Meta-Análise de Artigos Científicos Segundo Critérios Estatísticos: Um estudo de caso no WSCAD*

Alessander Osorio, Marina Dias, Gerson Geraldo H. Cavalheiro

[1]Programa de Pós-Graduação em Computação
Universidade Federal de Pelotas
Pelotas – RS – Brasil

{alessander.osorio,mldias,gerson.cavalheiro}@inf.ufpel.br

Abstract. *This paper presents the systematization of the results of the meta-analysis of the publications of the 18 editions of the WSCAD according to the categories: statistics, metrics and tests. The goal is not to invalidate the results, but to alert to how they are described and presented.From a sample of 426 publications, 93 % referred at least one of the terms searched, showing that there is some care in the demonstration of results, even inadequate or incomplete given the low occurrence of only 3 % of statistical tests confirmed by a second sample of 30 articles. It is necessary not only to focus on the development of the research object itself, but also on the results demonstrations.*

Resumo. *Este artigo apresenta a sistematização dos resultados da meta-análise das publicações das 18 edições do WSCAD segundo as categorias: estatística, métricas e testes. O objetivo não é invalidar os resultados, mas alertar para a forma como são descritos e apresentados. Da amostra de 426 publicações analisadas, 93% fizeram referência a pelo menos um dos termos pesquisados, mostrando que existe a preocupação na demonstração dos resultados, mesmo que inadequada ou incompleta dada a baixa ocorrência de apenas 3% de testes estatísticos confirmada por uma segunda amostra de 30 artigos. É preciso não apenas colocar foco no desenvolvimento do objeto de pesquisa em si, mas também na aferição e apresentação dos resultados.*

1. Introdução

A apresentação de uma nova técnica ou algoritmo usualmente é acompanhada de uma análise de desempenho. Deve-se atentar para que o estudo de desempenho seja realizado de forma a que o ganho, caso exista, possa ser atestado. Se observa que, em muitos casos, pesquisadores dedicam grande quantidade de tempo para execução dos experimentos mas, em contrapartida, limitam-se a apresentar dados de desempenho sem realizar um estudo estatístico que os valide.

O estudo estatístico a ser realizado, como considerado neste artigo, é aquele estudo em que é realizada a análise da consistência e coerência dos dados de desempenho obtidos. Esta etapa deve preceder a inferência de comportamentos, interpretações sobre os resultados coletados e as conclusões obtidas. Portanto, a efetiva realização do estudo

*O presente trabalho foi realizado com apoio do Programa Nacional de Cooperação Acadêmica da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES/Brasil.

estatístico permite associar confiabilidade aos resultados apresentados em qualquer documento de divulgação científica.

Neste trabalho é realizada uma meta-análise qualitativa, por meio de processo automatizado de mineração de dados, dos artigos dos últimos 18 anos do Simpósio de Sistemas Computacionais de Alto Desempenho (WSCAD) para extrair a síntese sobre os métodos estatísticos e métricas utilizados nestas publicações. O objetivo do trabalho é fazer um levantamento de como o estudo estatístico está sendo caracterizado nos artigos deste simpósio, com vistas a trazer indicativos de como qualificar suas submissões nos próximos anos.

Além de apresentar o levantamento da apresentação dos estudos estatísticos nos artigos do WSCAD, o presente artigo também contribui por caracterizar métodos estatísticos relevantes para avaliação de desempenho em processamento de alto desempenho e indicar técnicas estatísticas aplicáveis nesta área.

Este trabalho esta dividido em 6 seções. Na Seção 2 são caracterizados trabalhos relacionados ao estudo apresentado neste artigo. Os critérios de análise quantitativa são expostos na Seção 3 e a discussão sobre a metodologia utilizada é apresentada na Seção 4. Os resultados da coleta realizadas são discutidos na Seção 5. A Seção 6 conclui o trabalho.

2. Trabalhos Relacionados

Após uma pesquisa bibliográfica sobre o tema, foram encontrados trabalhos relacionados desenvolvidos ao longo dos anos. Estes trabalhos dizem respeito a outros trabalhos que também tiveram como objetivo a identificação de metodologias de validação de resultados em artigos científicos da grande área de Computação. Nesta seção, os trabalhos são organizados em ordem cronológica, de forma que a evolução das pesquisas ao longo dos anos possa ser visualizada e comparada com os resultados obtidos neste trabalho.

