# Estatísticas de citações[1]

*Robert Adler[2], John Ewing[3] e Peter Taylor[4]*

## Resumo

Este artigo é um relatório acerca do emprego e uso de citações na avaliação de pesquisas científicas. A idéia de que a avaliação da pesquisa deve ser feita empregando métodos "simples e objetivos" é cada vez mais prevalecente hoje no mundo acadêmico, o que tem gerado uma "cultura de números", sustentada no pressuposto de que tais avaliações são mais precisas e buscam superar julgamentos subjetivos da revisão por pares. No entanto, tais considerações são analisadas por profissionais que "lidam com números" – os matemáticos e os estatísticos. A convicção nas estatísticas deixa de ter fundamento quando estas são mal aplicadas ou mal interpretadas, como pode ocorrer no caso de uso de estatística para classificar periódicos, documentos, pessoas, programas e disciplinas. Os autores chamam a atenção para a objetividade ilusória dos números e para o fato de que a avaliação baseada em dados de citações pode fornecer uma visão limitada,

[1] Texto publicado sob o título inglês *Citations Statistics*, em junho de 2008, como Relatório da *International Mathematical Union* (IMU), em cooperação com o *International Council of Industrial and Applied Mathematics* (ICIAM) e o *Institute of Mathematical Statistics* (IMS). Disponivel em: http://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf
O Comitê Editorial de *Mediações* agradece a Martin Groetschel, Secretário da *International Mathematical Union* (IMU), por autorizar a publicação deste relatório. Tradução de Anselmo Rodrigues da Costa Filho. Revisão de Sávio Cavalcante.

[2] Instituto de Tecnologia Technion-Israel.

[3] Sociedade Matemática Americana.

[4] Universidade de Melbourne. Sob a direção de John Ewing, os três autores compõem o Comitê Conjunto sobre a Avaliação Quantitativa de Pesquisa.

superficial e incompleta da qualidade da pesquisa.

**Palavras-chave:** Avaliação. Fator de impacto. Índice H. Pesquisa científica.

## Citations statistics

### Abstract

This article is a report on the use of citations in the evaluation of scientific research. The idea that the evaluation of research should be conducting using "simple and objective" measures is increasingly prevalent in today's academic world, which has generated a "culture of numbers", based on the fact that such evaluations are more precise and attempt to overcome the subjective judgements of peer review. Nevertheless, these considerations are analyzed by professionals who "deal with numbers" – mathematicians and statisticians. The belief in numbers lacks a sound basis when the numbers are poorly applied or interpreted, as can occur in the case of using statistics to classify journals, documents, people, programs and disciplines. The authors call attention to the illusory objectivity of numbers and to the fact that evaluation based on citation data can provide a limited, superficial and incomplete vision of the quality of research.

**Keywords:** Evaluation. Impact factor. H índex. Scientific research.

Este é um relatório sobre o emprego e uso indevidos de dados de citações na avaliação de pesquisa científica. A idéia que a avaliação da pesquisa deve ser feita empregando métodos "simples e objetivos" é cada vez mais prevalecente hoje. Os métodos "simples e objetivos" são amplamente interpretados como *bibliometria*, isto é, dados de citações e as estatísticas derivadas deles. Há uma convicção que as estatísticas de citações são basicamente mais precisas porque substituem números simples por julgamentos complexos, e conseqüentemente superam a possível subjetividade da revisão por pares. Mas esta convicção é sem fundamento.

- Confiar nas estatísticas não é mais correto quando as estatísticas forem usadas inadequadamente. Sem dúvida, as estatísticas podem enganar quando forem mal aplicadas ou mal interpretadas. Muito da bibliometria moderna parece confiar em experiência e intuição sobre a interpretação e validade das estatísticas de citações.

• Enquanto os números parecem ser "objetivos", a objetividade deles pode ser ilusória. O significado de uma citação pode ser até mesmo mais subjetivo que a revisão por pares. Porque esta subjetividade é menos óbvia para citações, aqueles que usam dados de citações são menos prováveis de entender as suas limitações.

• A confiança exclusiva em dados de citações proporciona na melhor das hipóteses uma compreensão incompleta e geralmente superficial da pesquisa – uma compreensão que só é válida quando reforçada através de outros julgamentos. *Números não são basicamente superiores a julgamentos legítimos*.

Utilizar dados de citações para avaliar a pesquisa significa, em última análise, utilizar estatísticas baseadas em citações para classificar coisas – periódicos, documentos, pessoas, programas, e disciplinas. As ferramentas estatísticas usadas para classificar estas coisas são muitas vezes mal-compreendidas e usadas indevidamente.

