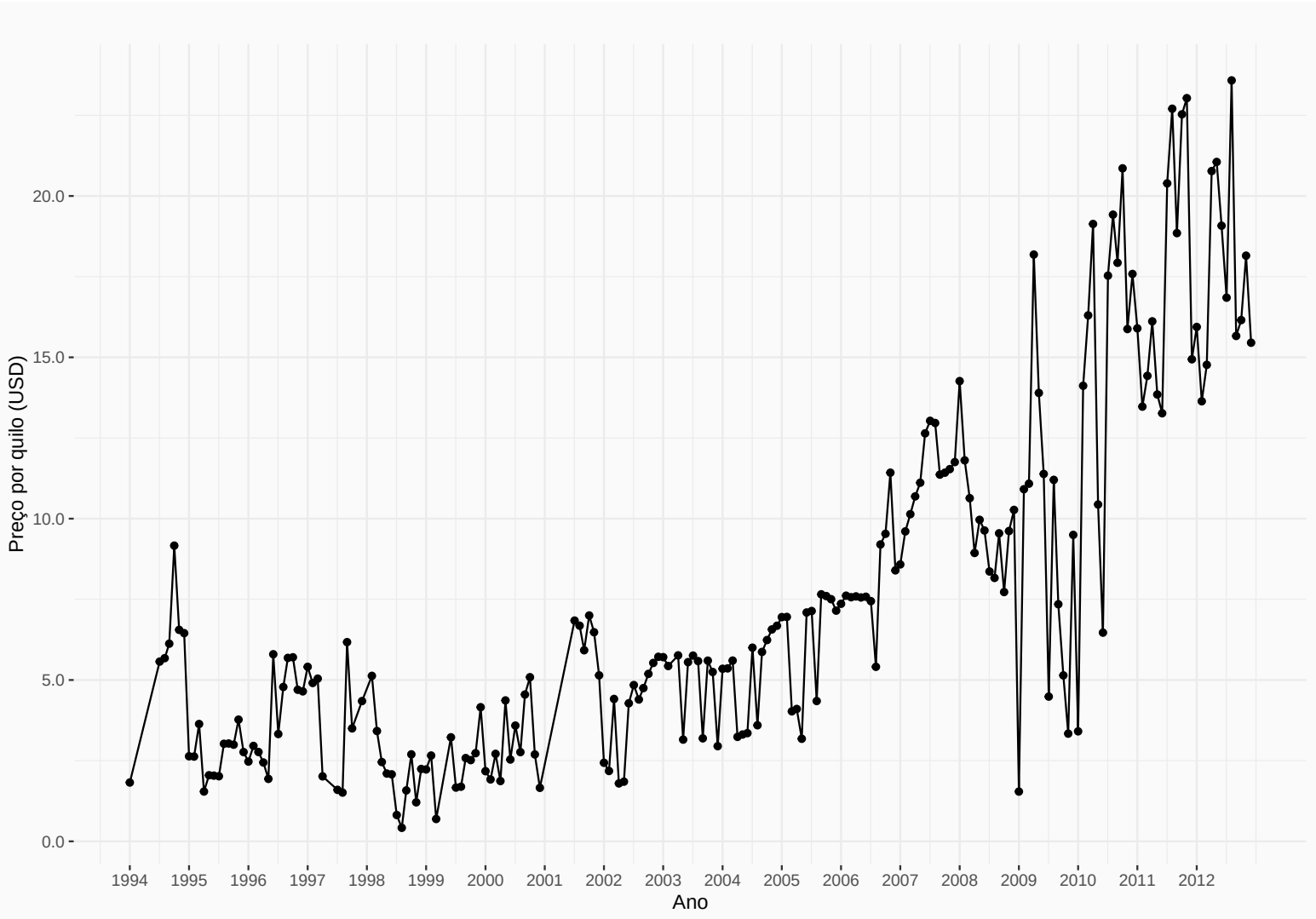


Figure 2: Rio Madeira e suas represas

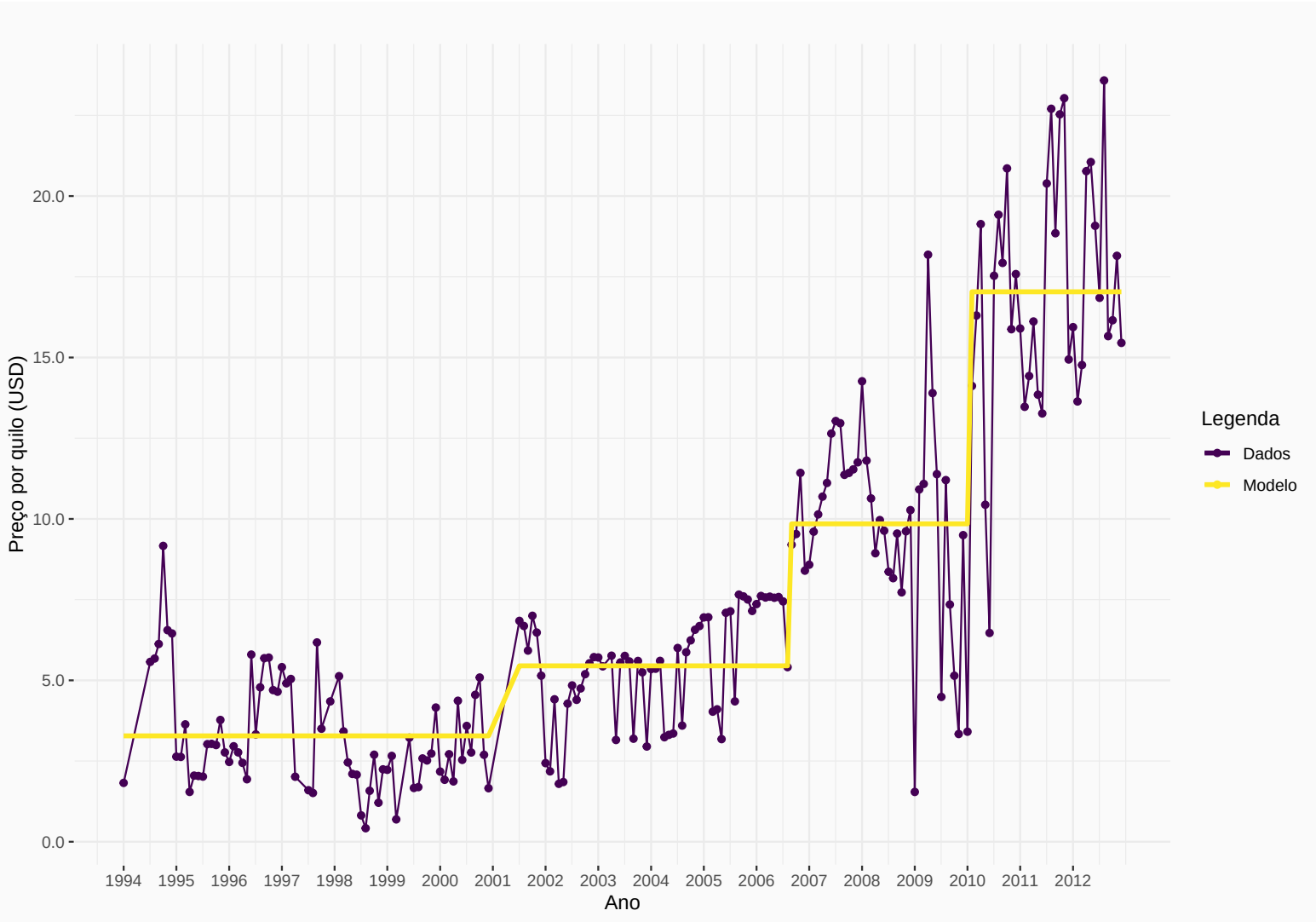
Captura de Peixes



Captura de Peixes



Captura de Peixes



Cienciometria

- **Cienciometria:** ciência que, a partir de aspectos quantitativos e qualitativos da publicação científica, busca o entendimento dos múltiplos fatores que afetam a produção acadêmica e consequentemente determinam a qualidade de cientistas, periódicos e instituições.
- **Objetivo:** analisar a produção científica da UFRN no período de 2014-2017 e encontrar quais as variáveis (*drivers*) que determinam esta produção. Auxiliar na formulação de estratégias de gerenciamento e financiamento.