[Prechelt 1994] avaliou 190 artigos publicados em 1993 e 1994 e evidenciou que apenas 1/3 dos artigos não tinham comparação quantitativa com técnicas previamente conhecidas.[Tichy et al. 1995] analisou 400 artigos para determinar se cientistas da computação apoiam seus resultados com avaliação experimental. O estudo descobriu que 40% dos artigos não possuía qualquer tipo de avaliação. [Wainer et al. 2009] replicou a pesquisa de [Tichy et al. 1995] analisando 147 artigos publicados no ano de 2005 concluindo que 33% dos artigos encontram-se na mesma situação. O trabalho de [Tedre and Moisseinen 2014] reforça essa afirmação e cita como causa a falta de experiência da comunidade na correta análise dos dados a fim de produzir evidências estatísticas que comprovem os resultados.

Mais recentemente, o trabalho de [Adler et al. 2015] investigou 183 artigos do IPIN (International Conference on Indoor Positioning and Indoor Navigation) e concluiu que embora em muitas das publicações houvesse alguma preocupação na avaliação dos resultados, a qualidade da descrição dos métodos de análise era pobre. Apenas 35% relatam de maneira clara não só a metodologia do experimento em si, mas o que efetivamente os resultados querem estatisticamente dizer.

Considerando os resultados nos trabalhos mencionados acima, nosso objetivo é avaliar as publicações do Simpósio de Sistemas Computacionais de Alto Desempenho

(WSCAD) quanto à forma como são descritas em seus artigos as análises estatísticas de desempenho de maneira que possa ser atestado o ganho, se houver.

3. Critérios de Análise Quantitativa

A pesquisa em Ciência da Computação, envolve na maioria dos casos, o desenvolvimento de um modelo, aplicação, algoritmo ou sistema computacional novo [Wainer et al. 2007], que ainda não foi totalmente estudado à exaustão como objeto de pesquisa. Dentro deste processo, o objeto de pesquisa é comparado a seus pares para efetiva avaliação de desempenho da solução proposta. Esta avaliação deve ser feita pela análise quantitativa dos resultados sumarizados, obtidos pela utilização de *dados sintéticos* e técnicas estatísticas de comparação de conjuntos de medidas [Wainer et al. 2007, Bukh 1992].

Dados sintéticos, obtidos por *workloads*, *benchmarks*, simulações e competições, são classificados em três categorias. A primeira é utilizada para avaliar o tempo de resposta de uma solução, a segunda para avaliar se uma solução consegue obter o resultado (eficácia) e a terceira para avaliar a qualidade da resposta da solução (eficiência) [Wainer et al. 2007].

Os experimentos realizados em uma solução precisam ter efetiva significância estatística, de acordo com o tipo de medição realizada e o teste estatístico correto para analisar essa medição. Os tipos de medida são classificados em categóricas ou nominais, ordinais, intervaláveis e de razão [Wainer et al. 2007]. Testes estatísticos são procedimentos usados para testar a hipótese nula, cujo pressuposto é de que não há diferença ou relação entre os grupos de dados ou eventos testados no objeto da pesquisa e que as diferenças encontradas se devem ao acaso, bem como a hipótese alternativa, na qual o pressuposto é que existem diferenças estatisticamente significantes entre as medidas.

Calculando a probabilidade da hipótese nula ser verdadeira ou não, por meio do teste adequado ao tipo de medição realizada, chega-se ao chamado valor-p ou *p-value*. Quando o nível de significância representado por este valor é inferior a um terminado indicador, sendo 0,05 (5%) o valor mais empregado, rejeita-se a hipótese nula e aceita-se hipótese alternativa, de que realmente existe a diferença e esta não foi encontrada ao acaso [Wainer et al. 2007] [Dean et al. 1999]. Ainda que o teste de hipóteses seja útil, quando se comparam valores obtidos em experimentos diferentes o teste de hipóteses não é suficiente. É necessário saber o quanto esses valores efetivamente diferem, para isso utiliza-se o chamado intervalo de confiança. Em [Montgomery 2017], este intervalo de confiança é definido em, pelo menos, 95%, o qual representa o maior e o menor valores assumíveis garantindo um p-valor de 0,05 [Wainer et al. 2007]. O intervalo de confiança não se sobrepõe, ou invalida, a medida de desvio padrão. Este último corresponde à indicação do quanto os dados do experimento podem variar em relação a média e é utilizado como parâmetro em alguns testes.