• Para periódicos, o fator de impacto é mais freqüentemente usado para classificação. Esta é uma média simples derivada da distribuição das citações para uma coleção de artigos no periódico. A média captura apenas uma pequena quantidade de informações sobre aquela distribuição, e é uma estatística bastante rudimentar. Além disso, há muitos fatores confundidores ao avaliar periódicos através das citações, e qualquer comparação de periódicos requer precaução ao usar fatores de impacto. Usar somente o fator de impacto para avaliar um periódico é como usar o peso sozinho para avaliar a saúde de uma pessoa.

• Para documentos, em vez de confiar na contagem real das citações para comparar documentos individuais, as pessoas frequentemente substituem o fator de impacto de periódicos nos quais os documentos aparecem. Elas acreditam que fatores de impactos mais altos têm que significar contagens de citação mais altas. Mas este frequentemente não é o caso! Este é um uso indevido difundido da estatística que precisa ser desafiado sempre que e onde quer que aconteça.

• Para cientistas individuais, registros completos de citações podem ser difíceis de comparar. Como consequência, houve tentativas de encontrar estatísticas simples que capturam a complexidade total do registro de citações de um cientista com um único número. O mais notável destes é

o índice h, que parece estar ganhando em popularidade. Mas até mesmo uma inspeção casual do índice h e suas variantes mostram que estas são tentativas ingênuas para entender registros complicados de citações. Enquanto elas capturam uma quantidade pequena de informações sobre a distribuição das citações de um cientista, elas perdem informações cruciais que são essenciais para a avaliação da pesquisa.

A validade das estatísticas tais como o fator de impacto e o índice h não são nem bem entendidos nem bem estudados. A conexão destas estatísticas com a qualidade da pesquisa às vezes é estabelecida na base da "experiência." A justificativa para se confiar nelas é que elas estão "prontamente disponíveis." Os poucos estudos que foram feitos destas estatísticas focalizavam estreitamente em mostrar uma correlação com alguma outra medida de qualidade em vez de determinar como se pode melhor extrair informações úteis dos dados das citações.

Não rejeitamos as estatísticas de citações como uma ferramenta para avaliar a qualidade da pesquisa – os dados e as estatísticas de citações podem fornecer algumas informações valiosas. Reconhecemos que a avaliação deve ser prática, e por este motivo as estatísticas de citações facilmente derivadas quase que seguramente farão parte do processo. Mas os dados das citações fornecem apenas uma visão limitada e incompleta da qualidade da pesquisa, e as estatísticas derivadas dos dados das citações às vezes são entendidas precariamente e usadas indevidamente. A pesquisa é demasiadamente importante para medir seu valor com apenas uma única ferramenta rudimentar.

Esperamos que os envolvidos em avaliação leiam tanto o comentário como os detalhes deste relatório para entender não apenas as limitações das estatísticas de citações, mas também como usá-las melhor. Se fixarmos padrões altos para a gestão da ciência, certamente deveríamos fixar padrões igualmente altos para avaliar sua qualidade.

## Introdução

A pesquisa científica é importante. A pesquisa constitui a base de muito progresso em nosso mundo moderno e proporciona a esperança que possamos resolver alguns dos problemas aparentemente obstinados que a humanidade enfrenta, do meio ambiente até a nossa população crescente. Por causa disto, os governos e as instituições ao redor do mundo fornecem apoio financeiro considerável para a pesquisa científica. Naturalmente, querem saber se o seu

dinheiro está sendo investido sabiamente; querem avaliar a qualidade da pesquisa pela qual pagam a fim de tomar decisões informadas sobre investimentos futuros.

Isto não é muita novidade: as pessoas têm avaliado a pesquisa por muitos anos. O que *é* novo, entretanto, é a noção que a boa avaliação deve ser "simples e objetiva", e que isto pode ser alcançado confiando-se principalmente em métrica (estatística) derivada dos dados de citações ao invés de uma variedade de métodos, incluindo avaliações pelos próprios cientistas. O parágrafo de abertura de um relatório recente declara esta visão perfeitamente:

> É a intenção do Governo que o método atual para determinar a qualidade da pesquisa universitária - o Exercício de Avaliação da Pesquisa do REINO UNIDO (RAE) – deve ser substituído depois que o próximo ciclo seja completado em 2008. A métrica, em vez da revisão por pares, será o foco do novo sistema e é esperado que a bibliometria (usando contagens de artigos de periódicos e as suas citações) será um índice de qualidade central neste sistema (EVIDENCE REPORT, 2007, p.3).

Aqueles que argumentam por esta objetividade simples acreditam que a pesquisa é demasiadamente importante para se confiar em avaliações subjetivas. Acreditam que a métrica baseada em citações traz clareza ao processo de classificação e elimina ambigüidades inerentes em outras formas de avaliação. Acreditam que as métricas cuidadosamente escolhidas sejam independentes e livres de tendenciosidade. Mais que tudo, acreditam que tais métricas nos permitem comparar todas as partes do empreendimento da pesquisa – publicações, documentos, pessoas, programas, e até mesmo disciplinas inteiras – simples e efetivamente, sem o uso da revisão subjetiva de pares.

