Meta-Análise de Artigos Científicos Segundo Critérios Estatísticos: Um estudo de caso no WSCAD 2018 a 2021

Bruno L. S. Rech[1], **Fabiano C. Dicheti**[1], **Gustavo R. Malacarne**[1], **Rodrigo S. Numberg**[1], **Thiago S. Elias** [1]

[1]Programa de Pós-Graduação em Ciência da Computação (PPGComp)
Universidade Estadual do Oeste do Paraná (UNIOESTE)
R. Universitária, 1619 - Universitário Cascavel - PR - Brasil - CEP: 85819-110

{bruno.rech, fabiano.dicheti, gustavo.malacarne}@unioeste.br

Abstract. *Statistical methods use data to understand a problem. They can be used for collecting, organizing, analyzing and interpreting data. In the scientific environment, such methods allow conclusions to be drawn about the characteristics of the sources from which the data were taken, in order to better understand the situations. This article presents the systematization of the results of the meta-analysis of publications from the last 4 editions (2018, 2019, 2020 and 2021) of WSCAD according to the categories: statistics, metrics and tests. For that, we manually extracted text from PDF files for the mining of statistical terms. From the sample of 70 publications analyzed, the search words were divided into three large groups: statistical terms, metrics and tests.*

Resumo. *Métodos estatísticos utilizam dados para compreender um problema. Podem ser empregados para coleta, organização, análise e interpretação de dados. No meio científico, tais métodos permitem tirar conclusões sobre as características das fontes em que os dados foram retirados, para melhor compreender as situações. Este artigo apresenta a sistematização dos resultados da meta-analise das publicações das 4 últimas edições (2018, 2019, 2020 e 2021) do WSCAD segundo as categorias: estatística, métricas e testes. Para tanto, realizou-se a extração manual de textos de arquivos PDF para a mineração de termos estatísticos. Da amostra de 70 publicações analisadas, as palavras de busca foram divididas em três grandes grupos: termos estatísticos, métricas e testes.*

1. Introdução

A estatística, desde seu início até os dias de hoje, passou de uma forma de sofisticação do processo de pesquisa, para se tornar um requisito básico, garantidor da confiabilidade dos processos e dos resultados de qualquer estudo científico, essencial na produção e difusão do conhecimento.

Uma forma de se aferir o emprego de ferramentas essenciais de estatística, é a observação e quantificação do vocabulário utilizado por pesquisadores, sendo a contagem do jargão próprio, um tipo de análise que assenta no positivismo lógico com referências ao conjunto de métodos quantitativos utilizados na análise e descrição de um fenômeno.

Em computação, a análise de desempenho é uma etapa fundamental na concepção de uma técnica ou algoritmo. A avaliação sob diversas métricas permite que pesquisadores utilizem de meios dispostos convenientemente para atestar o desempenho da solução.

Através da estatística, um conjunto de equações matemáticas podem ser usadas para analisar dados e determinar qual é a influência que cada fator apresenta no resultado final.

Embora aspectos como legibilidade, simplicidade e modularidade de uma solução sejam importantes para a sua manutenibilidade, o desempenho de uma solução é muito relevante para a sua adoção. Sendo assim, a análise de consistência e coerência dos dados é de suma importância para associar confiabilidade aos resultados apresentados e então levar a obtenção da inferência de comportamentos, interpretações sobre os resultados coletados e as conclusões obtidas.

Neste trabalho é realizada uma meta-análise qualitativa, por meio de processo de mineração de dados, dos artigos dos últimos 4 anos do Simpósio de Sistemas Computacionais de Alto Desempenho (WSCAD) para extrair a síntese sobre os métodos estatísticos e as métricas utilizadas nestas publicações. O objetivo do trabalho é fazer um levantamento de como o estudo estatístico esta sendo caracterizado nos artigos deste simpósio, com vistas a trazer indicativos se houve mudanças significativas das submissões realizadas em comparação aos anos anteriores.

Este trabalho está dividido em 6 seções. A Seção 2 apresenta os trabalhos relacionados ao estudo apresentado neste artigo. Na Seção 3 são apresentados os critérios de análise quantitativa. A Seção 4 apresenta a discussão sobre a metodologia utilizada. Os resultados da coleta realizada são discutidos na Seção 5 e a Seção 6 conclui o trabalho.

2. Trabalhos Relacionados

Em [Tichy et al. 1995] os autores buscaram artigos de pesquisa em ciência da computação, do ano de 1993, para determinar se os pesquisadores apoiavam seus resultados com avaliação experimental. Como conclusão, foi possível constatar que há uma falta considerável de apoio para os resultados, sendo que 40% da amostra não possuía nenhuma avaliação. Posteriormente, em 2005, os autores de [Wainer 2007], replicaram a pesquisa de Tichy com 147 artigos publicados, menos da metade do número da pesquisa original. No entanto, verificaram que 33% dos artigos ainda encontravam-se na mesma situação.

Estudos mais recentes também enfatizam a importância da experimentação e a falta dela em várias áreas da ciência da computação. Por exemplo, para os autores de [Andujar et al. 2012], a experimentação deve ser melhor compreendida e apreciada como uma metodologia chave em Informática. As ideias básicas do "método experimental" devem ser incluídas no aprendizado de Informática, como também em outros currículos de ciências. Mais tarde, Tedre e Moisseinen buscaram responder em [Tedre e Moisseinen 2014] o que os cientistas da computação querem dizer quando falam sobre experimentos em ciência da computação.