Os testes indicados, e, por consequência os mais utilizados, para comparação de até dois conjuntos de medições e obtenção do valor-p são: Teste T, Teste T Pareado, Teste U de *Mann-Whitney* ou *Wilcoxon rank-sum test*, *Wilcoxon signed-rank test*, Chi-quadrado e Teste Exato de *Fisher*. Para comparações múltiplas, com mais de dois conjuntos de valores, são: Teste ANOVA e *Kruskal-Wallis* [Wainer et al. 2007].

Sabendo que o estudo estatístico deve ser aplicado sobre uma coleção de n amostras de desempenho coletadas, o problema que resta é como definir o valor de n para

um determinado experimento. O Teorema do Limite Central é o resultado mais importante em estatística, do qual muitos métodos estatísticos comummente usados se baseiam para terem validade [Navidi 2012]. Este teorema diz que se for extraída uma amostra suficientemente grande o comportamento das médias tende a ser uma distribuição normal [Navidi 2012, Bukh 1992] ou gaussiana [Neto et al. 2010]. A distribuição normal (ou gaussiana), é o modelo estatístico que melhor representa o comportamento natural de um experimento, onde uma variável aleatória pode assumir qualquer valor dentro de um intervalo definido [Neto et al. 2010].

Amostra aqui refere-se às medições dos objetos de pesquisa, número de repetições ou iterações realizadas nos testes ou experimentos. Dependendo do tipo do objeto de pesquisa exitem cálculos específicos para o tamanho da amostra, porém, o teorema do limite central sugere que para a maioria dos casos uma amostra de tamanho 30 ou mais é suficientemente grande para que a aproximação normal seja adequada [Navidi 2012].

Para cada objeto de pesquisa existem critérios de medições aplicáveis. Segundo [Fortier and Michel 2003] a mensuração de desempenho pode ser classificada como medidas orientadas ao sistema ou ao usuário. As medições orientadas ao sistema transitam tipicamente no entorno da taxa de transferência (*Throughput*) e utilização. Taxa de transferência é definida como uma média por intervalo de tempo, sejam tarefas, processos ou dados. Utilização é a medida do intervalo de tempo em que um determinado recurso computacional está ocupado. Já as medições orientadas ao usuário compreendem o tempo de resposta (*response time*) e *turnaround time*.

Dentro desse conceito, é possivel perceber que métricas especializadas como *Reaction Time, Strech Fator*, MIPS (*Millions of Instructions Per Second*), MFLOPS (*Millions of Floating-Point Operations Per Second*), PPS (*packets per second*), BPS (*bits per second*), TPS (*Transactions Per Second*), *Nominal Capacity, Bandwidth, Usable Capacity, Efficiency, Idle Time, Reliability, Availability, Downtime, Uptime,* MTTF (*Mean Time To Failure*), *Cost/Performance Ratio* [Bukh 1992] [Fortier and Michel 2003] estão inseridas dentro destas duas generalizações.

A termologia das métricas é bastante extensa e variada. Pode também variar seu uso conforme o entendimento dos autores das publicações ou conforme a região. No entanto [Bukh 1992] e [Fortier and Michel 2003] são elucidativos tanto na definição, quanto no uso de cada métrica supra mencionada.

4. Metodologia

Para minerar os dados, todos os trabalhos publicados dos anos de 2000 a 2017 foram salvos localmente, numerados e separados por pasta segundo o ano de publicação. Trabalhos no formato resumo e aqueles escritos em língua estrangeira não fizeram parte da amostra das 426 publicações analisadas. Isto se deve ao fato de que, dada a natureza do formato resumo, certos detalhes da análise dos dados do objeto estudado poderiam ser suprimidos o que certamente geraria um viés. Bem como trabalhos em língua estrangeira aumentaria consideravelmente o número de termos de pesquisa se sua diversidade.

