Visualização de dados: a autonomia do usuário na construção de gráficos

Paulo George Miranda Martins[a], *Danila Fernandes Alencar*[b] *e Ricardo César Gonçalves Sant'Ana*[c]

1. Introdução

O volume de dados gerados nas últimas décadas têm proporcionado mudanças significativas na forma de como os dados são recuperados e utilizados por usuários de sistemas informatizados, o que tem se tornado um desafio cada vez maior no mundo contemporâneo frente a abundância de dados e informações que, em sua maioria, não estão estruturados adequadamente na internet, necessitando de tratamento especializado.

Considerando tal cenário, a informação desempenha um papel fundamental no desenvolvimento humano, seja no compartilhamento de informações entre os diversos grupos sociais, seja no uso estratégico pelas organizações e/ou unidades de informação. O acesso a dados públicos constitui-se um fator

a Doutorando em Ciência da Informação pela Universidade Estadual Paulista – UNESP. Inspetor Sênior do Controle da Qualidade na LATAM Airlines Brasil. E-mail: paulo.george@unesp.br. ORCID: https://orcid.org/0000-0002-3922-5069.

b Mestranda em Ciência da Informação pela Universidade Estadual Paulista – UNESP. Arquivista no Centro de Documentação da Universidade de Marília – UNIMAR. E-mail: danila.fernandes@unesp.br.

c Doutor em Ciência da Informação pela Universidade Estadual Paulista – UNESP. Professor na Universidade Estadual Paulista – UNESP. E-mail: ricardo.santana@unesp.br. ORCID: https://orcid.org/0000-0003-1387-4519.

fundamental para a geração de conhecimento e, em um contexto democrático de ações e informações transparentes, o "[…] acesso a dados e as muitas possibilidades de construção de visualização para estes dados", constitui-se um elemento fundamental para a "[…] construção de uma participação cidadã e para a concretização de uma relação mais transparente", entre os mais diversos órgãos governamentais e a sociedade civil (SANT'ANA; RODRIGUES, 2013, p. 58).

Para o *Open Government Working Group* (OGD), grupo de estudo multidisciplinar formado em 2007 na cidade americana de Sebastopol, o acesso aos dados públicos de forma transparente e livre parte do pressuposto de que tais informações devem estar em formato aberto. Para nortear o uso dos dados abertos, o grupo estabeleceu um conjunto de 8 Princípios na qual os dados devem ser: completos, primários, atuais, acessíveis, processáveis por máquinas, de acesso não-discriminatório, de formato não-proprietário e com licenças livres (ODG, 2007).

Nesse contexto, a Ciência da Informação desempenha papel fundamental na perspectiva de acessibilidade informacional ao proporcionar melhores práticas nos processos de recuperação de recursos e na instrumentalização através da visualização de informações que podem ser utilizadas para a prática decisória nas instituições informacionais, empresariais, indivíduos e/ou para o próprio processo de retroalimentação de coleta informacional.

O processo de visualização dos dados deve ser calcado em métodos estatísticos como o modelo proposto por Card,

Mackinlay e Shneiderman (1999), que ilustra a sequência de transformação de dados brutos até sua forma visual, passando por fases de tratamento e transformações de dados e o mapeamento visual necessário para apresentação gráfica.

Nessa perspectiva e considerando o cenário de múltiplas formas em se obter dados, a presente pesquisa apresenta a seguinte questão: quais são os pré-requisitos necessários para ampliação da autonomia dos usuários na construção de visualizações gráficas de dados?

O objetivo central da pesquisa é identificar barreiras na construção de visualização gráfica de dados a partir da utilização do software R, e para tanto, os objetivos específicos são: a) selecionar um conjunto de dados disponibilizados em formato que permita reutilização, b) construção de uma nova visualização gráfica dos dados e c) identificar pré-requisitos necessários para a construção do gráfico.

Esse estudo pretende trazer elementos para uma reflexão sobre a autonomia do usuário na construção e visualização gráfica de dados, além de subsídios para o entendimento sobre a ferramenta.

2. Procedimentos metodológicos

Por se tratar de uma análise a partir da recuperação de dados, a metodologia adotada para o desenvolvimento da pesquisa se caracteriza como exploratória, através da coleta de conjunto de dados disponíveis no sítio Terrabrasilis a partir dos resultados obtidos nos termos de busca "Desmatamento" e "Cerrado", no

mecanismo de busca, tendo como ponto específico as premissas de disseminação da informação, destacadas no Ciclo de Vida dos Dados no âmbito da CI (SANT'ANA, 2016).