Mas esta fé na precisão, independência, e eficácia da métrica é inapropriada.

- Primeiro, a precisão dessas métricas é ilusória. É um axioma comum que a estatística pode mentir quando forem incorretamente usadas. O abuso das estatísticas de citações é difundido e notório. Apesar de repetidas tentativas de advertir contra tal uso impróprio (por exemplo, o uso impróprio do fator de impacto), governos, instituições, e até mesmo os próprios cientistas continuam extraindo conclusões não comprovadas ou até mesmo falsas da má aplicação das estatísticas de citações.

- Segundo, a confiança exclusiva em métricas baseadas em citações substitui um tipo de julgamento por outro. Em vez da revisão subjetiva de pares tem-se a interpretação subjetiva do significado da citação. Aqueles que promovem confiança exclusiva em métrica baseada em citações

implicitamente pressupõem que cada citação significa a mesma coisa sobre a pesquisa citada – seu "impacto." Esta é uma suposição que não é comprovada e é muito provavelmente incorreta.

• Terceiro, enquanto as estatísticas são valiosas para entender o mundo em que vivemos, elas fornecem apenas uma compreensão parcial. Em nosso mundo moderno, às vezes é moda afirmar um ponto de vista místico que as medições numéricas são superiores a outras formas de percepção. Aqueles que promovem o uso das estatísticas de citações como uma *substituição* para uma percepção mais completa da pesquisa implicitamente defendem tal ponto de vista. Não apenas precisamos usar as estatísticas *corretamente* – precisamos usá-las sabiamente também.

Não debatemos com o esforço para avaliar a pesquisa, mas certamente com a exigência que tais avaliações confiam predominantemente em métrica baseada em citações "simples e objetivas" – uma exigência que é frequentemente interpretada como requerendo números fáceis de calcular os quais classificam publicações ou pessoas ou programas. A pesquisa normalmente tem múltiplas metas, tanto a curto quanto em longo prazo, e, portanto é razoável que seu valor deva ser avaliado através de critérios múltiplos. Os matemáticos sabem que há muitas coisas, tanto reais quanto abstratas, que não podem ser simplesmente ordenadas, no sentido que cada uma das duas pode ser comparada. A comparação frequentemente requer uma análise mais complicada, que às vezes deixa alguém indeciso sobre qual das duas coisas é "melhor". A resposta correta para "Qual é a melhor?" às vezes é: "Depende!".

O argumento para usar métodos múltiplos para avaliar a qualidade da pesquisa foi feito antes (por exemplo MARTIN, 1996 ou CAREY COWLING TAYLOR, 2007). As publicações podem ser julgadas de muitas formas, não apenas através de citações. Medidas de opinião favorável tais como convites, associação em grupos de editoress, e prêmios frequentemente medem a qualidade. Em algumas disciplinas e em alguns países, a concessão de fundos pode exercer uma função. E a revisão por pares – o julgamento de cientistas contemporâneos – é um componente importante da avaliação. (Não devemos descartar a revisão por pares meramente porque às vezes ela se torna falha por influências, mais do que devemos descartar as estatísticas de citações porque elas às vezes se tornam falhas por uso indevido.) Esta é uma pequena amostra dos múltiplos modos nos quais a avaliação pode ser feita. Há muitas possibilidades para a boa avaliação, e a importância relativa delas varia entre as disciplinas. Apesar disto, a estatística baseada em citação "objetiva"

se torna repetidamente o método preferido para a avaliação. O chamariz de um processo simples e números simples (preferivelmente um único número) dão certa impressão de superar o bom senso e o bom julgamento.

Este relatório é escrito por cientistas matemáticos para tratar do uso indevido de estatísticas em avaliar a pesquisa científica. Por certo, este uso indevido às vezes é dirigido para a própria disciplina de matemática, e isso é uma das razões para escrever este relatório. A cultura especial de citações da matemática, com baixas contagens de citações para periódicos, documentos, e autores, a torna especialmente vulnerável para o abuso das estatísticas de citações. Porém, acreditamos que *todos* os cientistas, como também o público geral, devem estar ansiosos para usar métodos científicos seguros ao avaliar a pesquisa.

Alguns na comunidade científica dispensariam completamente as estatísticas de citações em uma reação cínica para abuso passado, mas fazendo assim significaria deixar de lado uma valiosa ferramenta. As estatísticas baseadas em citações podem exercer uma função na avaliação da pesquisa, contanto que elas sejam usadas corretamente, interpretadas com precaução, e constituírem apenas parte do processo. As citações fornecem informações sobre periódicos, documentos, e pessoas. Não queremos esconder essa informação; queremos esclarecê-la.