Em seguida, por meio de processo automatizado via software (NVivo[1]), foram pesquisadas citações de termos, utilizados em análise estatística bem como métricas e

[1] http://www.software.com.br/p/qsr-nvivo

testes utilizados para aferição de resultados, assim categorizados neste artigo. Estas categorias referem-se às boas práticas na coleta de dados e análise de resultados em pesquisa científica, conforme Tabelas 1, 2 e 3.

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

Tabela 2. Métricas selecionadas para coleta de dados

Descrição	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 Utilizável (CU)	”usable capacity”or ”capacidade utilizável”
Confiabilidade (CO)	Reliability OR Confiabilidade
Cost/Perfornace 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 Reação (TR)	reaction time or tempo de reação
TPS (TP)	TPS

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”

Após a tabulação, os dados foram sumarizados por ano conforme a categoria do termo. As Tabelas 4, 5 e 6 mostram os resultados obtidos na pesquisa de termos estatísticos, métricas e testes respectivamente. Note que somente termos que obtiveram resultados são sumarizados, aqueles onde não houve ocorrência foram suprimidos. Ainda na Tabela 4, os resultados do termo frequência foram suprimidos devido ao viés que os resultados obtiveram. Frequência é citada como unidade de medida do *clock* de processadores. Os resultados dos termos distribuição normal e gaussiana foram sumarizados juntos por se tratarem do mesmo objeto.

Tabela 4. Citações de Termos Estatísticos por ano

Ano	n	AM	DP	DN	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	-

Tabela 5. Citações de Métricas por ano

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

Tabela 6. Citações de Testes por ano

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

A Tabela 7 apresenta a sumarização dos resultados das Tabelas 4, 5 e 6, bem como a distribuição de citações por categoria do termo e ano. Nela é possível visualizar o número de artigos total no ano, o número de artigos que contém pelo menos uma ocorrência do termo. Também é identificado o número de ocorrências dentro dos artigos separados por categoria de termo.

Tabela 7. Distribuição de citações por tipo do termo e ano

Ano	n	Estatística		Métricas		Testes	
	-	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	-	-

5. Discussão

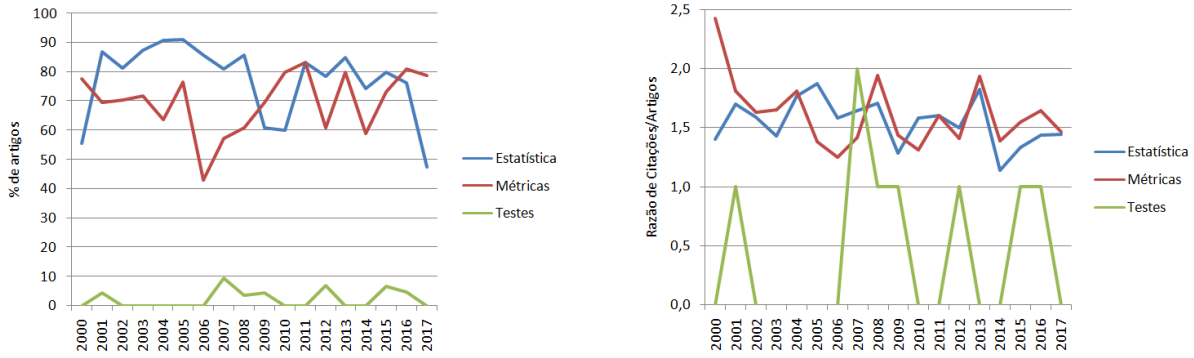
O objetivo aqui não é desmerecer ou invalidar os resultados obtidos nos experimentos realizados, uma vez que a prática permite identificar resultados coerentes com o esperado. O que se espera é verificar qual o cuidado dos autores ao publicarem esses resultados assim como a evolução da pesquisa científica uma vez que se tratam de quase 20 anos de WSCAD. Assume-se que todos os possíveis erros assim como as questões para evitá-los já foram levados em conta [Bukh 1992].

Nesta análise, foi utilizada a proporção de artigos com citações em relação ao total de artigos no ano e não o número de artigos para possibilitar a comparação entre os anos, uma vez que o número de artigos por ano não é o mesmo tendo grande variabilidade Figura 1(a). Do total de 426 artigos analisados, 79% citaram termos estatísticos, 67% métricas e apenas 2% testes estatísticos.

