ESTUDO COMPARATIVO ENTRE OS SOFTWARES R E MATLAB NA ESTATÍSTICA

COMPARATIVE STUDY BETWEEN R AND MATLAB SOFTWARE IN STATISTICS

ESTUDIO COMPARATIVO ENTRE LOS SOFTWARES R Y MATLAB EN ESTADÍSTICA

Raquel Nailê Brinkhus[1]
Guilherme Augusto Pianezzer[2]

Resumo

O presente trabalho realizou uma análise comparativa entre os softwares R e Matlab. Inicialmente, verificou-se que ambos apresentam vantagens similares, como a fácil utilização e a boa interface gráfica; quanto às desvantagens, constatou-se que o Matlab tem um alto custo de aquisição e o *software* R um tempo de processamento maior — quando comparado a outras linguagens. Subsequentemente, realizou-se um levantamento das principais funções da análise estatística, a saber: a média, mediana, valor máximo, valor mínimo, desvio padrão e variância. Efetuaram-se dez testes de tempo de processamento de cada uma destas rotinas nos *softwares*, para duas populações de dados; logo após, calculou-se a média dos dados encontrados. Os resultados indicaram que, para o cálculo da média, mediana, valor máximo e valor mínimo, o tempo de execução foi menor no *software* Matlab; já para o cálculo do desvio padrão e variância, o tempo menor ocorreu no *software* R. Destarte, para os casos estudados, a diferença entre resultados aumenta com o crescimento da população; isto é, caso o volume de dados a serem analisados aumente, a diferença no tempo de execução tende a ser maior.

Palavras-chave: linguagem R; linguagem Matlab; estatística; velocidade de processamento.

Abstract

The present work performed a comparative analysis between R and Matlab software. Initially, it was found that the programs have similar advantages, such as ease of use and good graphical interface; as for the disadvantages, it was found that Matlab has a high acquisition cost and the R software a longer processing time — when compared to other languages. Subsequently, a survey of the main functions of the statistical analysis was carried out, namely: the mean, median, maximum value, minimum value, standard deviation and variance. Ten processing time tests were performed for each of these routines in the software, for two data populations; soon after, the average of the data found was calculated. The results indicated that for calculating the mean, median, maximum and minimum values, the execution time was shorter in Matlab software; for calculating of standard deviation and variance, the execution time was shorter in the R software. Thus, for the cases studied, the difference between results increases with population growth; that is, if more data is analyzed, the tendency is for the difference in execution time to be even greater.

Keywords: R language; Matlab language; statistics; processing speed.

Resumen

El presente trabajo realizó un análisis comparativo entre los *softwares* R y Matlab. Inicialmente, se verificó que ambos presentan ventajas similares, como una fácil utilización y buena interfaz gráfica; sobre las desventajas, se constató que el Matlab tiene un alto costo de adquisición y el *software* R un tiempo de procesamiento más largo — cuando comparado a otros lenguajes. Luego, se hizo una recopilación de las principales funciones del análisis estadístico, a saber: media, mediana, valor máximo, valor mínimo, desviación estándar y varianza. Se hicieron diez pruebas de tiempo de procesamiento para cada una de esas rutinas en los *softwares*, con dos poblaciones de datos; en seguida, se calculó la media de los datos encontrados. Los resultados indicaron que, para el cálculo de la

[1] Mestre em Cálculo Estrutural pela UFRGS, professora da CESURG - Sarandi e acadêmica do curso de bacharelado de Matemática do Centro Universitário Internacional UNINTER. E-mail: quelchi@gmail.com.
[2] Doutor em Métodos Numéricos em Engenharia pela UFPR. Docente na área de Exatas do Centro Universitário Internacional UNINTER. E-mail: guilherme.pi@uninter.com.

media, mediana, valor máximo y valor mínimo, el tiempo de ejecución fue más corto en el *software* Matlab; en cambio, para el cálculo de la desviación estándar y la varianza, el tiempo más corto se dio en el *software* R. Así, para los casos estudiados, la diferencia entre resultados aumenta con el crecimiento de la población; es decir, si el volumen de datos a ser analizados aumenta, la diferencia en el tiempo de ejecución tiende a ser más largo.

Palabras-clave: lenguaje R; lenguaje Matlab; estadística; velocidad de procesamiento.