Este é o objetivo deste relatório. As primeiras três seções tratam os modos nos quais os dados de citações podem ser empregados (e usados indevidamente) para avaliar periódicos, documentos, e pessoas. A próxima seção discute os significados variados das citações e as consequentes limitações em estatísticas baseadas em citações. A última seção recomenda o uso inteligente das estatísticas e estimula que as avaliações moderem o uso das estatísticas de citações com outros julgamentos, mesmo que isto torne as avaliações menos simples.

"Tudo deve ser feito tão simples quanto possível, mas não mais simples", Albert Einstein certa vez disse[5]. Este conselho de um dos cientistas preeminentes do mundo é especialmente apropriado ao avaliar pesquisa científica.

---

[5] Esta citação foi atribuída a Einstein na *Reader`s Digest*, Outubro de 1977. Parece ser derivada da efetiva citação dele: "Quase que não pode ser negado que a meta suprema de toda teoria seja tornar os elementos básicos irredutíveis tão simples e tão poucos quanto possíveis sem ter que se render à representação adequada de um único dado de experiência." Proveniente de "*On the Method of Theoretical Physics*". A palestra de Herbert Spencer, proferida em Oxford (10 de junho de 1933); também publicada em *Philosophy of Science*, Vol. 1, no. 2 (Abril de1934), pp. 163 169.

## Classificando os periódicos: o fator de impacto[6]

O fator de impacto foi criado nos anos de 1960 como um modo para medir o valor dos periódicos calculando-se o número médio de citações por artigo durante um período específico de tempo (GARFIELD, 2005). A média é computada de dados coletados por *Thomson Scientific* (previamente chamado de Instituto para Informações Científicas), o qual publica o *Journal Citation Reports* (Relatórios de Citações em Periódicos). O *Thomson Scientific* extrai, por ano, referências de mais de 9000 periódicos, acrescentando informações sobre cada artigo e suas referências ao seu banco de dados (THOMSON: SELECTION). Usando estas informações, pode-se contar com que frequência um artigo específico é citado por artigos subsequentes que são publicados na coleção de periódicos relacionados. (Observamos que o *Thomson Scientific* relaciona menos da metade dos periódicos de matemática cobertos pelo *Mathematical Reviews* e *Zentralblatt*, os dois maiores periódicos de revisão em matemática[7]).

Para um periódico e ano específicos, o fator de impacto do periódico é computado calculando-se o número médio de citações para os artigos no periódico durante os dois anos anteriores de todos os artigos publicados naquele determinado ano (na coleção específica de periódicos relacionados pelo *Thomson Scientific*). Se o fator de impacto de um periódico for 1,5 em 2007, significa que em média os artigos publicados durante 2005 e 2006 foram citados 1,5 vezes através de artigos na coleção de todos os periódicos relacionados publicados em 2007.

O próprio *Thomson Scientific* usa o fator de impacto como um fator ao selecionar quais periódicos a relacionar (THOMSON: SELECTION). Por outro lado,

---

[6] Enquanto nos concentramos no fator de impacto do *Thomson Scientific* nesta seção, notamos que *Thomson* promove o uso de duas outras estatísticas. Igualmente, estatísticas semelhantes baseadas em contagens médias de citações para periódicos podem ser derivadas de outros bancos de dados, incluindo *Scopus*, *Spires*, *Google Scholar*, e (para matemática) o banco de dados de citações do *Math Reviews*. Este último consiste de citações de mais 400 periódicos de matemática do período de 2000 até o presente, identificados como artigos que foram listados em *Math Reviews* desde 1940; ele inclui mais de 3 milhões de citações.

[7] O *Thomson Scientific* indica (Março de 2008) que indexa periódicos nas seguintes categorias: MATEMÁTICA (217), MATEMÁTICA APLICADA (177), MATEMÁTICA INTERDISCIPLINAR (76), FÍSICA, MATEMÁTICA (44), PROBABILIDADE E ESTATÍSTICA (96).
As categorias se sobrepõem e o número total de periódicos é de aproximadamente 400. Por contraste, o *Mathematical Reviews* inclui artigos de mais de 1200 periódicos a cada ano, e considera mais de 800 periódicos como "núcleo" (no sentido que todo artigo no periódico esteja incluído em *Math Reviews*). *Zentralblatt* abrange um número semelhante de periódicos de matemática.

*Thomson* promove o uso do fator de impacto mais geralmente para comparar periódicos.