Em 2015, [Adler et al. 2015] investigou 183 artigos do IPIN (*International Conference on Indoor Positioning and Indoor Navigation*). A partir disso, descobriram que 95% de todos os artigos relacionados ao IPIN realizaram algum tipo de avaliação. A experimentação física desempenha um papel importante e é usada em 77% de todos os artigos relacionados. No entanto, embora o uso da avaliação seja alta, a qualidade da descrição dos métodos é fraca, concluem os autores.

Considerando os resultados dos trabalhos mencionados acima, o objetivo deste

trabalho é complementar e atualizar a pesquisa de [Osorio 2018]. Nessa obra, as publicações do Simpósio de Sistemas Computacionais de Alto Desempenho (WSCAD) foram avaliadas quanto a forma como são descritas em seus artigos as análises estatísticas de desempenho de maneira que possa ser atestado o ganho, se houver. Além disso, os autores analisaram uma amostra total de 426 artigos de todas as edições do WSCAD. Da amostra analisada, 398 publicações fizeram referência a pelo menos um dos termos pesquisados, correspondendo a 93% do total.

3. Critérios de Análise Quantitativa

Considerando a natureza comparativa deste trabalho, os artigos publicados no evento WSCAD até a publicação de [Osorio 2018], estão sendo comparados com as publicações no mesmo evento nos três anos subsequentes. Pode-se afirmar que a presente pesquisa trata-se de uma investigação causal comparativa. Segundo [Lakatos e Marconi 2003] uma pesquisa causal comparativa busca a descoberta de possíveis causas e efeitos, um padrão de comportamento, buscando explicações para semelhanças ou divergências. Sendo assim, os critérios analíticos considerados foram os seguintes:

- Frequência absoluta do uso de termos relativos à estatística em publicações do WSCAD (delimitadas na seção 4);
- Mudanças no padrão de escrita dos trabalhos publicados no WSCAD determinados pelas medias e desvios-padrões;
- Analise da variância dos grupos de palavras;
- Correlação entre a ocorrência concomitante de termos de cunho estatístico;

4. Metodologia

O desenvolvimento do presente trabalho, seguiu os seguintes atos contínuos:

- Foram minerados dados do WSCAD dos anos de 2000 a 2017 que já se encontravam tratados em repositório público[1], conforme [Osorio 2018];
- Foram minerados dados do WSCAD dos anos de 2018 a 2021;
- Foi feito o tratamento dos arquivos de 2018 a 2021;
- Foram tratadas e conferidas inconsistências de dados, manualmente;
- Os dados minerados, assim que tratados foram salvos no formato de arquivo com valores separados por vírgulas (CSV);
- Os arquivos CSV foram processados nas linguagens de programação Python e R;
- Foram feitas as quantificações comparativas ao trabalho de [Osorio 2018] ;
- Foi calculada, via software R, a analise de variância para três grupos de palavras em função do ano de observação;
- Adicionalmente foi verificado se há correlação entre os dados observados;
- Todos os arquivos de código fonte e CSV foram salvos em repositório público no github[2].

Os dados dos trabalhos publicados nos Anais do WSCAD entre os anos de 2000 a 2017 forma extraídos do artigo publicado por [Osorio 2018], já para os dados dos trabalhos publicados dos Anais do WSCAD entre os anos de 2018 a 2021, os trabalhos foram

[1]https://github.com/alessanderosorio/SMPE-UFRGS
[2]https://github.com/brunoluizs/ADIMS

salvos localmente, numerados e separados por pasta segundo o ano de publicação. Utilizando os mesmos critérios adotados por [Osorio 2018], trabalhos no formato de resumo ou escritos em língua estrangeira foram eliminados do processamento, dessa forma, do total de 136 artigos, restaram 70 publicações para análise.

A extração automatizada, de textos de arquivos PDF, utilizando pacotes Python, se mostrou mais complicada do que o antecipado. Os testes foram efetuados utilizando-se 3 pacotes diferentes, a saber PyPDF2[3], pdfpumbler[4] e PyMuPDF[5], cada um apresentando uma deficiência diferente na leitura de caracteres acentuados. O PyPDF2 inseria espaços, em número variável, entre caracteres acentuados além de deixar outros acentos "soltos" no restante da linha. O pdfpubler, em sua configuração padrão, falhava ao reconhecer o espaçamento entre as palavras. O PyMuPDF, das soluções testadas, foi a que apresentou o melhor desempenho, apesar de não acentuar corretamente as palavras, não inseria outros artefatos no texto.

Assim, seria possível remover a acentuação também dos termos de busca e comparar com o texto não acentuado extraído do artigo, porém isso introduziria outros problemas, por exemplo, na busca pelo termo média, em sua versão não acentuada, gerou correspondências com palavras como imediatamente e "media". Apesar de não parecer um problema, "media" não se referia apenas ao termo estatístico média mas também à tradução, em língua inglesa, da palavra mídia. Se buscássemos uma combinação exata, para evitar correspondência dos termos com palavra maiores e não relacionadas, ainda incorreríamos em problemas como o da correspondência com a palavra inglesa "media", além disso, não seriam geradas correspondências para os plurais dos termos de busca.

Dessa forma, como o número de artigos selecionados era relativamente pequeno, optou-se por conduzir um processo de busca manual, por meio do qual foram realizadas buscas, nos textos dos artigos em arquivos PDF, pelos termos utilizados em análise estatística bem como métricas e testes utilizados para aferição de resultados, assim categorizados neste artigo. Os termos de busca são apresentados nas Tabelas 1, 2 e 3.

A Tabela 1, contém as palavras de busca referentes aos "termos" estatísticos, bem como suas respectivas siglas. Correspondem à palavras empregadas na rotina estatística.