1 Introdução

Na atualidade, inúmeros *softwares* são capazes de resolver os mais diversos problemas; no entanto, todos apresentam aspectos diferentes, tais como, facilidade de utilização, performance e quantidade de funções disponíveis. É fulcral escolher cuidadosamente o *software* a ser utilizado, para que as soluções desejadas sejam alcançadas. Com o crescimento da tecnologia, a quantidade de dados disponíveis é cada vez maior; por isso, há uma crescente necessidade de ferramentas que não só analisem estes dados, mas que o faça no menor tempo possível. Entre as linguagens mais utilizadas para análise de dados e estatística, podemos citar o software R e o Matlab, que possuem funções prontas e permitem a criação de rotinas — que variam entre análises simples até as mais complexas. O objetivo geral deste trabalho é realizar uma análise comparativa entre o software R e Matlab no campo da estatística; o intuito é comparar a facilidade de uso, suas vantagens e desvantagens, além do tempo de processamento de funções estatísticas. Desta forma, o problema de pesquisa a ser respondida neste trabalho é: *quais as considerações importantes sobre o software R e o Matlab para o uso na estatística?*

Já os objetivos específicos da investigação serão: realizar uma pesquisa bibliográfica sobre o *software* R; realizar uma pesquisa bibliográfica sobre o software *Matlab*; documentar as vantagens e desvantagens destes dois softwares; e comparar a performance deles para solucionar problemas comuns da estatística.

Para o embasamento teórico, buscou-se, inicialmente, artigos e materiais relacionados ao tema; destarte, foi possível definir os dois *softwares* e descrever suas vantagens e desvantagens. A partir destas informações, propõe-se uma rotina com funções estatísticas para analisar o tempo de processamento de duas populações de dados, em cada um dos programas, e, por fim, comparar os resultados encontrados.

Gilat (2012) apresenta o *software* Matlab, enquanto Chapman (2016) pontua as vantagens do seu uso; em relação ao *software* R, Oliveira (2018) o apresenta, Barros (2018) aponta as vantagens de uso e Faria e Parga (2020) as desvantagens.

O estudo está organizado da seguinte forma: inicialmente, serão apresentados os *softwares* Matlab e R e as suas vantagens de uso; subsequentemente, serão apresentadas as

funções estatísticas disponíveis nos *softwares*; e, para finalizar, será mostrado e comparado o tempo de processamento de cada função em cada *software*, para cada uma das populações.

2 Software R e Matlab

Na seção a seguir, apresenta-se a revisão básica da literatura, que embasará o presente estudo. Foram utilizados livros, artigos e trabalhos acadêmicos que abordam os *softwares* R e Matlab e sobre funções estatísticas.

2.1 Fundamentação teórica

Os *softwares* R e Matlab são linguagens de programação criadas para análise matemática; ou seja, é possível desenvolver, por meio deles, rotinas e utilizar funções matemáticas para resolver os mais diversos problemas.

Segundo Gilat (2012, p. 12), o Matlab "é bastante versátil em cálculos matemáticos, modelagens e simulações, análises numéricas e processamentos, visualização e gráficos, desenvolvimentos de algoritmos, etc.".

Conforme Chapman (2016), as vantagens de uso do Matlab são inúmeras, como, por exemplo: a facilidade de uso; a independência da plataforma; funções predefinidas; representações gráficas independentes de dispositivos; interface gráfica; e o compilador Matlab. O mesmo autor ainda apresenta algumas das desvantagens do uso do *software*, como a linguagem interpretada e o alto custo de aquisição.

O *software* R, amplamente utilizado para análise estatística, foi criado justamente com este propósito. De acordo com Oliveira (2018, p. 9), "Podemos entender o R também como um conjunto de pacotes e ferramentas estatísticas, munido de funções que facilitam sua utilização, desde a criação de simples rotinas até análises de dados complexas, com visualizações bem-acabadas".

Barros (2018) cita algumas vantagens da linguagem R, tais como uma linguagem simples, eficaz e instalações gráficas para análise. A autora também complementa que é uma: "grande coleção coerente e integrada de ferramentas intermediárias para análise de dados." (BARROS, 2018, p. 1).

Oliveira (2018) apresenta, também, inúmeras vantagens da linguagem R: "é completamente gratuito e de livre distribuição; muito fácil de se aprender; enorme quantidade de tutoriais e ajuda disponíveis gratuitamente na internet; amplamente utilizado pela comunidade acadêmica e pelo mercado." (OLIVEIRA, 2018, p. 9).

Como desvantagens da linguagem R, os autores Faria e Parga (2020, p. 5) citam que "essas linguagens (Python e R) não são particularmente rápidas se comparadas com outras linguagens que lhe obrigam a especificar cada componente de sua análise".

As duas linguagens são utilizadas na criação de rotinas matemáticas e tem como vantagens a fácil utilização e boa interface gráfica. O Matlab tem como desvantagem o seu elevado custo de aquisição, enquanto o software R é gratuito; entretanto, apresenta a desvantagem de não ser tão rápido, se comparado a outras linguagens.