Cienciometria

- Angelini e Nunes (202?)
- Quais fatores influenciam a produção acadêmica dos professores da UFRN?
- Analisamos dados entre 2014 e 2017

Cienciometria - Produção Total

Média	Desvio Padrão
6.92	10.23

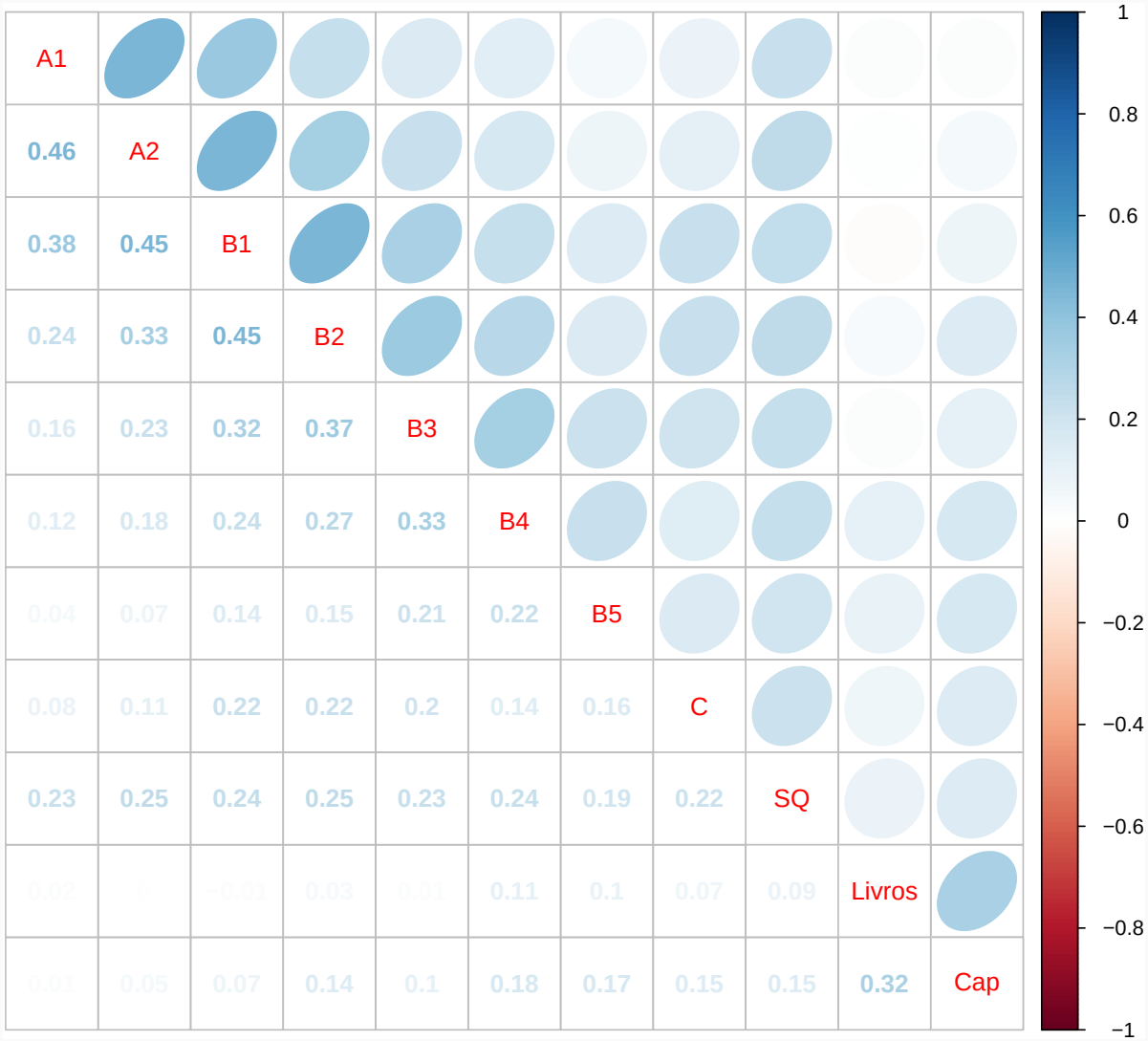
# artigos	Qtde	%
0	327	18.31
1	227	12.71
2	187	10.47
3	168	9.41
4	126	7.05
5	92	5.15
6	81	4.54
7	59	3.30
8	64	3.58
9	38	2.13
10+	417	23.35