Na Tabela 2, encontram-se palavras de uso mais restrito, agrupadas sob a referência de "métricas". Tratam-se de palavras empregadas em medidas estatísticas de desempenho no contexto da computação.

Já na Tabela 3, estão os termos de busca e siglas relativos aos testes estatísticos, relacionados à validação estatística, e agrupados sob o nome "testes", que é a palavra de referência no decorrer do trabalho.

5. Discussão

Na presente sessão, apresentam-se as afirmações de comparação entre os trabalhos publicados entre os anos de 2000 a 2017 com os dos anos de 2018 a 2021 no WSCAD. Para possibilitar a comparação entre os anos foi utilizada a proporção de artigos com citações

[3]https://pypi.org/project/PyPDF2/

[4]https://pypi.org/project/pdfplumber/0.1.2/

[5]https://pymupdf.readthedocs.io/en/latest/

Tabela 1. Termos estatísticos selecionados para coleta de dados

Descrição	Chave de Pesquisa
Amostra (AM)	amostra
Desvio Padrão (DP)	desvio padrão
Distribuição Normal (DN)	distribuição normal
Frequência (FR)	frequência OR frequência
Gaussiana (GA)	gaussiana
Intervalo de Confiança (IC)	intervalo de confiança
Média (ME)	média
Num. Execuções (NE)	número de execuções
Num. Iterações (NI)	numero de iterações
Teste/Experimento (TE)	teste OR experimento OR simulação
Variância (VR)	variância

Fonte: [Osorio 2018]

Tabela 2. Métricas selecionadas para coleta de dados

Discrição	Chave de Pesquisa
Bandwidth (BW)	“bandwidth” OR “largura de banda”
BPS (BP)	“bits por segundo” OR “bps”
Capacidade Nominal (CN)	“nominal capacity” OR “capacidade nominal”
Capacidade Utilizavel (CU)	“usable capacity” OR “capacidade utilizável”
Confiabilidade (CO)	Reliability OR Confiabilidade
Cost/Performance Ratio (CP)	“cost ratio” OR “performance ratio”
Disponibilidade (DI)	availability OR disponibilidade
Downtime/Uptime (DU)	downtime OR uptime
Eficiência/Acurácia (EA)	eficiência OR eficácia OR accuracy
Fator de Estiramento (FE)	“strech factor” OR “fator de estriamento”
Tempo Ocioso (TO)	“Idle time” OR “tempo ocisoso”
MFLOPS (MF)	MFLOPS
MIPS (MI)	MIPS
MTTF (MT)	MTTF
PPS (PP)	PPS
Speed up (SU)	“speedup” OR “speed-up” OR “speed up”
Tempo de Reaçao (TR)	“reaction time” OR “tempo de reação”
TPS (TP)	TPS

Fonte: [Osorio 2018]

Tabela 3. Testes estatísticos selecionados para coleta de dados

Descrição	Chave de Pesquisa
P-Valor (PV)	"p-valor" OR "p-value" OR "valor p"
Teste ANOVA (AN)	anova
Teste Chi-quadrado (CH)	"chi-quadrado" OR "qui-quadrado"
Teste de Wilcoxon (TC)	wilcoxon signed-rank
Teste Exato de Fisher (FI)	"teste exato de fisher" OR "fisher"
Teste Kruskal-Wallis (KR)	kruskal-wallis
Teste T (TT)	"teste t" OR "teste-t" OR "teste de student" OR "Student"
Teste U (TU)	"teste U" OR "mann-whitney" OR "wilcoxon rank-sum"

Fonte: [Osorio 2018]

em relação total de artigos no ano e não o número de artigos, uma vez que o número de artigos por ano não é o mesmo.

Após a tabulação, os dados foram sumarizados por ano conforme a categoria do termo. As Tabelas 4, 5, 6, 7, 8 e 9 apresentam os resultados de termos estatísticos, métricas e testes. Importante ressaltar que, somente os termos que obtiveram resultados foram sumarizados, e em relação às tabelas 4 e 5, os resultados do termo frequência foram excluídos devido ao viés que os resultados apresentaram, pois constantemente se referia à grandeza física (oscilações por segundo em Hz) e não ao termo estatístico. Os resultados dos termos distribuição normal e gaussiana foram agregados por se tratarem do mesmo objeto.

Pela tabela 4, constata-se que o total de ocorrência de termos estatísticos na primeira janela de tempo (2000-2017), apresentados na coluna "n" tem média de 29,34 utilizações por ano, com um desvio padrão de 14,28 entre os anos. Na segunda janela (2018-2021), obtém-se, pela tabela 5, uma média consideravelmente maior, sendo verificadas 40 aparições de termos estatísticos em média para cada ano, com uma menor oscilação, denotada por um desvio padrão de 11,94 pontos. Cabe ressalvar que o número total de observações para a segunda janela de tempo é consideravelmente menor, devido ao menor período temporal e consequentemente ao menor número de artigos analisados.

Comparando-se as tabelas 4 e 5, verifica-se que não é possível atribuir a mudança de comportamento à publicação de [Osorio 2018], visto que no ano de 2005 ocorreu o maior "n" de 58 ocorrências, seguido pelos anos de 2004 e 2019 ambos com 53 termos. A partir do ano de 2018, verifica-se uma troca do emprego maioritário de "IC" e "NI" por "ME" e "TE", mesmo que o valor de "n" não tenha sofrido alterações.