A razão foi obtida por meio dos dados Tabela 7, dividindo o número de artigos com citação pelo número de citações. A média de citações das três categorias foi de 1,46 citações por artigo. Individualmente observa-se que para termos estatísticos (1,57) e métricas (1,59) os valores ficaram acima da média geral em quase todos os anos.

Dos testes pesquisados apenas o Teste T ou de *Student* obteve resultado com 9 ocorrências, e, sendo que das duas ocorrências do nível de significância estatística (p-valor), nenhuma delas foi de artigos com ocorrência do Teste T.

Devido à baixa ocorrência de testes estatísticos nos resultados, em torno de 2%, resolveu-se investigar mais a fundo este dado. Para isso foi calculada uma segunda amostra dos 426 trabalhos, desconsiderando os 9 trabalhos onde já se encontrou a ocorrências



(a) Proporção de artigos com citações por ano

(b) Razão de citações por artigo ano

Figura 1. Gráficos dos Resultados Encontrados

destes termos, com intervalo de confiança de 95% e erro amostral de 5%, chegando a um total de 30 artigos sorteados aleatoriamente.

A segunda amostra de artigos foi encaminhada a dois revisores, que leram e analisaram a metodologia e análise dos resultados destes artigos segundo alguns critérios. Experimentação, se foi realizada ou não; Método amostral referindo-se ao método utilizado para encontrar o tamanho da amostra, pela utilização de *benchmarks*, estimação ou cálculo; Tamanho da amostra representado pelo número de experimentos realizados, repetições, instruções, jobs e afins; Se fica evidente o tamanho da amostra e se seu tamanho é adequado; Utilização de métricas; Se houve comparação com outra técnica e esta foi demonstrada adequadamente de alguma maneira.

Dos artigos analisados na segunda amostra, *nenhum* evidenciou igualidade ou deficiência em seus resultados, o pior caso apenas aponta ”indícios de melhora no desempenho”. Desta amostra, 24 deles realizaram experimentação, 6 são modelos teóricos ou memoriais descritivos de implementações de sistemas computacionais sem análise numérica de resultados. Daqueles que realizaram experimentação 10 usaram *benchmarks*, 13 estimaram o tamanho da amostra e 1 não citou números.

Os tamanhos de amostra foram considerados adequados para todos aqueles que usaram *benchmarks*, uma vez que se tratam de grandes coleções de registros e fica implícito que para cada registro de entrada há uma medição e registro de saída, também por serem um método amplamente aceito na comunidade científica [Wainer et al. 2007]. Daqueles trabalhos onde houve estimação da amostra (13) em 11 deles o tamanho da amostra foi considerado adequado, dado o tamanho amostral e em 2 foi considerado inadequado por estarem abaixo do mínimo estipulado pelo TLC [Navidi 2012]. É preciso estar atento, pois um *benchmark* pode representar a carga de entrada a que um experimento é submetido e não o montante de dados coletados, olhando por este aspecto apenas 8 trabalhos tiveram amostras adequadas com poder estatístico de análise especificado no texto.

Todos os artigos onde houve experimento (24) houve a utilização de métricas, deste total 15 compararam seus resultados a outras técnicas e 9 não o fizeram. Confirmando os resultados para utilização de métricas da primeira amostra em 70%.

Na segunda amostra apenas em 3 artigos a demonstração dos resultados foi embasada em critérios científicos (métodos estatísticos e modelos matemáticos), porém apenas 1 artigo utilizou minimamente métodos estatísticos (p-valor e DP) para confirmarem seus resultados. Percentualmente os valores para este critério de avaliação da primeira e segunda amostras são similares 2% e 3% respectivamente.

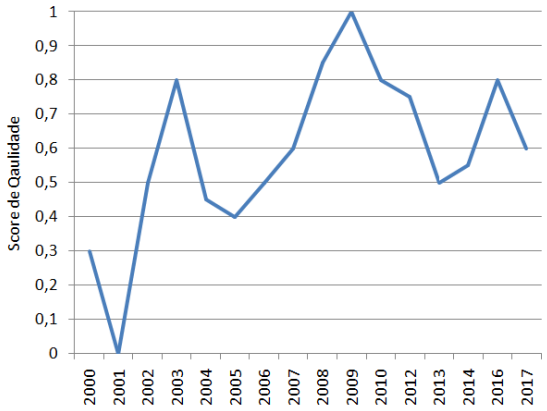


Figura 2. Score de Qualidade por Ano

Para calcular a evolução do WSCAD em termos de qualidade das publicações, a cada resultado positivo das categorias analisadas na segunda amostra foi atribuído peso 1 e peso zero para resultados negativos ou inexistentes. Calculados os somatórios de cada ano e as médias chegou-se em um *score* que compreende valores entre 0 e 1. A Figura 2 mostra a linha evolutiva do *score* de qualidade das publicações, evidenciando o amadurecimento do evento ao longo do tempo. Embora existam alternâncias na linha, a tendência da curva é ascendente com a maioria dos anos acima da média.