Tais ferramentas possuem diversas funções matemáticas e estatísticas prontas; para os problemas mais complexos, em que uma única função não resolva, é possível realizar uma rotina (sequência) de comandos, por exemplo.

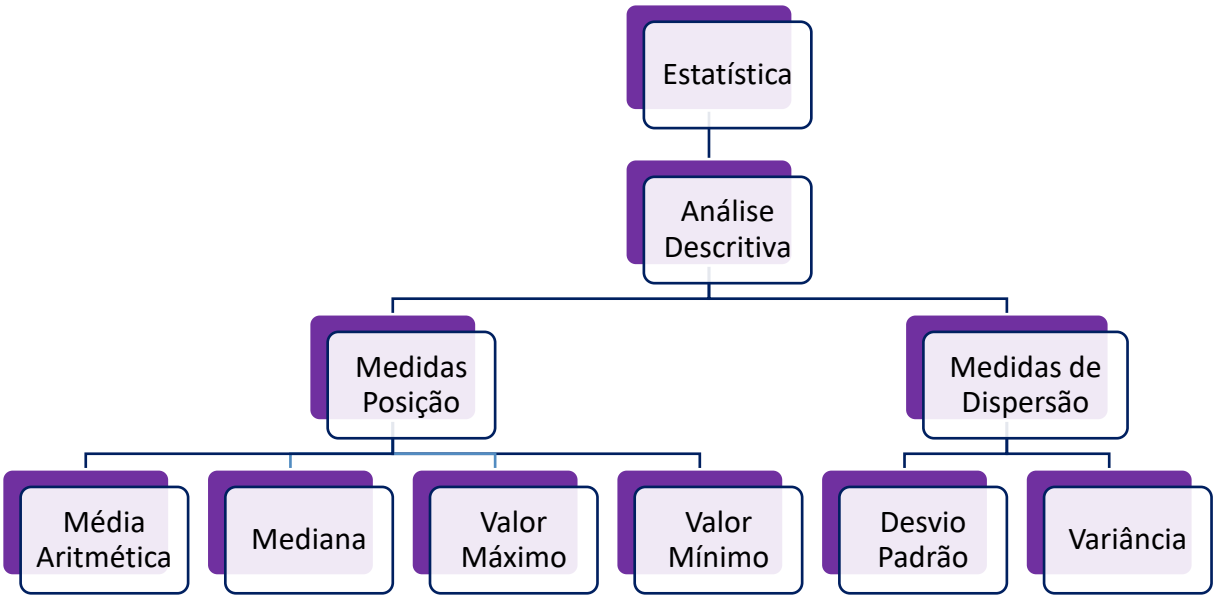
Em relação à estatística, Silva, Grams e Silveira (2018, p. 13) apontam que "toda evolução humana se dá em virtude de descobertas e invenções, que podem ser criadas ou adaptadas para contribuir e descomplicar a vida do homem, seja na área da saúde, engenharia, economia, comunicação, entre outras". Desta forma, percebemos que todas as áreas a necessitam para o avanço da descoberta e invenções. As autoras complementam que "essa evolução se deve, em grande parte, à análise de dados coletados nas mais diversas áreas. " (SILVA, GRAMS; SILVEIRA, 2018, p. 13); concluímos, assim, que a análise de dados pode contribuir com todas as áreas.

Silva, Grams e Silveira (2018, p. 13) versam que "coletar e analisar tais dados são funções da estatística, embasando decisões, planejamentos, sabendo como obter dados úteis e, principalmente, o que fazer com eles". Observa-se, então, que a estatística é fundamental para nortear as tomadas de decisões e para a melhor garantia de resultados.

Na estatística, há diversas técnicas que auxiliam desde o processo de captação de dados até a sintetize e análise destes. A estatística descritiva é aquela auxilia no que fazer com os dados coletados; ou seja, como sintetizar e analisar o comportamento de um grupo de elementos.

As técnicas da estatística descritiva são divididas em dois grupos: as medidas de posição e as medidas de dispersão. Conforme Virgillito (2017, p. 77), "medidas de posição que ajudam a interpretar inicialmente a maioria dos eventos estudados", como, por exemplos, os cálculos de médias, mediana, valor máximo e valor mínimo. O autor também versa sobre as medidas de dispersão, "quando se fala em variabilidade ou oscilação" (VIRGILLITO, 2017, p. 77), como, por exemplo, o cálculo do desvio padrão e da variância. Isto posto, cada técnica supracitada será melhor apresentada na sequência. A Figura 1 mostra esta área da estatística e apresenta algumas das técnicas.

