Desenvolvimento de uma Aplicação Interativa para Visualização e Análise de Dados

Samantha Rico Gonçalves [1], Jorge Luis Boeira Bavaresco [1], Carlos Alexandre Silva dos Santos [1]

[1] Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense (IFSUL), Campus Passo Fundo

sah.rico@gmail.com, jlbavaresco@gmail.com, carlos.al.silva@live.com

Abstract. *Increasingly seek ways to improve visualization and data analysis. The R programming language supports, for example, the generation of several types of graphs and statistical informations, useful to extract informations from data. Shiny package was developed to complement the R language, allowing you to create Web applications using the language. This paper presents the stages of development of an interactive application using the R language with R Shiny package, to facilitate the visualization and analysis of a data set.*
Pavras-chave: *Facility. Statistic. Charts. Programming.*

Resumo. *Cada vez mais se buscam maneiras de se aperfeiçoar a visualização e a análise de dados. A linguagem de programação R auxilia, por exemplo, na geração de diversos tipos de gráficos e de informações estatísticas, que são úteis para que se possam extrair informações dos dados. Já o pacote* Shiny *foi desenvolvido para complementar a linguagem R, permitindo a criação de aplicações Web utilizando a linguagem. Este artigo apresenta as etapas do desenvolvimento de uma aplicação interativa utilizando a linguagem R com o pacote* Shiny*, para facilitar a visualização e a análise de um conjunto de dados. Como resultado da pesquisa obteve-se o protótipo de uma aplicação interativa para visualização e análise de dados, que permitiu que o usuário final interagisse com seus dados de forma dinâmica e facilitada.*
Pavras-chave: *Facilidade. Estatística. Gráficos. Programação.*

1. Introdução

A tecnologia tem papel fundamental em diversas áreas do mundo e está sempre em busca de inovações e formas de facilitar a vida das pessoas. Atualmente, existe uma grande

disponibilidade de dados sobre os mais diversos temas, inclusive de forma pública, como é o caso de alguns sites governamentais como o Portal Brasileiro de Dados Abertos e o Instituto de Pesquisa Econômica Aplicada, entre outros, que disponibilizam dados sobre diferentes temas, como economia, educação, geográficos, meteorológicos e assim por diante.

Além disso, podem-se citar as redes sociais, que como MENDES [2011] explica, estão muito presentes na vida de todos, principalmente das novas gerações que já começam a vida teclando e vivenciando um mundo rápido, instantâneo, com troca de dados a cada instante, convivendo com um enorme volume de informações. Levando em consideração as afirmações anteriores, percebe-se a importância de se possuir uma forma que facilite a extração de informações dos dados. Assim como IDREOS *et al.* [2015] diz, a exploração de dados tem como objetivo a extração eficaz das informações dos dados, mesmo que o usuário não saiba o que está procurando.

Nesse contexto, o desenvolvimento de uma aplicação que permita a visualização e a análise interativa de um conjunto de dados pode auxiliar os usuários que necessitam extrair, de forma mais facilitada, o conteúdo informacional dos dados. Para desenvolver a aplicação, foi utilizada a linguagem de programação R para a visualização dos dados, geração de gráficos e informações estatísticas. E o pacote *Shiny* que permitirá o desenvolvimento e a disponibilização da aplicação na *Web*.

2. Revisão da Literatura

2.1. Linguagem de Programação R

MATLOFF [2009] explica que a linguagem de programação R "é uma linguagem de script para manipulação de dados e análise estatística", ou seja, uma linguagem voltada para fins científicos, matemáticos e estatísticos. O R é um software livre, desta forma, possui código fonte aberto e pode ser modificado ou implementado com novos recursos por seus usuários contando com grande número de colaboradores das mais diversas áreas [SOUZA *et al.*, 2008]. A linguagem R, além de todas as características já citadas, possui a capacidade de ler dados compactados (zip, tar.gz etc.) e de ler arquivos de diversos formatos (CSV, xls, tab etc). Com o R, é simples obter dados de aplicações em formatos de *Web* (dados JSON, XML, HTML, etc), além de possuir a capacidade de interagir com diferentes bancos de dados (Postgres, MongoDB etc.) [RADU *et al.*, 2014].