Tabela 4. Citações de Termos Estatísticos de 2000 a 2017

Ano	n	AM	DP	DN	GA	IC	ME	NE	NI	TE	VR
2000	7	-	-	-	-	2	-	1	4	-	
2001	34	4	1	-	-	10	1	1	16	1	
2002	35	1	2	-	-	9	1	2	20	-	
2003	40	2	3	-	-	11	1	2	21	-	
2004	53	-	4	-	-	19	1	4	24	1	
2005	58	3	6	-	1	17		2	28	1	
2006	38	-	2	1	-	7	1	3	24		
2007	28	-	1	-	-	7	1	1	17	1	
2008	41	1	3	1	1	12	-	1	22	-	
2009	18	-	-		-	4	-	-	14	-	
2010	19	2	-	1	-	5	-	-	11	-	
2011	8	-	-	-	-	3	-	-	5	-	
2012	33	-	1	1	1	7	1	-	21	1	
2013	31	2	2	2		11	-	-	14	-	
2014	33	-	-	-	-	-	3	2	27	1	
2015	16	-	1	1	1	1	-	1	11	-	
2016	23	1		1	1	2	-	1	16	1	
2017	13	2	1	-	-	2	-	-	8	-	

Fonte: [Osorio 2018]

Tabela 5. Citações de Termos Estatísticos de 2018 a 2021

Ano	n	AM	DP	DN	IC	ME	NE	NI	TE	VR
2018	45	3	4	-	2	15	-	2	19	-
2019	53	5	5	2	-	14	-	3	22	2
2020	37	2	7	1	-	11	-	-	16	-
2021	25	3	2	-	-	7	-	1	12	-

Fonte: do autor

Sobre o emprego de palavras que remetam a métricas estatísticas, obtém-se da Tabela 6 uma média 25,56 ocorrências por ano, entre 2000 e 2017, contra 30,25 observações/ano no período de 2018 a 2021, decorrente da Tabela 7. Os desvios-padrão para "métricas" não apresentaram diferenças significativas entre as duas janelas de tempo, sendo observados valores de 8,66 e 6,95 para o primeiro e segundo intervalo de tempos respectivamente.

Os maiores valores de observação de métricas se deu em 2003 e 2004 ("n" = 38) assim como em 2019 com o mesmo número, em ambos os intervalos as ocorrências predominantes foram: "EA", "CO" e "SU".

Tabela 6. Citações de Métricas de 2000 a 2017

Ano	n	Disp.	BW	BP	CP	DU	EA	ID	MF	MI	PP	CO	SU
2000	17	5	1	-	-	-	6	1	-	1	-	1	2
2001	29	7	3	-	-	-	9	-	-	1	-	2	7
2002	31	5	1	-	-	-	12	-	-	3	-	4	6
2003	38	11	2	-	-	-	8	-	3	2	-	5	7
2004	38	10		-	-	-	14	-	-	1	-	6	7
2005	36	10	1	-	-	-	9	1	-	1	-	4	10
2006	15	3	2	-	-	-	4	-	-	1	-	-	5
2007	17	3	1	-	-	-	3	1	-	2	-	-	7
2008	33	6	5	-	-	-	7	-	1	3	-	3	8
2009	23	6	1	1	-	-	4	1	1	1	-	1	7
2010	21	5	-	-	-	1	2	1	-	5	-	2	5
2011	8	3	-	-	-	-	1	-	-	1	-	-	3
2012	24	9	2	1	-	-	4	-	-	4	-	-	4
2013	31	5	-	1	-	3	5	1	-	3	-	2	11
2014	32	6	-	-	-	-	5	1	-	3	-	1	16
2015	17	1	-	-	-	-	3	1	1	2	1	-	8
2016	28	6	3	-	1	-	1	4	1	1	-	1	10
2017	22	5	1	-	-	1	-	2	-	1	-	1	11

Fonte: [Osorio 2018]

Tabela 7. Citações de Métricas de 2018 a 2021

Ano	n	Disp.	BW	BP	CP	DU	EA	ID	MF	MI	PP	CO	SU
2018	34	5	3	-	1	-	12	-	-	2	-	3	8
2019	38	4	2	1	-	-	14	-	-	-	-	3	14
2020	26	3	2	-	-	-	8	-	1	1	-	4	7
2021	23	2	4	-	-	-	10	-	-	-	-	2	5

Fonte: do autor

Tanto para o período de 2000 a 2017, quanto de 2018 a 2019, os termos relativos a testes estatísticos são os menos frequentes, com média de ocorrência de apenas 0,64 vezes ao ano para os anos anteriores a 2018 expressos na Tabela 8, e média de 2 observações anuais na Tabela 9, que representa o intervalo de 2018 a 2021 com desvios-padrão de 1,04 e 1,63 respectivamente, para a coluna "n" de ambas as tabelas (8 e 9).

Tabela 8. Citações de Testes de 2000 a 2017

Ano	n	PV	TC	KR	TT
2001	1	1	-	-	
2007	4	1	-	-	3
2008	1	-	-	-	1
2009	1	-	-	-	1
2012	2	-	-	-	2
2015	1	-	-	-	1
2016	1	-	-	-	1

Fonte: [Osorio 2018]

Tabela 9. Citações de Testes de 2018 a 2021

Ano	n	PV	TC	KR	TT
2018	2	-	-	1	1
2019	4	1	1		2
2020	2	-	-	-	2
2021	0	-	-	-	-

Fonte: do autor

A Tabela 10 apresenta a sumarização dos resultados das Tabelas 4, 6 e 8, enquanto a Tabela 11 apresenta a sumarização dos resultados das Tabelas 5, 7 e 9.