Figura 1: Estatística descritiva



Fonte: a autora (2021).

Para Virgillito (2017, p. 77), "a média é o cálculo estatístico mais elementar [...] talvez o cálculo mais instintivo do pesquisador e também dos profissionais"; desta forma, a média é a primeira técnica que deve ser realizada.

A média aritmética, ou apenas média como é chamada, é definida por Costa Neto (2002, p. 21) como "[...] o centro da distribuição de frequências, sendo, por isso, uma medida de posição. Em uma analogia de massas, a média corresponderia ao centro de gravidade da distribuição de frequências.", e a mesma pode ser calculada conforme a equação 1.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

Em que:

$\bar{x}$ é a média;

n é o número total de valores;

x_i é cada valor.

A média não é melhor solução para todos os casos; quando deseja-se encontrar o valor central de (ideia do valor típico), a Mediana é mais recomendada. Costa Neto (2002, p. 21) postula que "a mediana pode ser usada como alternativa, em relação à média, para caracterizar o centro do conjunto e dados. Em certos casos, efetivamente, seu uso é mais conveniente". Caso

o conjunto de valores analisados n seja ímpar, a Mediana pode ser calculada conforme a equação 2.

$$md = x_{(n+1)/2} \tag{2}$$

Em que:

md é a mediana.

Caso o conjunto de valores analisados n seja par, a Mediana pode ser calculada conforme a equação 3.

$$md = \frac{x_{(n/2)} + x_{(n/2+1)}}{2} \tag{3}$$

Além destas duas funções, na estatística, ainda há outras duas medidas que são importantes a serem definidas: os valores máximos e mínimos de um conjunto de dados.

O valor mínimo pode ser definido, conforme Becker (2015, p. 70), como "o mínimo de um conjunto de dados é definido simplesmente como o seu menor valor." Formalmente, pode ser definido conforme a equação abaixo:

$$min(X) = \min_i x_i \, , para \; i = 1, \dots, n \tag{4}$$

Em que:

min é o valor mínimo.

O Valor máximo também é definido conforme Becker (2015, p. 70): "O máximo de um conjunto de dados é definido simplesmente como o seu maior valor. " Formalmente pode ser definido conforme a equação abaixo:

$$max(X) = \max_i x_i \, , para \; i = 1, \dots, n \tag{5}$$

Em que:

max é o valor máximo.

Apesar da média ser o cálculo mais elementar e intuitivo, nem sempre apenas essas informações de medidas de posição são suficientes. Virgillito (2017, p. 81) afirma "Em muitos casos, o simples cálculo da média não proporciona uma visão do que realmente acontece com o comportamento dos dados observados e, portanto, não oferece informações para a tomada de

decisão". Desta forma, pode ser necessário utilizar também dos cálculos de medidas de dispersão (com por exemplo desvio padrão e variância).

O autor Virgillito (2017, p. 81) apresenta o conceito do desvio padrão: "medir o afastamento dos dados observados em relação à média da distribuição". O mesmo autor também apresenta a sua definição: "O desvio-padrão é definido como a "raiz quadrada da média, do quadrado dos desvios dos dados observados, em relação à média da distribuição destes", conforme a equação 6.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \tag{6}$$

Em que:

σ é o desvio padrão.

O valor do desvio padrão também está diretamente ligado à variância. Costa Neto (2002, p. 25) afirma que "A variância é uma medida de dispersão extremamente importante na teoria estatística. Do ponto de vista prático, ela tem o inconveniente de se expressar numa unidade quadrática em relação à variável em questão". A equação 7 apresenta a relação entre variância e desvio padrão.

$$var = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \tag{7}$$

Em que:

var é a variância.

2.2 Metodologia

A pesquisa utiliza uma abordagem qualitativa de investigação e quantitativa de estudo de caso. Analisaram-se os principais livros e artigos que abordem os seguintes temas: investigação sobre as definições e vantagens das linguagens R e Matlab e funções estatísticas.

Para o estudo quantitativo, realizou-se a comparação entre o tempo de execução de rotinas que usem funções comumente utilizadas na estatística, para uma população com uma mesma quantidade de valores e precisão, em ambas as linguagens.