De acordo com SOUZA *et al.* [2008], o R é uma importante ferramenta na análise e na manipulação de dados, com testes paramétricos e não paramétricos, modelagem linear e não linear, análise de séries temporais, análise de sobrevivência, simulação e estatística espacial, entre outros, além de apresentar facilidade na elaboração de diversos tipos de gráficos, no qual o usuário tem pleno controle sobre o gráfico criado.

2.2. Pacote *Shiny*

Segundo RADU *et al.* [2014], o R não serve apenas para análise quantitativa, mas é usado também para construir aplicativos de desktop e aplicações *Web*. E foi para permitir o desenvolvimento de aplicações *Web* interativas, com a utilização do R, que surgiu o pacote *Shiny*. O pacote *Shiny* permite que os usuários transformem suas análises em aplicações *Web* interativas e que podem ser acessadas por diversos outros usuários. E o mais interessante é que, para desenvolver a aplicação, não é necessário possuir um conhecimento prévio em HTML, CSS ou Javascript, já que a programação é realizada toda em R [RADU *et al.*, 2014].

As aplicações criadas com o *Shiny* são divididas em dois arquivos, o *ui.R* e o *server.R*, que são scripts que fazem parte do pacote *Shiny*. O script *ui.R* controla o layout e a aparência da aplicação, enquanto o script *server.R* contém as instruções que o computador precisa para construir a aplicação [RSTUDIO INC., 2014]. O pacote *Shiny* se tornou uma alternativa para os usuários que necessitam desenvolver aplicações *Web* interativas de forma facilitada, utilizando a linguagem R, podendo criar as mais diversas aplicações, para as mais diversas funções.

2.3. Formato de dados CSV

Segundo REPICI [2004], o formato CSV (Comma Separated Values) é usado frequentemente para trocar dados entre aplicações diferentes. Algumas características dos arquivos no formato CSV são citadas por SHAFRANOVICH [2005], como sua configuração, onde explica que cada registro se localiza em linhas separadas, delimitadas por uma quebra de linha. Pode haver um cabeçalho opcional na primeira linha do arquivo. Cada campo pode ou não estar com aspas duplas.

3. Materiais e Métodos

Os requisitos foram levantados com base nas principais necessidades dos usuários. O primeiro requisito é que a aplicação permita o envio de arquivos de dados no formato CSV ou TXT, deve possibilitar que o usuário informe as características do arquivo, como por exemplo, qual é o separador dos dados (ponto, vírgula, ponto e vírgula), se os dados possuem cabeçalho ou não, entre outras informações pertinentes.

Além disso, é de suma importância para o cumprimento do seu objetivo, exibir os dados do arquivo, enviado na tela, para que o usuário possa visualizá-los de forma interativa. O usuário, também, deverá ter a opção de gerar diferentes tipos de gráficos a partir dos dados do arquivo enviado, tais como *boxplot*, histograma e *plot*, além de realizar alguns cálculos básicos de estatística (média aritmética, mediana, quartis), para que se possa obter uma melhor análise sobre as informações. A partir dos requisitos, portanto, será

desenvolvida a aplicação utilizando a linguagem de programação R e sua IDE (*Integrated Development Environment*), o RStudio com o pacote *Shiny* para que seja possível disponibilizar a aplicação na *Web*.

A estrutura da aplicação está dividida em dois arquivos principais o "*ui.R*" e o "*server.R*". O arquivo *"ui.R"* contém o código referente à interface da aplicação e o arquivo *"server.R"* contém o código que realiza a comunicação para o funcionamento da aplicação.

Para cumprir com os requisitos da aplicação, a primeira função desenvolvida foi o envio de um arquivo de dados pelo usuário. Na linguagem R, o comando que realiza a leitura de dados é o *read.table()*. A função *reactive()* foi utilizada porque pode ser utilizada diversas vezes durante a aplicação , ou seja, essa função poderá ser reativada sempre que necessário . Para que a leitura de dados se torne interativa , o usuário terá a opção de informar alguns parâmetros , como *header* (cabeçalho), *quote* (aspas), *sep* (separador).

A figura 1 exibe a tela principal da aplicação e nela é possível visualizar as possibilidades de escolha de características que o usuário poderá informar de acordo com o arquivo enviado. Após o envio do arquivo, os dados são exibidos na tela, onde os resultados são paginados, ou seja, os dados são dispostos em páginas, cada página contendo uma determinada quantidade de dados para facilitar a navegação do usuário. Ainda, é possível efetuar consultas utilizando filtros sobre as colunas e buscar dados específicos por meio da busca, alcançando, assim, a interatividade da visualização de dados.