Caracteriza-se, também, como pesquisa de natureza aplicada uma vez que objetiva gerar "conhecimentos para aplicações práticas" (SILVEIRA & CÓRDOVA, 2009, p. 35) e busca "[…] compreender e analisar a realidade dos enfoques da pesquisa" (TREVIÑOS, 1987, p. 117).

Sant'Ana e Rodrigues (2013) pontuam que, o acesso aos dados primários favorece a construção de uma visualização mais próxima possível dos valores originais representados, evitando falhas de interpretação e permitindo caminhos variados no processo de análise, dada a liberdade de combinações que um dado primário oferece.

Para este estudo, realizou-se a coleta de dados sobre o desmatamento no Brasil a partir da fonte gráfica denominada "Incremento de desmatamento – Cerrado – Estados", disponível no portal TerraBrasilis[5], uma plataforma de disseminação de dados de desmatamento do INPE.

Para construção do arcabouço teórico foram pesquisadas fontes de informação primárias e secundárias, como artigos, livros e de estudos apresentados, além de pesquisas em bases de dados especializadas na temática abordada.

5 TerraBrasilis. Disponível em: http://terrabrasilis.dpi.inpe.br/. Acesso em: 10 out. 2021.

3. Coleta de dados e o processo de visualização

A coleta dos dados foi realizada no sítio TerraBrasilis, cuja disposição atende os princípios de acessibilidade a dados públicos de forma transparente e livre, estabelecidos pela OGD (2007) de que os dados devem ser acessíveis e processáveis por máquinas.

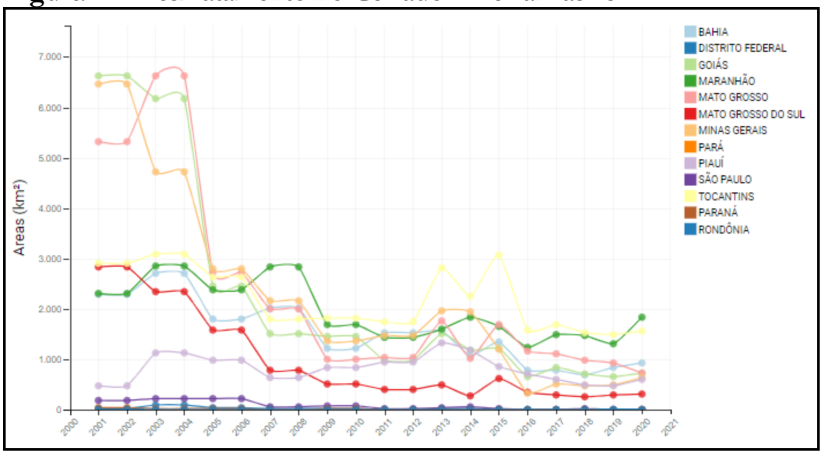
Os dados coletados foram armazenados em uma planilha de cálculo, para gerar um arquivo no formato ".xlsx" estruturado para posterior importação para o software R.

Sant'Ana (2016) destaca que algumas etapas predefinidas são essenciais para o tratamento pleno dos dados, que devem obedecer a um ciclo de vida estruturado para que gerem valor e atendam as expectativas de seus usuários, possibilitando desenvolver melhores formas de identificar, planejar e implementar serviços e produtos alinhados com as novas perspectivas tecnológicas contemporâneas.

Para a pesquisa foi selecionado o gráfico referente a linha do tempo de desmatamento e os dados correspondem aos nomes dos estados, áreas desmatadas em km² e ao ano referente aos dados levantados, como apresentado na Figura 1.

No intuito de propiciar condições para que o usuário possa ter liberdade de solucionar eventuais questões sobre os dados, percebe-se a necessidade de empoderá-los com ferramentas para que ele possa usufruir dos potenciais dados disponibilizados em expressiva quantidade na web.

Figura 1 – Desmatamento no Cerrado – TerraBrasilis



Fonte: TerraBrasilis PRODES (2021).

Para Wickham e Grolemun (2016), um processo de visualização eficiente mostrará coisas que o leitor não esperava ou fomentar a elaboração de novas questões sobre os dados, podendo, inclusive, indicar que o direcionamento a ser adotado está correto ou não, ou se é preciso coletar dados diferentes para consolidar as informações já processadas.

Os dados por si só não são o suficiente para representar algum contexto, sendo necessário, portanto, intervenções tecnológicas e humanas para dar sentido aos mesmos, transformando-os em informações relevantes que façam sentido e gerem valor para aqueles que os detém.