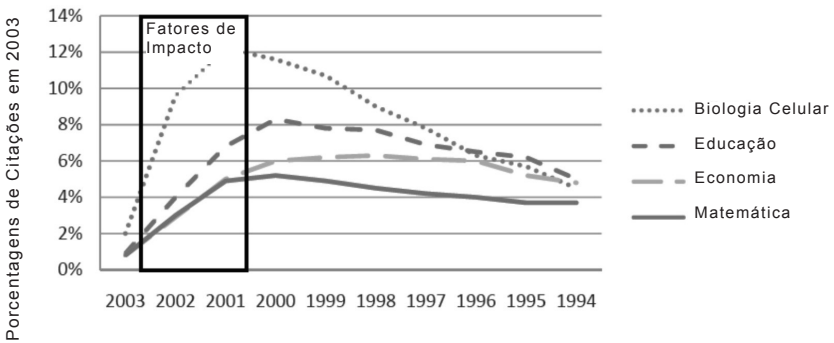
> Como uma ferramenta para a gestão de coleções de periódicos da biblioteca, o fator de impacto fornece ao administrador da biblioteca informações sobre periódicos já na coleção e periódicos sob consideração para aquisição. Estes dados também devem ser combinados com dados de custo e de circulação para tomar decisões racionais sobre compras de periódicos (THOMSON: IMPACT FACTOR).

Muitos escritores têm mostrado que não se deve julgar o valor acadêmico de um periódico usando apenas os dados de citações, e os autores atuais concordam muito com isso. Além desta observação geral, o fator de impacto também tem sido criticado por outras razões. (Ver SEGLEN, 1997; AMIN, MABE, 2000; MONASTERSKY, 2005; EWING, 2006; ADLER, 2007 e HALL, 2007).

> (i) A identificação do fator de impacto como uma média não é bastante correta. Porque muitos periódicos publicam artigos não substanciais tais como cartas ou editoriais, os quais raramente são citados, estes artigos não são contados no denominador do fator de impacto. Por outro lado, enquanto infrequentes, estes artigos às vezes são citados, e estas citações são contadas no numerador. Portanto, o fator de impacto não é totalmente as citações médias por artigo. Quando os periódicos publicam um grande número de tais artigos "não substanciais", esta divergência pode ser significativa. Em muitas áreas, inclusive matemática, esta divergência é mínima.
>
> (ii) O período de dois anos usado para definir o fator de impacto tinha a finalidade de tornar a estatística atual (GARFIELD, 2005). Para alguns campos, tais como as ciências biomédicas, isto é apropriado porque a maioria dos artigos publicados recebe a maioria das suas citações logo após a publicação. Em outros campos, como por exemplo, a matemática, a maioria das citações ocorre além do período de dois anos. Examinando uma coleção de mais de 3 milhões de recentes citações em periódicos de matemática (o banco de dados *Math Reviews Citation*) vê-se que aproximadamente 90% das citações para um periódico estão fora desta lacuna de dois anos. Por conseguinte, o fator de impacto é baseado em um mero 10% da atividade de citações e perde a vasta maioria das citações[8].

[8] O banco de dados de citações do Mathematical Reviews inclui (março de 2008) mais de 3 milhões de referências em aproximadamente 400 periódicos publicados de 2000 até o presente.

## Curvas de Citações[9]



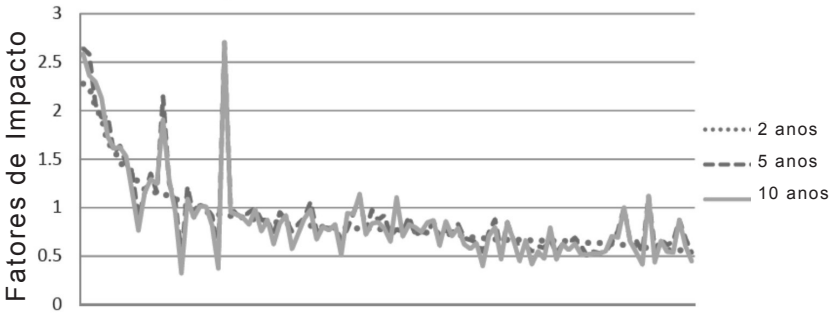

O intervalo de dois anos significa que o fator de impacto está equivocado? Para os periódicos de matemática a evidência é questionável. O *Thomson Scientific* computa fatores de impacto de 5 anos, os quais ele indica correlacionarem bem com os habituais fatores de impacto de dois anos (GARFIELD, 1998). Usando o banco de dados de citações da *Math Reviews*, podem-se computar os "fatores de impacto" (isto é, as citações médias por artigo) para uma coleção dos 100 periódicos de matemática mais citados utilizando períodos de 2, 5, e 10 anos. O gráfico abaixo mostra que fatores de impacto de 5 e 10 anos geralmente seguem o fator de impacto de 2 anos.

---

As referências são comparadas a artigos no banco de dados do MR e se estendem durante muitas décadas. Ao contrário do Índice de Citações de Ciência, as citações, tanto para livros quanto para periódicos, são incluídas. É um fato curioso que aproximadamente 50% das citações são para artigos que aparecem na década anterior; 25% citam artigos que aparecem na década antes disso; 12,5% citam artigos na década anterior; e assim por diante. Este tipo de comportamento é especial para cada disciplina, naturalmente.