Nas tabelas 10 e 11 é possível observar o número total de artigos no ano (coluna "n"), o número de artigos que contém pelo menos uma ocorrência do termo (colunas "Art."), bem como o número de citações aos termos (colunas "Cit."). Os dados estão agrupados por ano e pelas categorias dos termos. Por exemplo, no ano de 2000 foram analisados 9 trabalhos dos quais 5 apresentaram termos do grupo "estatística" e 7 apresentaram palavras do grupo "métricas".

Tabela 10. Distribuição de citações de 2010 a 2017 por tipo do termo e ano

Ano	n	Estatistica		Metricas		Teste	
		Art.	Cit.	Art.	Cit.	Art.	Cit.
2000	9	5	7	7	17	-	-
2001	23	20	34	16	29	1	1
2002	27	22	35	19	31	-	-
2003	32	28	40	23	38	-	-
2004	33	30	53	21	38	-	-
2005	34	31	58	26	36	-	-
2006	28	24	38	12	15	-	-
2007	21	17	28	12	17	2	4
2008	28	24	41	17	33	1	1
2009	23	14	18	16	23	1	1
2010	20	12	19	16	21	-	-
2011	6	5	8	5	8	-	-
2012	28	22	33	17	24	2	2
2013	20	17	31	16	31	-	-
2014	39	29	33	23	32	-	-
2015	15	12	16	11	17	1	1
2016	21	16	23	17	28	1	1
2017	19	9	13	15	22	-	-

Fonte: [Osorio 2018]

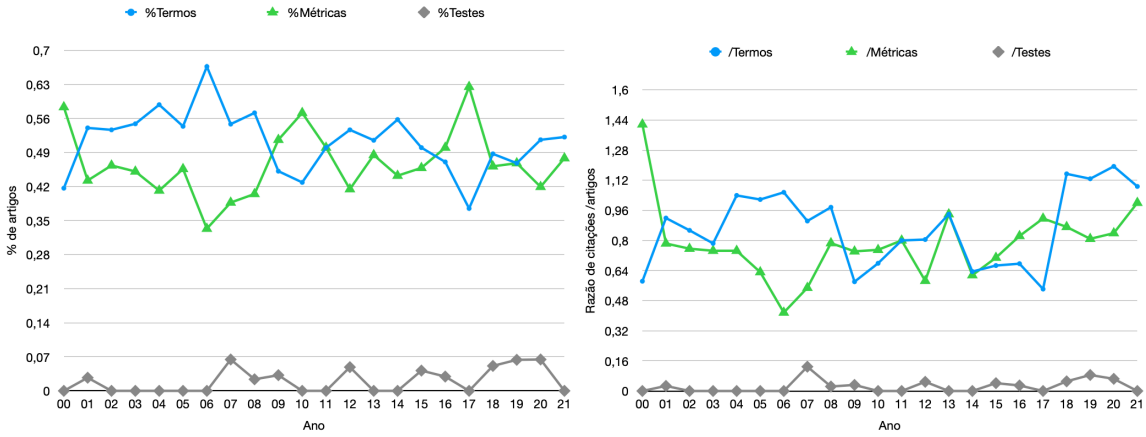
Tabela 11. Distribuição de citações de 2018 a 2021 por tipo do termo e ano

Ano	n	Estatistica		Metricas		Teste	
		Art.	Cit.	Art.	Cit.	Art.	Cit.
2018	19	19	45	18	34	2	2
2019	23	22	53	22	38	3	4
2020	16	16	37	13	26	2	2
2021	12	12	25	11	23	-	-

Fonte: do autor

Devido à densidade dos dados, foi elaborada a Figura 1. Na figura 1(a), tem-se o percentual de artigos com citações para cada um dos três grupos de palavras-chave, em relação ao número de artigos analisados por ano. Pode-se observar que não há qualquer padrão de linearidade ou mudança a partir do ano de 2018. A volatilidade visível na figura é expressa por desvio padrão de 6,33% para "termos", 6,65% para "métricas" e uma volatilidade de 2,51 pontos percentuais para "testes".

As mesmas considerações feitas para os valores percentuais, são válidas quando observamos a Figura 1(b), que contém a razão entre o número de ocorrência de cada um dos grupos de palavras e o volume de artigos para os respectivos anos.



(a) Proporção de artigos com citações por ano

(b) Razão de citações por artigo ano

Figura 1. Gráficos dos Resultados Encontrados

Na Figura 2, estão plotados os valores totais de ocorrência de termos dos três grupos, que deflagra uma distribuição volátil, sem haver qualquer tipo de tendencia aparente, embora haja no gráfico um aclive após o ano de 2018, o mesmo nível de valor pode ser observado entre os anos de 2004 e 2005. Observa-se que no ano de 2021 o valor de ocorrências atingiu nível significativamente baixo, ficando abaixo da média geral do gráfico de 58,54 aparições por ano, com um desvio padrão total de 21,85 pontos desde 2000 até 2021.