Cienciometria - Produção A1

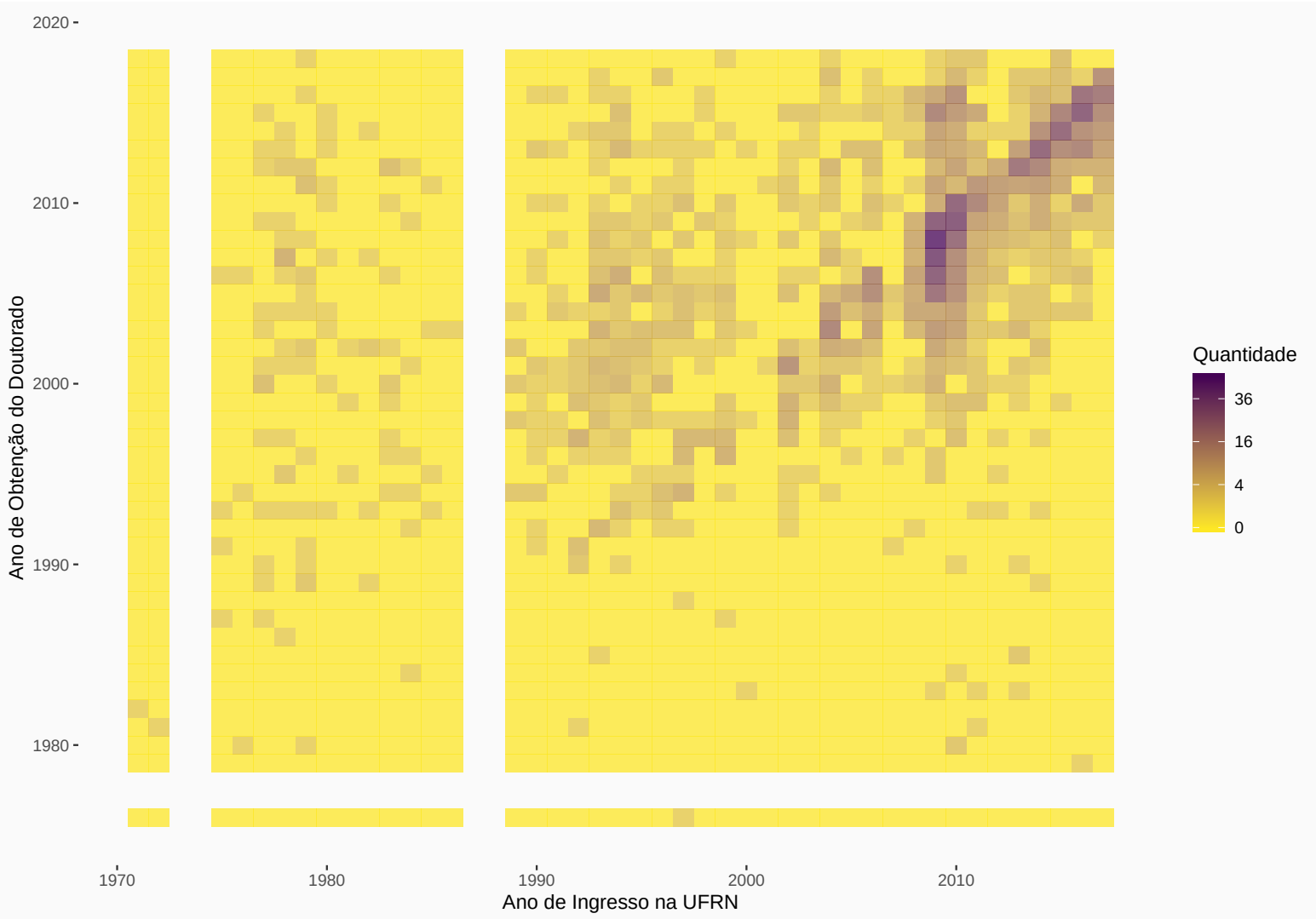
Média	Desvio Padrão
0.7	1.9

# artigos	Qtde	%
0	1311	73.40
1	239	13.38
2	84	4.70
3	53	2.97
4	25	1.40
5	28	1.57
6	13	0.73
7	7	0.39
8	5	0.28
9	4	0.22
10+	17	0.95

Cienciometria - Correlações entre os Tipos de Produção



Cienciometria - Heatmap



Cienciometria - Variáveis Utilizadas

- UNG: horas de ensino anuais na graduação
- GRA: horas de ensino anuais na pós-graduação
- SDOC: orientações de doutorado concluídas
- SMAS: orientações de mestrado concluídas
- SIC: orientações de iniciação científica concluídas
- SMON: número de TCCs orientados
- PFU: número de projetos financiados
- PNF: horas dedicadas à pesquisa
- OUT: número de eventos
- BEF: número de artigos publicados antes de obter o doutorado
- YDOC: ano de obtenção do doutorado
- YUFRN: ano de ingresso na UFRN
- GEN: gênero

Cienciometria - Modelagem

- Dados discretos, provenientes de contagens
- O natural seria escolher distribuições como Poisson ou Binomial Negativa para o ajuste do modelo
- Entretanto, a análiase exploratória nos sugere um excesso de zeros
- Isto nos leva a considerar o ajuste de um modelo hurdle

- Sendo assim, testamos quatro modelos diferentes:

1. Regressão Poisson: $\mathrm{E}(Y) = \mu; \mathrm{Var}(Y) = \mu$