O primeiro cuidado que se teve para realizar foi em relação à medição do tempo de processamento, pois existem várias funções que retornam o tempo de processamento de uma função. Na linguagem R, por exemplo, a função "system.time" retorna um vetor com 3 diferentes valores:

- Primeira coluna: tempo de processamento do usuário, ou seja, o tempo gasto no CPU apenas pelo processamento do software R;

- Segunda coluna: tempo do sistema, ou seja, tempo utilizado por outros processos que não são do software R (por exemplo, o sistema operacional do Windows ou outros programas abertos pelo usuário no computador);

- Terceira coluna: tempo decorrido, ou seja, tempo total sentido pelo usuário (soma da primeira e segunda coluna).

O tempo de processamento correto para comparação entre os softwares é aquele utilizado apenas pelo software R, desconsiderando outras aplicações que o usuário possa estar utilizando; por esta razão, utilizaram-se os valores retornados da primeira coluna da função "system.time".

No *software* Matlab, as funções de tempo são um pouco diferentes; não existe uma função que retorna diretamente o tempo utilizado apenas pelo software para a execução de uma função, mas existe uma função chamada "cputime" que retorna o tempo utilizado apenas pelo *software* Matlab, desde o início da sessão corrente. Desta forma, criou-se uma variável chamada "to" (representando o tempo inicial). Utilizou-se, nesta variável, a função "cputime", imediatamente antes de chamar a função que se desejava analisar; desta foram, a variável guardava o tempo de processamento gasto pelo Matlab até o momento anterior a chamada da função a ser analisada. Criou-se, também, outra variável chamada tf (representando o tempo final); nesta variável também era chamada à função "cputime", porém imediatamente após realizar a função que se desejava analisar. Ao realizar a diferença entre o tempo final e o tempo inicial, foi possível, então, obter o tempo gasto de processamento apenas pelo *software* Matlab, para calcular a função desejada.

A Tabela 1 apresenta algumas das rotinas de estatística disponíveis pelos softwares e o nome do comando utilizado para chamar estas rotinas em cada um dos dois softwares.

Tabela 1: Funções estatísticas no software R e Matlab

	Software	
Função	Matlab	R
Média	mean	mean
Mediana	median	median
Moda	mode	*

Valor Máximo	max	max
Valor Mínimo	min	min
Desvio Padrão	std	sd
Variância	var	var
Coeficiente de Variação	vc	*

* Sem função pronta (apenas com uso de bibliotecas)
Fonte: a autora (2021).

Como o objetivo é de comparar as funções prontas e de fácil acesso dos *softwares*, descartou-se o cálculo da moda e do coeficiente de variação. Desta forma, foram analisadas as seguintes funções: Média (aritmética), Mediana, Valor máximo, Valor mínimo, Desvio padrão e Variância.

O tamanho das populações (quantidade de dados a serem analisados) foram escolhidos de forma que o tempo de processamento para execução de nenhuma das rotinas fosse muito pequeno, ou seja, nenhum dos tempos retornasse zero (as funções retornam o tempo de processamento em segundos). Desta forma, foram escolhidos dois casos de populações:

- Caso A: uma população de 1e8 (cem milhões) de dados;
- Caso B: uma população de 1e9 (um bilhão) de dados.

Outra preocupação foi com relação à precisão dos dados. Para garantir que a comparação dos dados fosse igual entre dois *softwares*, utilizou-se de dados que variavam de 0 a 1 e com uma precisão de 4 casas decimais gerados aleatoriamente em ambos os *softwares*. Inicialmente, considerou-se utilizar exatamente os mesmos dados; ou seja, gerar os dados em um dos *softwares*, salvar estes dados em um arquivo de texto e após ler os dados no outro *software*. No entanto, a quantidade de dados era muito grande e acabou inviabilizando o processo; desta forma, optou-se em utilizar valores diferentes, mas, como citado, valores com a mesma precisão de casas decimais.

Além destas preocupações, foram consideradas, também, as quantidades de testes para uma mesma situação. É natural que mesmo se executando testes iguais, de uma mesma população de dados e em um mesmo *software*, os resultados de tempo de processamento não sejam exatamente os mesmos de um teste para outro. Desta forma, optou-se em repetir cada um dos testes 10 vezes e gerar uma média destes.

Após todas estas definições, foram realizados os roteiros para análise de tempo de execução das funções em cada um dos *softwares*. Destarte, a próxima seção apresenta os resultados encontrados.

2.3 Tempo de processamento

Após todas as considerações, apresentadas na metodologia, implementaram-se os códigos em ambos os *softwares* e realizados todos os testes. A Tabela 2 apresenta os resultados de tempo encontrados para a média de 10 execuções para cada função para o Caso A; a Tabela 3 apresenta os resultados para o Caso B.