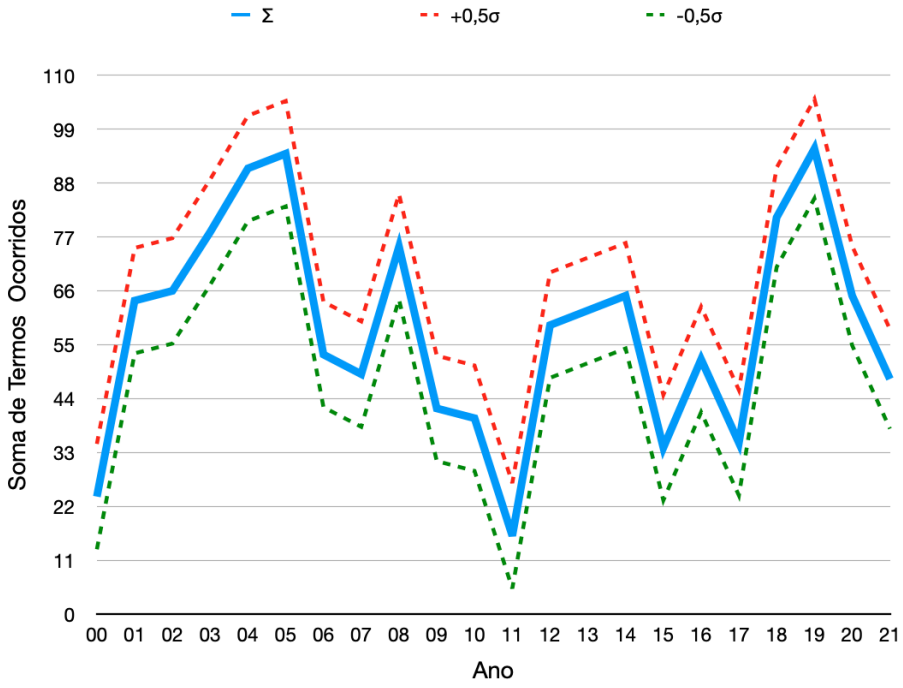


Figura 2. Soma de "n" Termos por ano, de 2000 a 2021

5.1. Análise de Variância dos Termos - ANOVA

Visto que as tabelas 10 e 11 sumarizam os valores totais relativos às ocorrências de termos estatísticos, foram performadas análises de variância (ANOVA) para cada um dos três grupos de palavras. Para tanto, foi considerado um alfa de 0,05 e as seguintes hipóteses:

$$H0 : \textit{após a publicação de } [\text{Osorio 2018}] \textit{ não houve mudança no padrão de uso de termos estatísticos em publicações do } WSCAD$$

$$H1 : \textit{após a publicação de } [\text{Osorio 2018}] \textit{ houve mudança no padrão de uso de termos estatísticos em publicações do } WSCAD$$

É valido ressaltar, que mesmo o ano de divisão entre as tabelas apresentadas seja o de 2018, pois [Osorio 2018] analisou trabalhos até o ano de 2017, para fins de analise de variância, o ano de 2018 foi considerado como "anterior" pois caso a hipótese de que o trabalho publicado em 2018 decorresse em mudança de comportamento, este comportamento só seria observado a partir do ano de 2019.

Sendo assim, o dataframe foi rotulado em duas classes, da seguinte maneira:

- situação = "anterior", para os trabalhos de 2000 até 2018
- situação = "novo", para os trabalhos de 2019 até 2021

Abaixo, pode-se visualizar os valores obtidos de analise da variâncias para cada grupo de palavras, na linguagem de programação "R":

```
alfa <- 5/100 # 5% = 0.05                                                    1
k <- 2 # tratamentos: anterior/novo                                          2
n <- length(df_dados$situacao) # repeticoes: 22 - 2000 a 2001               3
f_critico <- qf(1 - alfa, df1 = k - 1, df2 = n - k)                          4
f_critico                                                                    5
## [1] 4.351244                                                            6
                                                                             7
aov_metricas <- aov(metricas ~ situacao, df_dados)                           8
aov_testes <- aov(testes ~ situacao, df_dados)                               9
aov_termos <- aov(termos ~ situacao, df_dados)                               10
                                                                            11
summary(aov_metricas)                                                       12
##              Df Sum Sq Mean Sq F value Pr(>F)                            13
## situacao      1    1.8   1.762   0.062  0.806                            14
## Residuals    20  567.2  28.360                                           15
                                                                            16
summary(aov_testes)                                                         17
##              Df Sum Sq Mean Sq F value Pr(>F)                            18
## situacao      1  3.065  3.0654   4.008  0.059 .                          19
## Residuals    20 15.298  0.7649                                           20
                                                                            21
summary(aov_termos)                                                         22
##              Df Sum Sq Mean Sq F value Pr(>F)                            23
## situacao      1   11.1   11.10   0.186  0.671                            24
## Residuals    20 1196.4   59.82                                           25
```

Observa-se no bloco de códigos acima, que os p-valores (colunas "Pr(>F)") para as três categorias de termos é maior que o valor de alpha adotado (0,05) e portanto, aceita-se a hipótese nula ($H0$), desta forma não se verificou diferença entre as situações anterior e posterior à publicação de [Osorio 2018], para nenhum dos três grupos de palavras analisados.

O grupo "métricas", na linha 12 do código acima, resultou em uma estatística de 0,806 maior que o valor de 0,05, vide ultima coluna da linha 14.

Para o grupo "testes", na linha 17, observou-se uma estatística muito próxima de 0,05, mas o valor disposto na linha 19, de 0,059 aponta para a validade da hipótese nula ($H0$).

Do mesmo modo, para o terceiro grupo "termos" localizado na linha 22 do código acima, observou-se uma estatística de 0.671 na linha 24, também corroborando para $H0$.

5.2. Análise de Correlação Entre os Termos

Com o intuito de verificar se há uma uma relação de causa-efeito que justifique a união de dois termos estatísticos, foi calculado o índice de correlação de ϕk , desenvolvido por [Baak et al. 2019].