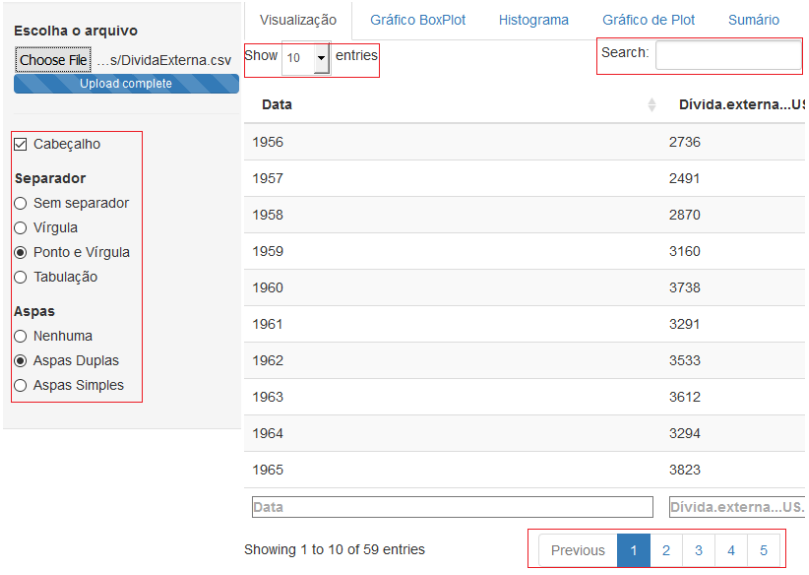

Figura 1. Características dos Dados.

É necessário destacar que a aplicação foi desenvolvida, dessa maneira, para que o usuário possa enviar e visualizar seus arquivos de dados de forma interativa, devido

ao fato de que os dados podem possuir diferentes tipos de configurações. Outra parte muito importante da aplicação é a geração de gráficos. A aplicação permite a geração de três tipos diferentes de gráficos que são o *boxplot*, o histograma e o *plot*, sendo que cada um deles possui uma implementação diferente. A exibição dinâmica das colunas dos dados foi realizada para os outros tipos de gráficos, também. Após gerar as opções para o usuário selecionar, foi gerado o gráfico em si, utilizando essas informações. Para que a exibição do gráfico fosse possível, utilizou-se a função *renderPlot()*, que permite renderizar o gráfico na tela. Após a chamada da função de leitura de dados, buscam-se os nomes das colunas dos dados e então, por meio da função *boxplot()*, gera-se o gráfico utilizando as informações escolhidas pelo usuário. Dessa forma, a geração de gráficos se torna dinâmica.

Para a geração dos outros tipos de gráficos, o histograma e o *plot*, utilizou-se o mesmo segmento. A exibição de ambos foi realizada da mesma forma que a do gráfico *boxplot*. O usuário seleciona as opções das colunas e o gráfico é exibido. A figura 2, mostra como é a exibição dos gráficos.

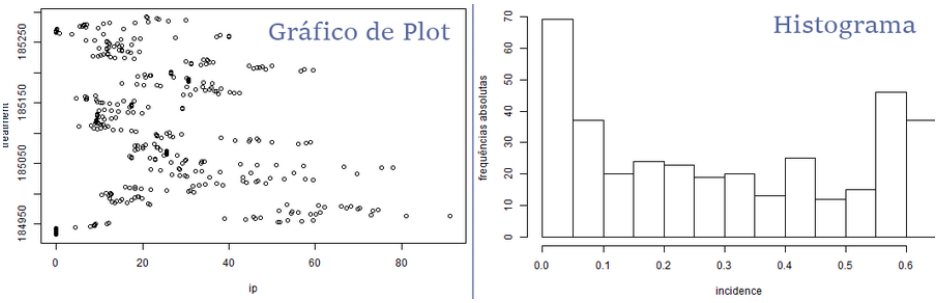

Figura 2. Exibição dos gráficos plot e histograma.