Tabela 2: Tempo de execução Caso A

	Software	
Função	R (seg)	Matlab (seg)
Média	0,211	0,156
Mediana	1,988	1,436
Máximo	0,166	0,097
Mínimo	0,16	0,108
Desvio Padrão	0,459	0,648
Variância	0,464	0,586

Fonte: da autora (2021).

Tabela 3: Tempo de execução Caso B

	Software	
Função	R (seg)	Matlab (seg)
Média	2,047	1,373
Mediana	31,019	25,922
Máximo	1,696	0,891
Mínimo	1,613	0,900
Desvio Padrão	4,475	19,153
Variância	4,474	18,597

Fonte: a autora (2021).

Como o objetivo do trabalho é a comparação entre os softwares, compete ressaltar que a execução dos testes de desvio padrão e variância do Caso B, de ambos os softwares, levaram o uso da memória RAM ao máximo. A memória RAM utilizada para o processamento destes testes foi de 15,9 GB dos 16 GB (99,4% da memória disponível).

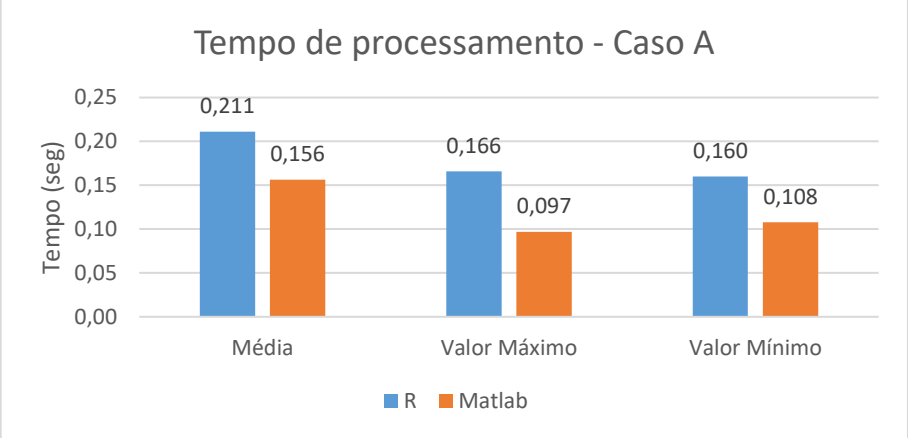
No *software* Matlab, o gerenciamento da utilização da memória é administrado pelo próprio *software*, ou seja, o usuário desenvolvedor da rotina não necessita realizar nenhuma adaptação no código ou intervenção. Já o software R não faz este gerenciamento sozinho. Por padrão, ele permite a utilização de apenas uma parte da memória, e quando o limite da memória é alcançado, ele para a execução e apresenta uma mensagem. Para contornar este problema e liberar o uso da memória, é necessário utilizar o comando "memory.limit" que expande o uso da memória.

Como os testes exigiram o uso de praticamente toda a memória RAM, no caso do software R houve a necessidade de realizar a adaptação na rotina para liberar o uso da memória, já no software Matlab não houve necessidade de realizar nenhuma alteração.

Ter a quantidade de uso da memória bloqueada tem aspectos positivos e negativos. O aspecto negativo, conforme já citado, é a interrupção da execução da rotina e adaptação do código. Já o lado positivo, é que evita que um *software* comprometa sozinho todo o uso da memória RAM. Nos casos em que a memória é utilizada apenas parcialmente, o usuário pode utilizar do restante da memória em outros softwares (realizando outras tarefas no computador); contudo, se a memória está totalmente liberada para o uso do software, o computador fica sem memória extra para utilizar em outras tarefas, fazendo com que o usuário tenha problemas caso tente realizar outras tarefas (demora excessiva).

A Figura 2 compara os resultados encontrados para o Caso A (cem milhões de dados). Para a função de média, o *software* R teve um tempo de processamento 35% mais lento se comparado ao Matlab; já para a função de valor máximo, o *software* R teve um tempo de processamento 71,31% mais lento. Para a função de valor mínimo, o *software* R teve um tempo de processamento 48,42% mais lento. Para este caso A, todas as funções analisadas tiveram um tempo de processamento menor no software Matlab.