6. Conclusão

Neste artigo foi apresentada uma pesquisa sobre a ocorrência de termos estatísticos, métricas e testes estatísticos utilizados para comprovação dos resultados em pesquisa científica. Foram analisadas as publicações de todas as edições do WSCAD numa amostra total de 426 artigos.

Da amostra analisada 398 publicações fizeram referência a pelo menos um dos termos pesquisados, correspondendo a 93% do total. Isso mostra que existe a preocupação na realização pesquisa e na demonstração dos resultados, mesmo que inadequada ou incompleta dada a ocorrência de apenas 3% de testes estatísticos confirmada por uma segunda amostra de 30 artigos. É preciso não apenas colocar foco no desenvolvimento do objeto de pesquisa em si, mas também na aferição e apresentação dos resultados.

Quase a totalidade dos artigos apenas apresenta as medições da técnica implementada porém o questionamento é inevitável: apenas simples mensuração da métrica é suficiente para comparação entre técnicas similares, desconsiderando fatores de erro e considerando que elas foram executadas dentro dos mesmos padrões?

A pesquisa aqui descrita serve de alerta à comunidade científica. Há claramente a necessidade de atenção nos pontos destacados aqui em pesquisas. Os achados deste trabalho devem promover a reflexão sobre os cursos de graduação e pós-graduação e a neces-

sidade da inclusão da metodologia de análise estatística de dados aplicada à computação nas disciplinas básicas de formação. Neste sentido, a contribuição do WSCAD poderia se dar na direção de indicar na sua chamada de trabalhos a necessidade de que artigos apresentando análise de desempenho venham acompanhados de validação estatística de seus resultados ao mesmo tempo em que solicita aos revisores destes artigos que observem se tal estudo foi devidamente apresentado.

Referências

Adler, S., Schmitt, S., Wolter, K., and Kyas, M. (2015). A survey of experimental evaluation in indoor localization research. In *Indoor Positioning and Indoor Navigation (IPIN), 2015 International Conference on*, pages 1–10. IEEE.

Bukh, P. N. D. (1992). The art of computer systems performance analysis, techniques for experimental design, measurement, simulation and modeling.

Dean, A., Voss, D., Draguljić, D., et al. (1999). *Design and analysis of experiments*, volume 1. Springer.

Fortier, P. and Michel, H. (2003). *Computer systems performance evaluation and prediction*. Elsevier.

Montgomery, D. C. (2017). *Design and analysis of experiments*. John Wiley & Sons.

Navidi, W. (2012). *Probabilidade e estatística para ciências exatas*. AMGH.

Neto, B. B., Scarminio, I. S., and Bruns, R. E. (2010). *Como Fazer Experimentos-: Pesquisa e Desenvolvimento na Ciência e na Indústria*. Bookman.

Prechelt, L. (1994). A quantitative study of experimental evaluations of neural network learning algorithms: Current research practice. *IEEE Transactions on Neural Networks*, 6.

Tedre, M. and Moisseinen, N. (2014). Experiments in computing: A survey. *The Scientific World Journal*, 2014.

Tichy, W. F., Lukowicz, P., Prechelt, L., and Heinz, E. A. (1995). Experimental evaluation in computer science: A quantitative study. *Journal of Systems and Software*, 28(1):9–18.

Wainer, J., Barsottini, C. G. N., Lacerda, D., and de Marco, L. R. M. (2009). Empirical evaluation in computer science research published by ACM. *Information and Software Technology*, 51(6):1081–1085.

Wainer, J. et al. (2007). Métodos de pesquisa quantitativa e qualitativa para a ciência da computação. *Atualização em informática*, 1:221–262.