Para o escopo dessa pesquisa, será utilizado o software R. Considerado um sistema para computação estatística e geração de gráficos, criada em 1993 por Ross Ihaka e Robert Gentleman, o

software R fornece, entre outras coisas, uma linguagem de programação, gráficos de alto nível, interfaces para outras linguagens e recursos de depuração (FARIA; PARGA, 2020).

4. Resultados e discussões

A partir do gráfico apresentado na Figura 1, buscou-se construir um gráfico similar, em linhas, e outro, com pequenas alterações no conjunto de dados, em barras. Dado o escopo deste trabalho, busca-se demonstrar qual nível de autonomia o usuário, mesmo com conhecimento mínimo sobre a ferramenta, possa ter para criar visualizações gráficas.

Os dados coletados foram armazenados em uma planilha de cálculo, gerando um arquivo ".xlsx" de valores dispostos em linhas e colunas deforma estruturada para que fosse possível a leitura no software R. O arquivo ".xlsx" foi usado para gerar os gráficos que correspondem, em síntese, ao gráfico da Figura 1 disponível no sítio TerraBrasilis. Neste exemplo, foram selecionados nove Estados que apresentam índices que compreendem os anos de 2001 até 2020.

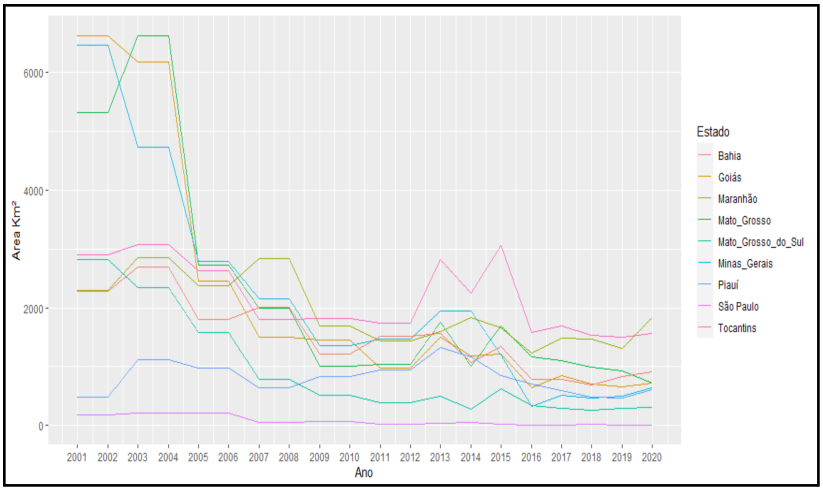
Foram utilizadas duas bibliotecas para importar arquivos e criar a visualização de dados, sendo: "*library(readxl)*", biblioteca necessária para leitura de arquivos no formato Excel; e "*library(ggplot2)*", usados para criação de visualizações.

O pacote "ggplot2" oferece uma ampla variedade de construção gráfica a partir de uma sintaxe compacta, possibilitando criar gráficos com uma melhor visualização, sendo

considerado especialmente útil para iniciantes na ferramenta R (WICKHAM, 2016).

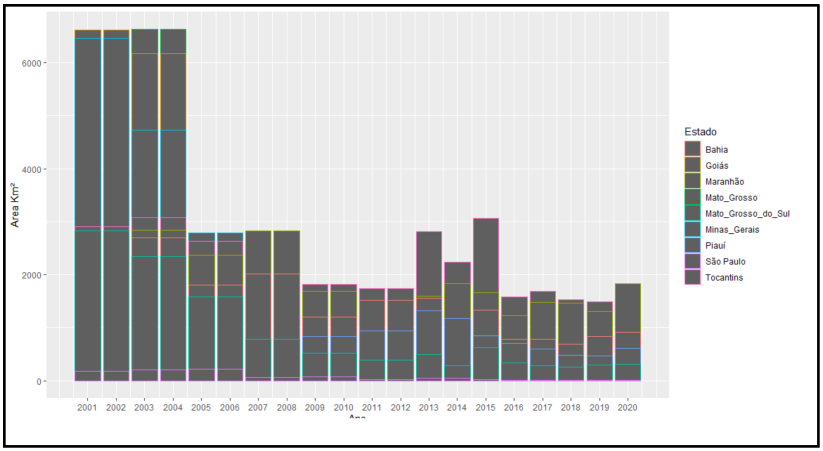
Após importar o arquivo ".xlsx" para o R, foram elaborados dois gráficos distintos, sendo um de linhas (Figura 2), e outro de barras verticais (Figura 3), com a finalidade de demonstrar que com as mesmas funções, fazendo-se pequenas alterações nas linhas de comando é possível gerar outras formas de visualizações.