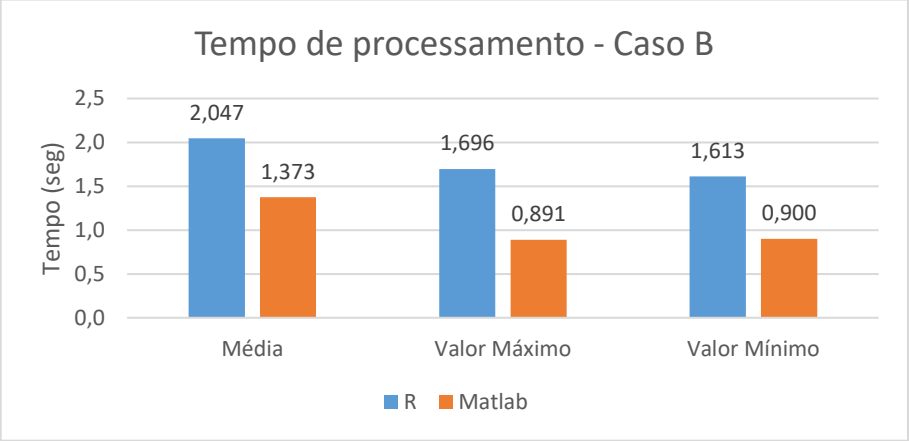
Figura 2: Tempo de processamento das funções média, valor máximo, valor mínimo para o Caso A



Fonte: da autora (2021).

A Figura 3 compara os resultados encontrados para o Caso B (um bilhão de dados). Para a função de média, o *software* R teve um tempo de processamento 49,05% mais lento quando comparado ao Matlab; para a função de valor máximo, o *software* R teve um tempo de processamento 90,43% mais lento, e para a função de valor mínimo, o *software* R teve um tempo de processamento 79,22% mais lento. Novamente, todas as funções analisadas tiveram um tempo de processamento menor no software Matlab. Além disto, percebeu-se que a comparação entre o tempo de processamento aumentou; isto é, para uma população maior, a diferença entre os tempos de execução também foi maior.

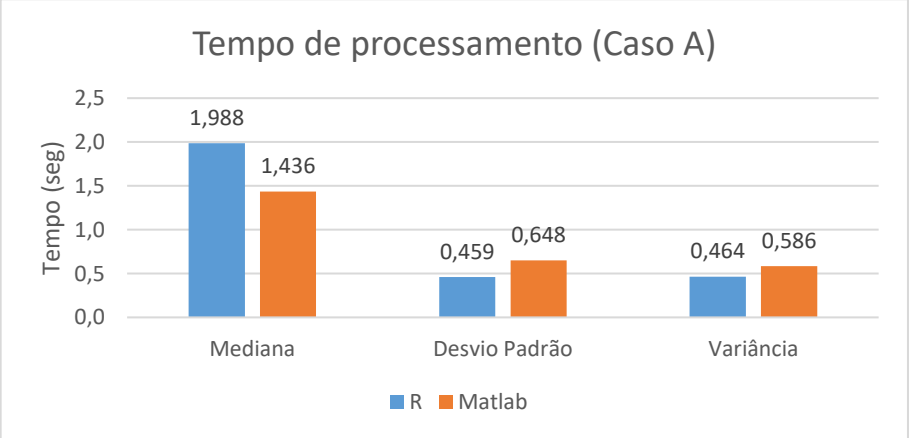
Figura 3: Tempo de processamento das funções média, valor máximo, valor mínimo para o Caso B



Fonte: a autora (2021).

A Figura 4 compara os resultados encontrados para o Caso A (cem milhões de dados). Para a função de mediana, o software R teve um tempo de processamento 38,45% mais lento quando comparado ao Matlab; para a função de desvio padrão, o *softwar*e R teve um tempo de processamento 29,21% mais rápido e, para a função de variância, o *software* R teve um tempo de processamento 20,81% mais rápido.

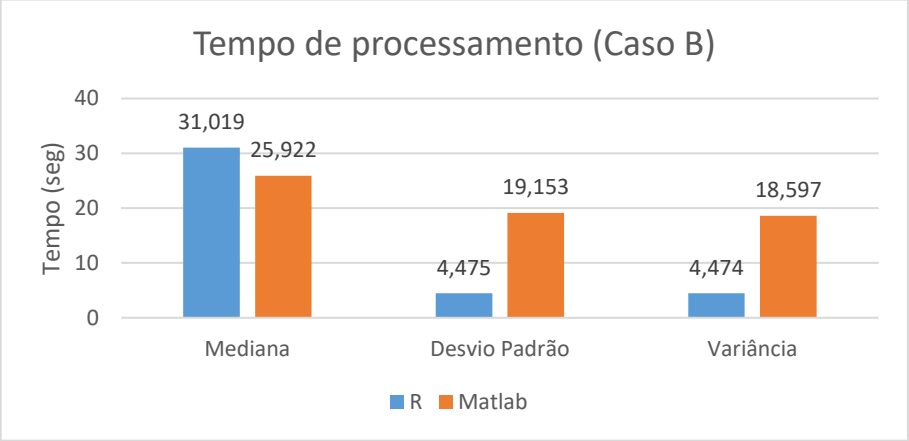
Figura 4: Tempo de processamento das funções Mediana, Desvio Padrão e Variância para o Caso A



Fonte: da autora (2021).