Para gerar os gráficos histograma e *plot* foram utilizadas, respectivamente, as funções *hist()* e *plot()*, tendo como parâmetros as informações escolhidas pelo usuário. Outra parte muito importante da aplicação é a geração de informações estatísticas como média, mediana, entre outras. Para isso, utiliza-se a função *summary()* que gera, automaticamente, essas informações. É relevante salientar a importância do sumário para que a aplicação cumpra com o seu objetivo, já que é por meio dessas informações, além dos gráficos, que o usuário poderá realizar uma melhor análise sobre as informações e encontrar possíveis inconsistências nelas. A figura 3 mostra como as informações são exibidas na aplicação. No caso, para cada coluna do conjunto de dados, está sendo exibida as suas informações de sumário (média, mediana, menor valor, maior valor, primeiro quartil e terceiro quartil).

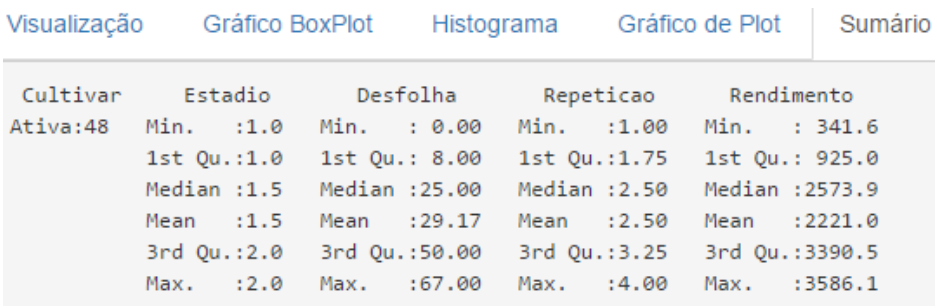

Figura 3. Exibição do sumário.

A aplicação foi testada, localmente, por intermédio do RStudio, mas o *Shiny* permite que a aplicação seja disponibilizada na *Web*, por meio de seu próprio servidor: o *Shiny* Server. O servidor *Shiny* possibilita que o usuário possua mais de uma aplicação hospedada nele, cada uma com uma URL própria. Essa é uma visão geral sobre o desenvolvimento da aplicação. O importante é que se buscou tornar a aplicação interativa o bastante para que o usuário possa visualizar e analisar seus dados de uma forma facilitada.

4. Resultados e Discussões

Ao fim do desenvolvimento da aplicação, atingiu-se o seu principal objetivo que é a interatividade, já que o usuário poderá analisar e visualizar os seus dados de uma forma mais facilitada do que apenas lê-los de uma maneira convencional ou recorrer à utilização da programação. Além disso, os gráficos são gerados de forma dinâmica, com a possibilidade da seleção das colunas do conjunto de dados para os valores dos eixos x e y dos gráficos. E um sumário com informações como média, menor e maior valor dos dados, entre outras. Todos esses recursos são encontrados na aplicação.

Uma das desvantagens da aplicação é a dificuldade na leitura de dados com tamanhos muito grandes ou com uma má formatação, o que acaba acarretando alguns erros , como a lentidão na hora de visualizar os dados, como gerar os gráficos e o sumário. Para validar que a aplicação cumpre com o que propõe , foram realizados alguns testes com conjuntos de dados. Utilizaram-se dados de saída de um modelo de simulação de doenças, no caso , a brusone no trigo . Na figura 4, pode-se visualizar como o conjunto de dados utilizado está estruturado

treatment	semeadura	ano	dia	city	latitude	longitude	cultivar	heading_date	ip	incidence
184933	2001-03-26	2001	26-3	CIANORTE	-23.4	-52.35	BRS-LOURO	2001-07-27	0	0
184934	2001-03-31	2001	31-3	CIANORTE	-23.4	-52.35	BRS-LOURO	2001-08-01	0	0
184935	2001-04-05	2001	5-4	CIANORTE	-23.4	-52.35	BRS-LOURO	2001-08-06	0	0
184936	2001-04-10	2001	10-4	CIANORTE	-23.4	-52.35	BRS-LOURO	2001-08-11	0	0
184937	2001-04-15	2001	15-4	CIANORTE	-23.4	-52.35	BRS-LOURO	2001-08-16	0	0

Figura 4. Dados Brusone no Trigo.

Nessa primeira análise, foi selecionada a coluna ip (inóculo potencial) e o ano, utilizando o gráfico *boxplot*, para visualizar a quantidade de inóculo que ocorreu a cada ano, como mostra a figura 5.