2. Regressão Binomial Negativa: $\mathrm{E}(Y) = \mu; \mathrm{Var}(Y) = \mu + \phi\mu^2$
3. Regressão Poisson Hurdle:
   $\mathrm{E}(Y) = \frac{1-p}{1-e^{-\mu}}\mu; \mathrm{Var}(Y) = \frac{1-p}{1-e^{-\mu}}(\mu + \mu^2) - \left(\frac{1-p}{1-e^{-\mu}}\mu\right)^2$
4. Regressão Binomial Negativa Hurdle:
   $\mathrm{E}(Y) = \frac{1-p}{1-P_0}\mu; \mathrm{Var}(Y) = \frac{1-p}{1-P_0}\left(\mu^2 + \mu + \frac{\mu^2}{k}\right) - \left(\frac{1-p}{1-P_0}\mu\right)^2$

em que p é a probabilidade de uma observação igual a zero ocorrer e $P_0 = \left(\frac{k}{\mu+k}\right)^k$

Cienciometria - Modelagem

	AIC	BIC
ajuste_negbin	9580.411	9662.727
ajuste_poisson	14508.942	14585.770
ajuste_hurdle_poisson	13234.043	13387.699
ajuste_hurdle_negbin	9463.204	9622.348

Cienciometria - Modelagem

Preditores	Coeficientes	Média	Desvio.Padrão	p.valor
(Intercept)	5.0415	NA	NA	0.0000
UNG	1.0013	13.9999	6.3341	0.9600
GRA	1.0763	3.4784	4.5456	0.0207
SDOC	1.3162	0.6988	1.7804	0.0000
SMAS	1.2484	2.2447	3.1184	0.0000
SIC	1.1551	2.0666	3.6978	0.0000
SMON	1.0558	3.3645	5.9741	0.0186
PFU	1.1478	3.6534	5.6673	0.0000
PNF	1.0776	21.7013	52.0439	0.0110
OUT	1.0114	5.8018	8.7833	0.6360
BEF	1.4226	3.2156	5.7848	0.0000
YDOC	0.9108	2007.2374	6.7722	0.0083
YUFRN	1.1558	2005.8970	9.7173	0.0000
GENMasculino	1.0695	NA	NA	0.1806

Obrigado