A Figura 5 compara os resultados encontrados para o Caso B (um bilhão de dados). Para a função de mediana, o software R teve um tempo de processamento 19,66% mais lento se comparado ao Matlab; para a função de desvio padrão, o *software* R teve um tempo de processamento 76,64% mais rápido e, para a função de variância, o *software* R teve um tempo de processamento 75,94% mais rápido.

Figura 5: Tempo de processamento das funções Mediana, Desvio Padrão e Variância para o Caso B



Fonte: da autora (2021).

Para as seis funções analisadas, tanto para o caso A quanto no caso B, quatro das funções foram mais rápidas no Matlab e duas das funções foram mais rápidas no *software* R. Observou-se, também, que com o aumento das quantidades de dados, maiores foram as diferenças entre os *softwares*; ou seja, com o aumento da quantidade de dados, maior foi a diferença entre o tempo de processamento das funções.

3 Considerações finais

O presente trabalho comparou os *softwares* Matlab e R. Ambos são linguagens de programação, utilizados para criação de rotinas matemáticas; tais ferramentas são de fácil utilização e possuem recursos de interface gráfica. Após comparações, verificou-se que o Matlab tem um alto custo de aquisição e o software R não é tão rápido, quando comparado a outras linguagens.

Subsequentemente, foram estudados quais os recursos disponíveis e mais utilizados na estatística para analisar um grupo de dados. A estatística descritiva é a divisão que faz esta análise de dados; nela, encontramos as medidas de posição (por exemplo, o cálculo de média, mediana, valor máximo e valor mínimo) e as medidas de dispersão (por exemplo, desvio padrão e variância).

Na sequência, criaram-se rotinas nas duas linguagens, para executar estas seis técnicas para duas populações de dados. Durante a execução dos testes, houve um problema de uso extremo da memória RAM, em que o software Matlab fez o gerenciamento do uso da memória de forma automática, enquanto o software R interrompeu a execução e precisou de adaptações no roteiro de cálculo, para expandir o uso da memória.

Conforme explicado, este bloqueio de memória tem lados positivos e negativos; ao realizar o gerenciamento sozinho, o *software* utiliza toda a memória RAM, fazendo com que o usuário perceba este uso excessivo e não consiga realizar outras tarefar de modo normal (e.g. demora para utilizar outros programas). Entretanto, quando *software* não gerencia sozinho o uso da memória, a rotina é interrompida, o que exige do desenvolvedor do roteiro conhecimentos extras para liberar este uso.

Quatro das seis técnicas analisadas (média, mediana, valor máximo e valor mínimo) apresentaram valores de tempo execução inferiores no *software* Matlab e com o aumento da população, as diferenças entre o tempo de processamento aumentou. Já nas outras duas técnicas (desvio padrão e variância), o tempo de processamento foi menor no *software* R, sendo que a diferença entre os softwares também cresce com o aumento da população.

Espera-se que este trabalho sirva de base para trabalhos futuros. Recomendamos que outras funções estatísticas e outros tamanhos de população de dados sejam estudados, com vistas a comparar melhor as linguagens e confirmar as tendências encontradas neste estudo.

Referências

BARROS, Anna Carolina *et al*. **Análise de séries temporais em R:** curso introdutório. Rio de Janeiro: Elsevier; FGV IBRE, 2018.

BECKER, João L. **Estatística básica:** transformando dados em informação. Porto Alegre: Bookman, 2015.

CHAPMAN, Stephen. J. **Programação em MATLAB para engenheiros**. 5. ed. São Paulo: Cengage Learning, 2016.

COSTA NETO, Pedro L. de O. **Estatística**. 3. ed. São Paulo: Edgard Blücher, 2002.

FARIA, Pedro D.; PARGA, **João 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: 03 ago. 2021.

GILAT, Amos. **MATLAB com aplicações em engenharia**. 4. ed. Porto Alegre: Bookman, 2012.

OLIVEIRA, P. F. de; GUERRA, S.; MCDONNELL, R. **Ciência de Dados com R:** Introdução. Brasília: IBPAD, 2018.

SILVIA, Juliane S. F. da; GRAMS, Ana L. B.; SILVEIRA, J. F. **Estatística**. Porto Alegre: SAGAH, 2018.

VIRGILLITO, Salvatore B. **Estatística aplicada**. 1. ed. São Paulo: Saraiva, 2017.