De forma simplificada, pode-se definir o algoritmo para obtenção do coeficiente de correlação ϕk da seguinte maneira:

• 1. Se as variáveis forem intervalares não categorizadas:

então: aplicar uma categorização a cada uma. Um dimensionamento razoável é geralmente específico do caso de uso, entretanto, por padrão, tomam-se 10 espaços de memória[6] por variável;

• 2. Alocar e preencher uma tabela de contingência para cada par de variáveis, contendo "N" registros de "r" linhas e "k" colunas;

• 3. Calcular o χ^2 (chi-quadrado) de Pearson[7], para a estatística de teste das variáveis qualitativas e nominais. Conforme a Equação 2, se obtém as estimativas de frequência estatisticamente dependentes;

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{1}$$

Onde:

O , é o número de observações;

E , é a frequência esperada (teórica);

χ^2 , é a estatística de teste cumulativa de Pearson, que assintoticamente se aproxima da distribuição chi-quadrado;

• 4. Interpretar o χ^2 como proveniente de uma distribuição normal bivariada, sem estatística, de acordo com a Equação 3[8]

$$X^2_{b.n.}(\rho, N, r, k) = \chi^2_{ped} + \left\{ \frac{\chi^2_{max}(N, r, k) - \chi^2_{ped}}{\chi^2_{b.n.}(1, N, r, K)} \right\} \chi^2_{b.n.}(\rho, N, r, K) \tag{2}$$

Onde: Se $\chi^2 < \chi^2_{ped}$ defina ρ como zero;

Senão, com N, r, k corrigidos, inverter a fincão $\chi^2_{b.n.}$ usando o método de Brent[9] e resolver numericamente para ρ entre [0,1];

A solução para ρ define o coeficiente de correlação ϕk

A Figura 3, apresenta um mapa térmico das correlações obtidas pelo índice de Phik. Em primeira análise, pode-se enumerar vários termos que possivelmente possuam correlação positiva com os demais, entretanto alguns termos como "Cost ou Performance Ratio" que apresentou índice de correlação de 0,686 com os termos "bps, MFLOPS, p-valor, e Teste de Wilcoxon", bem como chegou ao índice 1,00 (totalmente correlacionado)

[6]Para fins de calculo da integral ou da área de distribuição, considera-se este dimensionamento como as "bins", distância entre as variáveis no eixo X(abscissas).

[7]O teste qui-quadrado de Pearson (ou teste chi-quadrado de Pearson) é um teste estatístico aplicado a dados categóricos para avaliar quão provável é que qualquer diferença observada aconteça ao acaso. É adequado para amostras não pareadas/emparelhadas.

[8]$b.n.$: leia-se distribuição bivariada normal.

[9]

$$Z = \sqrt{u - log\ u};\ u = -2log(p\sqrt{2\pi})$$

com o termo “disponibilidade”, não devem ser considerado significativos, pois, aparecem apenas uma vez entre os anos de 2018 e 2021, logo, não se pode afirmar qualquer tipo de causa e efeito deste resultado. O mesmo pode ser aplicado aos termos “MFLOPS” que ocorreu apenas em um trabalho em 2020,“Teste Kruskal-Wallis” e “Teste de Wilcoxon” presentes em apenas uma publicação dos anos de 2018 e 2019 respectivamente.

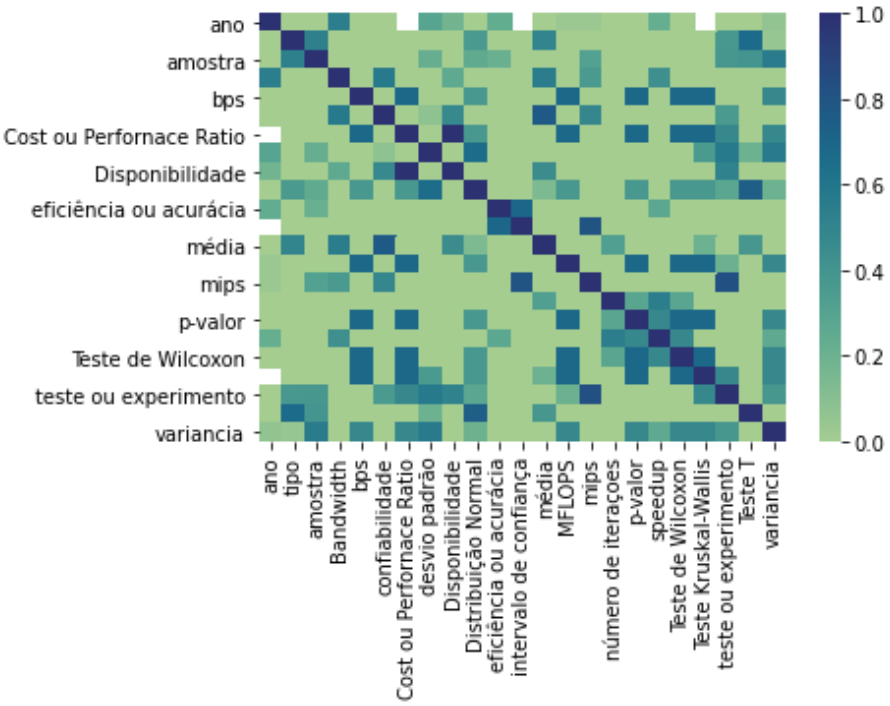


Figura 3. Mapa térmico de correlações entre as palavras

Feita a primeira análise com base na Figura 3, pode-se sumarizar os índices realmente significativos, isto é, de termos que aparecem em ao menos 3 publicações e que apresentem entre si correlações superiores a 0,70. Na Tabela 12 estão sintetizados 7 termos que apresentaram as seguintes correlações:

- Confiabilidade x Média, $\phi k = 0{,}775$;
- Distribuição Normal x Teste T, $\phi k = 0{,}750$;
- Mips x Intervalo de Confiança, $\phi k = 0{,}807$;
- Mips x Teste ou Experimento, $\phi k = 0{,}826$;
- Intervalo de Confiança x Teste ou Experimento, $\phi k = 0{,}000$.