Figura 2 – Gráfico desenvolvido no software R para ilustrar o desmatamento no Cerrado



Fonte: Autores.

Figura 3 – Gráfico desenvolvido no software R para ilustrar o desmatamento no Cerrado



Fonte: Autores.

O código utilizado para a elaboração dos gráficos foi construído em linhas de comando repetidas 2 vezes com parâmetros diferentes para carregar as bibliotecas de funções (*library*).

Para execução do gráfico existem apenas duas linhas de comando: uma para carga do conteúdo da planilha para a variável de memória e outra para execução do gráfico. A primeira com 1 parâmetro e a segunda com 9 parâmetros, sendo que um deles (*geom_line*) foi necessário mais de uma vez (Figura 4).

Figura 4 – Código utilizado para construir o gráfico da Figura 2.

```
library(readxl)
library(ggplot2)
dados <- read_excel("arquivo.xlsx")
ggplot(dados)+aes(x=2001:2020,y=Goias)+
  geom_line(aes(y = Mato_Grosso,col="Mato_Grosso")) +
  geom_line(aes(y = Goias,col="Goiás")) +
  geom_line(aes(y = Minas_Gerais,col="Minas_Gerais")) +
  geom_line(aes(y = Tocantins,col="Tocantins")) +
  geom_line(aes(y = Maranhao,col="Maranhão")) +
  geom_line(aes(y = Bahia,col="Bahia")) +
  geom_line(aes(y = Piaui,col="Piauí")) +
  geom_line(aes(y = Mato_Grosso_do_Sul,col="Mato_Grosso_do_Sul")) +
  geom_line(aes(y = Sao_Paulo,col="São Paulo")) +
  scale_x_continuous("Ano", labels = as.character(2001:2020), breaks = 2001:2020)+
  labs(x = "Ano",y = "Area Km²",color="Estado")
```

Fonte: Autores.

Para a construção de um gráfico alternativo (Figura 3), foram alterados os parâmetros da segunda linha de comando, modificando o parâmetro "*geom_line*", no gráfico em linhas, para "*geom_col*", no gráfico de colunas, demonstrando-se, nesse caso, a flexibilidade em se obter diferentes visualizações, substituindo apenas um parâmetro, conforme Figura 5.

Figura 5 – Código utilizado para construir o gráfico da Figura 3

```
library(readxl)
library(ggplot2)
dados <- read_excel("arquivo.xlsx")
ggplot(dados)+aes(x=2001:2020,y=Goias)+
  geom_col(aes(y = Mato_Grosso,col="Mato_Grosso")) +
  geom_col(aes(y = Goias,col="Goiás")) +
  geom_col(aes(y = Minas_Gerais,col="Minas_Gerais")) +
  geom_col(aes(y = Tocantins,col="Tocantins")) +
  geom_col(aes(y = Maranhao,col="Maranhão")) +
  geom_col(aes(y = Bahia,col="Bahia")) +
  geom_col(aes(y = Piaui,col="Piauí")) +
  geom_col(aes(y = Mato_Grosso_do_Sul,col="Mato_Grosso_do_Sul")) +
  geom_col(aes(y = Sao_Paulo,col="São Paulo")) +
  scale_x_continuous("Ano", labels = as.character(2001:2020), breaks = 2001:2020)+
  labs(x = "Ano",y = "Area Km²",color="Estado")
```

Fonte: Autores.

Apresenta-se no Quadro 1, uma síntese das linhas de comando, funções, parâmetros e pré-requisitos utilizados para a criação da visualização gráfica da Figura 2 e Figura 3. Além dos pré-requisitos necessários, apresentados no quadro, e com base nas barreiras encontradas, verificou-se que os usuários devem possuir competências básicas no idioma predominante do R e raciocínio lógico para correlacionar os parâmetros do software.

Quadro 1 – Pré-requisitos para construção dos gráficos

Linhas de comando	**Função**	**Parâmetros**	**Descrição**
library(readxl) *library(ggplot2)*	*library()*	Pacote de bibliotecas	Permite carregar bibliotecas de funções
		readxl	Conjunto de funções para acesso e interação com planilhas no Excel
		ggplot2	Conjunto de funções para construção de gráficos
dados <-*read_excel* ("arquivo.xlsx")	*read_excel*()	"nome_do_arquivo"	Retorna o conteúdo do arquivo que neste caso é atribuído a uma variável denominada 'dados'
ggplot(dados […])	*ggplot*()	dados, aes(), *geom_line*(), *geom_col*(), *scale_x_continuous*(), *labs*()	Executa o gráfico com as definições estabelecidas nos respectivos parâmetros

Fonte: Autores.