[9] O gráfico mostra a época das citações de artigos publicados em 2003 abrangendo quatro campos diferentes. Citações para artigo publicado em 2001 / 2002 são aqueles contribuindo para o fator de impacto; todas as outras citações são irrelevantes ao fator de impacto. Dados do Thomson Scietinfic.

## Os 100 Principais Periódicos de Matemática[10]

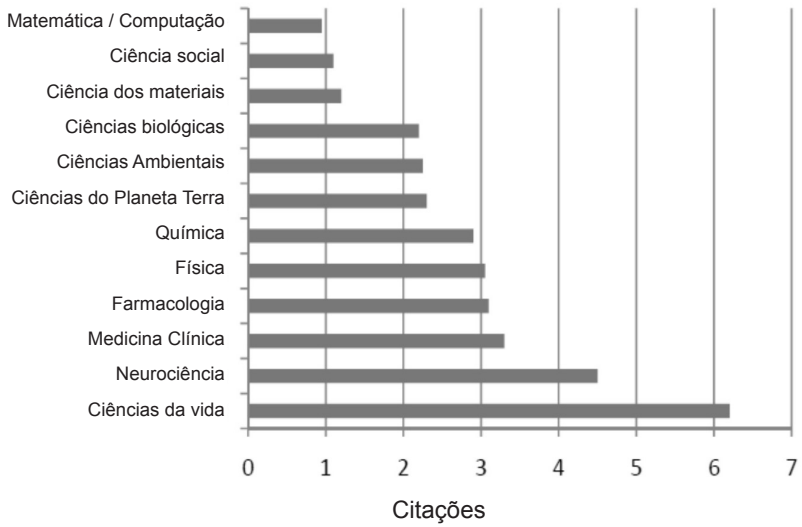



Aquele grande valor discrepante é um periódico que não publicou documentos durante parte deste tempo; os menores valores discrepantes tendem a ser periódicos que publicam um número relativamente pequeno de documentos a cada ano, e o gráfico meramente reflete a variabilidade normal em fatores de impacto para tais periódicos. É evidente que mudar o número de "anos alvo" ao calcular as mudanças do fator de impacto muda a classificação dos periódicos, mas as mudanças são geralmente modestas, exceto para periódicos pequenos, onde os fatores de impacto também variam ao mudar o "ano da fonte" (veja abaixo).

> (iii) O fator de impacto varia consideravelmente entre as disciplinas (AMIN, MABE, 2000). Parte desta diferença se origina da observação (ii): Se em algumas disciplinas ocorrerem muitas citações fora da lacuna de dois anos, os fatores de impacto para os periódicos serão bem mais baixos. Por outro lado, parte da diferença é simplesmente que as culturas de citação diferem de disciplina para disciplina, e os cientistas citarão os documentos em diferentes padrões e por diferentes razões. (Acrescentamos detalhes a esta observação posteriormente porque o significado das citações é extremamente importante). Segue que não se pode em qualquer modo significativo comparar dois periódicos em disciplinas diferentes usando fatores de impacto.

[10] "Fatores de impacto" para 2, 5, e 10 anos para os 100 periódicos de matemática. Informações do banco de dados de citações da *Math Reviews*.

## Citações médias por artigo[11]




(iv) O fator de impacto pode variar consideravelmente de ano para ano, e a variação tende a ser maior para os periódicos menores (AMIN, MABE, 2000). Para periódicos que publicam menos que 50 artigos, por exemplo, a *mudança* média no fator de impacto de 2002 a 2003 foi quase 50%. Isto é esperado completamente, é claro, porque o tamanho da amostra para periódicos pequenos é pequeno. Por outro lado, geralmente se compara os periódicos por um ano fixo, sem levar em conta a variação mais alta para periódicos pequenos.

(v) Os periódicos que publicam artigos em idiomas que não seja o Inglês provavelmente receberão menos citações porque uma grande parte da comunidade científica não sabe (ou não) os lê. E o tipo de periódico, ao invés de apenas a qualidade, pode influenciar o fator de impacto. Os periódicos que publicam artigos de revisão, por exemplo, receberão freqüentemente muito mais citações do que os periódicos que não publicam, e, portanto tem mais altos (às vezes, substancialmente mais altos) fatores de impacto (AMIN, MABE, 2000).

(vi) A crítica mais importante do fator de impacto é que seu significado não é bem entendido. Ao usar o fator de impacto para comparar dois periódicos,

[11] Citações médias por artigo para diferentes disciplinas, mostrando que as práticas de citação diferem acentuadamente. Dados do *Thomson Scientific* (AMIN, MABE, 2000).

não há um modelo a *priori* que defina o que significa ser "melhor". O único modelo deriva do próprio fator de impacto - um fator de impacto maior significa um periódico melhor. No paradigma estatístico clássico, define-se um modelo, formula-se uma hipótese (de nenhuma diferença), e então se acha uma estatística, a qual dependendo de seus valores permite-se aceitar ou rejeitar a hipótese.