Tabela 12. Correlações significativas 2018 a 2021

Termo	CO	DN	IC	ME	MI	TE	TT
Confiabilidade	1,00	0,000	0,000	0,775	0,488	0,360	0,000
Distribuição Normal	0,000	1,00	0,000	0,143	0,000	0,283	0,750
Intervalo de confiança	0,000	0,000	1,00	0,000	0,807	0,000	0,000
Média	0,775	0,143	0,000	1,00	0,000	0,000	0,384
Mips	0,488	0,000	0,807	0,000	1,00	0,826	0,000
Teste ou Experimento	0,360	0,283	0,000	0,000	0,826	1,00	0,000
Teste T	0,000	0,750	0,000	0,384	0,000	0,000	1,00

Fonte: do autor

6. Conclusão

Uma base adequada de afirmações qualificadas é fundamental para o desenvolvimento científico, neste contexto, o uso da estatística deve ser especifico e eficaz, fornecendo subsídios para a parametrização e reprodutibilidade de todos os tipos de experimentos, pois, limitações, ou falhas no rigor estatístico podem comprometer o alcance e a validade dos resultados de uma pesquisa.

Neste trabalho foi apresentado o resultado das pesquisa sobre os "termos estatísticos", "métricas" e "testes estatísticos". Em complemento ao trabalho de [Osorio 2018], foram analisadas 70 publicações adicionais, completas e em português, submetidas nas edições de 2018 a 2021 do WSCAD, e chegou-se as mesmas conclusões daquele, onde se evidenciou a carência do uso correto e metódico de estatística nas ciências da computação, sobretudo nas publicações do evento supracitado, esta deficiência pode ter fonte em diversos fatores, embora o estudo destes fatores não faça parte do escopo deste trabalho, podem-se enumerar que possivelmente alguns motivos para o pouco emprego correto da estatística sejam a pouca oferta de conteúdo de estatística nos cursos de graduação, falta de planejamento adequado nas pesquisas e fatores relacionados à falta de professores, cada vez mais escassos devido à falta de politicas públicas de valorização e reconhecimento.

Entre os anos de 2000 a 2017 foram feitas em média 55,5 menções a termos relativos à estatística, com um desvio padrão de 21,52 (Tabelas 4, 6 e 8). Nos anos seguintes, de 2018 a 2021, houve um incremento de 30,18%, com uma media de 72,25 menções por ano e com desvio padrão de 20,29 pontos. Os anos onde as publicações atingiram a maior contagem de aplicação de termos estatísticos foram 2004, 2005, 2018 e 2019 com 91, 94, 81 e 95 menções respectivamente, bem como, os piores anos de acordo com o mesmo critério foram 2000, 2011 e 2017 com 24, 16 e 35 palavras, nessa ordem.

Embora a ANOVA aponte para uma não existência de causa e efeito entre a publicação de [Osorio 2018] e o comportamento das publicações observadas, com relação às mudanças no comportamento do uso de termos entre os três grupos de palavras, podem ser feitas as seguintes observações: Antes de 2018, para o grupo "termos", as palavras chave: Num. de Iterações e Intervalo de Confiança eram as mais utilizadas, a partir deste ano, os termos mais utilizados são Média e Teste/Experimento. Para o grupo "métricas", observa-se independente da janela de tempo as palavras mais recorrentes são

Eficiência/Acurácia, Confiabilidade e Speed-Up. Para o grupo "testes" devido ao baixo índice de uso, não são possíveis considerações de usualidade.

Quando correlacionadas as palavras de busca, pode-se afirmar que há uma intersecção de ocorrência entre as palavras "Confiabilidade e Média", bem como, entre "Distribuição Normal e Teste T", que em trabalhos futuros, ou pra fins de refinamento, poderiam ser agrupadas.

Por fim, conclui-se que não houve qualquer relação de causa-efeito entre o comportamento das publicações do WSCAD e a publicação de [Osorio 2018], visto os scores da ANOVA dos três grupos de palavras, nem mesmo se pode estabelecer um padrão previsível de comportamento com o decorrer do tempo conforme exposto pela Figura 2.

Referências

Adler, S., Schmitt, S., Wolter, K., e Kyas, M. (2015). A survey of experimental evaluation in indoor localization research.

Andujar, C., Schiaffonati, V., Schreiber, F. A., Tanca, L., Tedre, M., van Hee, K., e van Leeuwen, J. (2012). The role and relevance of experimentation in informatics.

Baak, M., Koopman, R., Snoek, H., e Klous, S. (2019). A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics. *https://doi.org/10.48550/arXiv.1811.11440*.

Lakatos, E. M. e Marconi, M. d. A. (2003). *Fundamentos de metodologia científica*. Atlas.

Osorio, A. (2018). Meta-análise de artigos científicos segundo critérios estatísticos: Um estudo de caso no wscad. *WSCAD 2018 - XIX Simpósio de Sistemas Computacionais de Alto Desempenho*, 19.

Tedre, M. e Moisseinen, N. (2014). Experiments in computing: A survey. *TheScientificWorldJournal*, 2014:549398.

Tichy, W., Lukowicz, P., Prechelt, L., e Heinz, E. (1995). Experimental evaluation in computer science. *Journal of Systems and Software - JSS*.

Wainer, J. (2007). Métodos de pesquisa quantitativa e qualitativa para a ciência da computação. *Instituto de Computação – UNICAMP*.