Considerado uma linguagem de programação bem desenvolvida, o R é um software livre que oferece um ambiente para análise interativa de dados, podendo ser utilizado em uma ampla variedade de plataformas, sendo considerado simples e eficaz que inclui condicionais com funções recursivas definidas pelo usuário (R FOUNDATION, 2021).

Contudo, ressalta-se que do ponto de vista da práxis, faz-se necessário um conhecimento mínimo de ferramentas tecnológicas que propiciam o desenvolvimento de uma matriz estruturada dos dados em planilhas de cálculos, como Excel, que seja interpretável por uma ferramenta como o R. Para seu uso se faz necessário conhecer as bibliotecas que podem ser utilizadas para atender demandas específicas, e qual a sintaxe para operar as funções disponíveis no pacote. E, a partir das funções que se pretende usar, quais os parâmetros e os argumentos necessários para executar a função.

Assim, para que o usuário tenha autonomia na construção de visualizações gráficas, é necessário um esforço considerável, uma vez que requer competências que demandam conhecimento específico da ferramenta que será utilizada para que se possa interpretar os conjuntos de elementos intrínsecos ao software.

5. Considerações finais

Compreender as implicações que envolvem as tecnologias utilizadas para o tratamento de grandes volumes de dados em suas aplicações, não se restringe apenas ao uso sistêmico da

informação, mas também para as relações que os usuários têm com as ferramentas que realizam tais operações.

Neste trabalho, buscou-se avaliar barreiras para construção de visualizações gráficas a partir do software R e na busca pela autonomia do usuário, que pretende usar a ferramenta para gerar visualização de dados.

Constatou-se que, o uso do R em suas funções básicas requer, ainda, um considerável volume de conhecimento sobre suas principais funções e parâmetros, além, é claro, sobre seus pacotes de bibliotecas, bem como, conhecimento relacionado à essência dos processos que envolvem as etapas do Ciclo de Vidados Dados como a coleta, armazenamento e recuperação dos dados. Assim, mesmo com todas as vantagens de simplificações oferecidas por recursos como o R, ainda existe um longo caminho na construção do encapsulamento da complexidade a ser percorrido na busca por maior autonomia do usuário na construção de visualizações de dados.

Referências

BATINI, C.; SCANNAPIECO, M. **Data and information quality: Dimensions, Principles and Techniques**. Cham, Switzerland: Springer International Publishing, 2016.

CARD, S. K.: MACKINLAY, J. D.: SHNEIDERMAN, B. **Readings in information visualization**: using vision to think. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1999.

FARIA, P. D.; PARGA, J. P. F. A. **Introdução à Linguagem R**: seus fundamentos e sua prática. 1. ed. Belo Horizonte: [s.n.], 2020.Disponível em: https://pedro-faria.netlify.app/pt/publication/book/introducao_linguagem_r/. Acesso em: 21 maio 2021.

OGD – Open Government Data. **The Annotated 8 Principles of Open Government Data**. 2007. Disponível em: https://opengovdata.org/. Acesso em: 21 maio 2021.

R FOUNDATION. **What is R?**. 2020. Disponível em: https://www.r-project.org/about.html. Acesso em: 24 maio, 2021.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, [S.l.], v. 21, n. 2, p. 116–142, dez. 2016. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/27940/20124. Acesso em: 5 jul. 2020.

SANT'ANA, R. C. G.; RODRIGUES, F. A. Acessando dados para visualização de afinidades nas votações entre parlamentares do senado. **Informação & Sociedade**: Estudos, v. 23, n. 1, 2013. Disponível em: http://hdl.handle.net/20.500.11959/brapci/92985. Acesso em: 21 maio 2021.

SILVEIRA, D. T., CÓRDOVA, F. P. **A pesquisa científica**. In: Gerhardt, T. E., & Silveira, D. T. (Orgs.). Métodos de pesquisa. Porto Alegre: Editora da UFRGS, 2009.

TREVIÑOS, A. N. S. **Introdução à pesquisa em ciências sociais**: a pesquisa qualitativa em educação. São Paulo: Atlas, 1987, p.117.

TERRABRASILIS. **PRODES Desmatamento.** 2021. Disponível em: http://terrabrasilis.dpi.inpe.br/app/dashboard/deforestation/biomes/cerrado/increments. Acesso em: 20 maio 2021.

WICKHAM, H.; GROLEMUN, G. **R for Data Science.** 2016. Disponível em: https://r4ds.had.co.nz. Acesso em: 24 maio, 2021.

WICKHAM, H. **ggplot2**: Elegant Graphics for Data Analysis. 2ª Edition. Springer, 2016.