Derivar Informações (e possivelmente um modelo) dos próprios dados é uma abordagem legítima para a análise estatística, mas neste caso não está claro que informação foi derivada. Como o fator de impacto mede a qualidade? É a melhor estatística para medir a qualidade? O que precisamente ele mede? (Nossa discussão posterior sobre o significado de citações é pertinente aqui.) Notavelmente pouco se sabe sobre um modelo para a qualidade do periódico ou como ele pode relacionar-se com o fator de impacto.

As seis críticas acima sobre o fator de impacto são todas válidas, mas elas significam apenas que o fator de impacto é rudimentar, não inútil. Por exemplo, o fator de impacto pode ser utilizado como um ponto de partida em classificar os periódicos em grupos usando-se os fatores de impacto inicialmente para definir os grupos e então empregando outros critérios para aperfeiçoar a classificação e verificar que os grupos fazem sentido. Mas usar o fator de impacto para avaliar os periódicos requer precaução. O fator de impacto não pode ser usado para comparar os periódicos através das disciplinas, por exemplo, e deve-se olhar de perto para o tipo de periódicos ao usar o fator de impacto para classificá-los. Também se deve prestar muita atenção às variações anuais, especialmente para periódicos menores, e entender que pequenas diferenças podem ser fenômenos puramente aleatórios. E é importante reconhecer que o fator de impacto pode não refletir com precisão a gama completa da atividade de citações em algumas disciplinas, tanto porque nem todos os periódicos são relacionados e porque o período de tempo é muito curto. Outras estatísticas baseadas em mais longos períodos de tempo e mais publicações podem ser melhores indicadores de qualidade. Finalmente, as citações são apenas um modo para julgar os periódicos, e devem ser acrescentados com outras informações (a mensagem central deste relatório).

Estas são todas as precauções semelhantes àquelas que se faria para qualquer classificação baseada em estatísticas. Classificar periódicos descuidadamente de acordo com os fatores de impacto para um ano especificado é um uso indevido da estatística. A seu crédito, *Thomson Scientific* concorda com esta declaração

e (gentilmente) previne aqueles que usam o fator de impacto sobre estas coisas.

> *Thomson Scientific* não depende apenas do fator de impacto em avaliar a utilidade de um periódico, e ninguém mais deve também. O fator de impacto não deve ser usado sem atenção cuidadosa para os muitos fenômenos que influenciam os padrões de citação, como por exemplo, o número médio de referências citadas no artigo médio. O fator de impacto deve ser usado com a revisão de pares informada (THOMSON: IMPACT FACTOR).

Infelizmente, este conselho é muito frequentemente ignorado.

## Classificando documentos

O fator de impacto e as estatísticas semelhantes baseadas em citações podem ser usados indevidamente ao classificar os periódicos, mas há um uso indevido mais fundamental e mais insidioso: usar o fator de impacto para comparar documentos individuais, pessoas, programas, ou até mesmo disciplinas. Este é um problema crescente que se estende por muitas nações e muitas disciplinas, tornado pior por recentes avaliações nacionais de pesquisa.

De certo modo, este não é um fenômeno novo. Os cientistas são chamados frequentemente para fazer julgamentos sobre registros de publicação, e ouvem-se comentários tais como, "Ela publica em bons periódicos" ou a "Maioria dos documentos dele está em periódicos de baixo nível". Estas podem ser avaliações sensatas: a qualidade dos periódicos nos quais um cientista geralmente (ou constantemente) publica é um dos muitos fatores que se pode usar para avaliar a pesquisa total do cientista. Porém, o fator de impacto aumentou a tendência para designar as propriedades de um periódico individual para *cada* artigo dentro daquele periódico (e para *cada* autor).

*O Thomson Scientific* implicitamente promove esta prática*:*

> Talvez o uso mais importante e recente de impacto esteja no processo de avaliação acadêmica. O fator de impacto pode ser usado para prover uma aproximação total do prestígio dos periódicos nos quais os indivíduos foram publicados (THOMSON: IMPACT FACTOR).

Aqui estão alguns exemplos dos modos nos quais as pessoas interpretaram este conselho, reportados por matemáticos ao redor do mundo:

> Exemplo 1: Minha universidade apresentou recentemente uma nova classificação de periódicos usando os periódicos do *Science Citation Index*

*Core*. Os periódicos são divididos em três grupos baseados apenas no fator de impacto. Há 30 periódicos na lista principal, não contendo periódico de matemática. A segunda lista contém 667, que incluem 21 periódicos de matemática. A publicação na primeira lista faz o apoio universitário de pesquisa triplicar; a publicação na segunda lista, dobrar. A publicação na lista principal premia 15 pontos; a publicação em qualquer periódico coberto por *Thomson Scientific* premia 10. A promoção requer um número mínimo fixo de pontos.