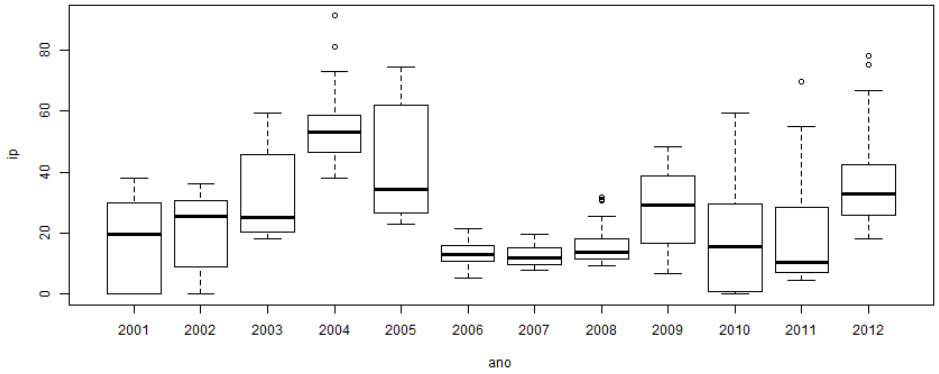

Figura 5. Analisando dados por meio de gráficos.

Por meio desta análise, é possível que o usuário possa ter uma melhor noção sobre os anos em que a incidência da doença foi maior ou menor, auxiliando aos pesquisadores a extração da informação sobre o comportamento da doença ao longo dos anos. Os resultados obtidos, portanto, estão de acordo com o que se propôs alcançar com o desenvolvimento da aplicação.

5. Conclusão

O objetivo desse projeto foi o desenvolvimento de uma aplicação interativa capaz de aperfeiçoar e facilitar a visualização e a análise de um conjunto de dados utilizando a linguagem de programação R e o pacote *Shiny*. A implementação da aplicação foi realizada com o propósito de permitir que o usuário interaja com seus dados de forma dinâmica e facilitada.

É interessante salientar a notável facilidade com que é possível desenvolver uma aplicação *Web* utilizando a linguagem de programação R e seu pacote *Shiny*. Os resultados, que foram obtidos por meio do desenvolvimento, não teriam sido possíveis sem os diversos

recursos que essas tecnologias disponibilizam. Em termos de conhecimento, pode-se afirmar que aprender a linguagem R pode ser de muita utilidade, quando se trata de visualização e de análise de dados.

A maior dificuldade encontrada durante o desenvolvimento foi a busca por formas de tornar a aplicação interativa pois em termos de desenvolvimento é mais fácil desenvolver uma aplicação para um conjunto de dados específico do que para dados de tipos variados, uma vez que não há como saber a forma como os dados enviados para a aplicação estão estruturados, o que gera dificuldade para o desenvolvimento.

Como proposta para trabalhos futuros, é possível realizar a integração da aplicação com o banco de dados, já que a linguagem R permite esse tipo de interação, para que, assim, o usuário possa buscar seus dados diretamente do banco de dados. Além disso, pode ser realizada a adição de novos recursos e funcionalidades à aplicação existente, como novos gráficos e informações estatísticas.

Referências

IDREOS, S., PAPAEMMANOUIL, O., e CHAUDHURI, S. (2015). Overview of Data Explo- ration Techniques. In *ACM SIGMOD International Conference on Management of Data, Tutorial*, Melbourne, Austrália.

MATLOFF, N. (2009). The art of R programming. Disponível em: <http://heather.cs.ucdavis.edu/~matloff/132/NSPpart.pdf>. Acesso em: 01/03/2018.

MENDES, A. (2011). As redes sociais e sua influência na sociedade. Disponível em: <http://goo.gl/6P5YSB>. Acesso em: 29/10/2015.

RADU, M., MURESAN, I., e NISTOR, R. (2014). Using R To Get Value Out Of Public Data. Disponível em: <http://goo.gl/GQQckw>. Acesso em: 10/06/2015.

REPICI, J. (2004). HOW-TO: The Comma Separated Value (CSV) File Format. Disponível em: <http://goo.gl/PFDgH>. Acesso em 11/05/2015.

RSTUDIO INC. (2014). Shiny: Web Application Framework for R. SHAFRANOVICH, Y. Disponível em: <http://goo.gl/EDyxyt>. Acesso em: 11/06/2015.

SOUZA, E. F. M., PETERNELLI, L. A., e MELLO, M. P. D. (2008). Software Livre R: aplicação estatística. Disponível em: <http://goo.gl/xQqK27>. Acesso em: 08/06/2015.