Exemplo 2: Em meu país, o corpo docente de uma universidade com cargos permanentes são avaliados a cada seis anos. As avaliações sequenciais bem sucedidas são a chave para todo o sucesso acadêmico. Além de um *curriculum vitae*, o maior fator em avaliação diz respeito classificar cinco documentos publicados. Em anos recentes, a estes são dados 3 pontos se eles aparecem em periódicos no terço superior da lista do *Thomson Scientific*, 2 pontos se no segundo terço, e 1 ponto no terço inferior. (As três listas são criadas usando o fator de impacto.)

Exemplo 3: Em nosso departamento, cada membro do corpo docente da universidade é avaliado por uma fórmula que envolve o número de documentos equivalentes de um único autor, multiplicado pelo fator de impacto dos periódicos nos quais eles aparecem. As promoções e contratações são baseadas parcialmente nesta fórmula.

Nestes exemplos, como também em muitos outros reportados a nós, o fator de impacto está sendo usado ou explicita ou implicitamente para comparar documentos individuais junto com os seus autores: se o fator de impacto do periódico A for maior do que aquele do periódico B, então seguramente um documento no A deve ser superior a um documento em B, e o autor A superior ao autor B. Em alguns casos, este raciocínio é estendido para classificar os departamentos ou até mesmo disciplinas inteiras.

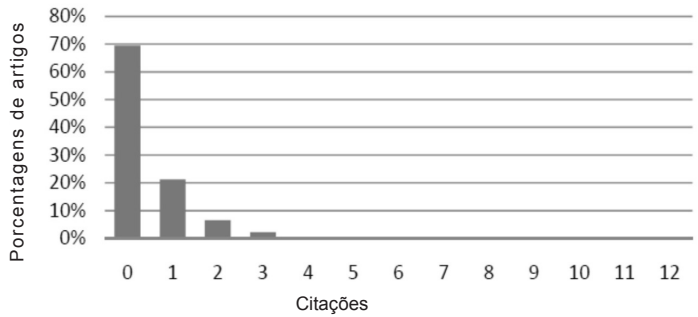
Há muito tempo já se sabe que a distribuição das contagens de citações para documentos individuais em um periódico é altamente enviesada, assemelhando-se a uma assim chamada lei de poder (SEGLEN, 1996; GARFIELD, 1987). Isto tem consequências que podem ser tornadas exatas com um exemplo.

A distribuição para documentos no *Proceedings of the American Mathematical Society* durante o período 2000 - 2004 pode ser vista abaixo. O *Proceedings* publica documentos curtos, normalmente mais curtos que

dez páginas de extensão. Durante este período, publicou 2.381 documentos (aproximadamente 15.000 páginas). Usando 2005 periódicos no banco de dados de citação da *Math Reviews*, a contagem de citações média por artigo (isto é, o fator de impacto) é 0,434.

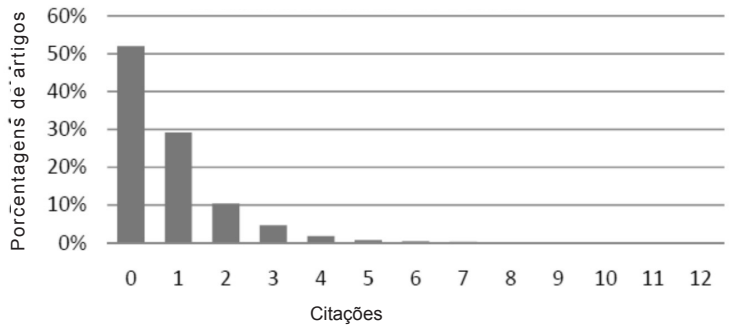
## PROCEEDINGS OF THE AMS

(AMS - *American Mathematical Society* – Sociedade Americana de Matemática)




O *Transactions of the AMS* publica artigos mais longos que são geralmente mais substanciais, tanto em extensão quanto conteúdo. Durante o mesmo período de tempo, o *Transactions* publicou 1.165 documentos (mais de 25.000 páginas), com contagens de citações variando de 0 a 12. O número médio de citações por artigo estava 0, 846 – quase duas vezes aquele do *Proceedings*.

## TRANSACTIONS OF AMS




Agora considere dois matemáticos, um que publica um documento no *Proceedings* e o outro um documento no *Transactions*. Usando algumas das práticas institucionais citadas acima, o segundo seria julgado superior ao primeiro,

publicando um documento em um periódico com o fator de impacto mais alto – de fato, duas vezes mais alto! Isto é uma avaliação válida? Os documentos no *Transaction of the AMS* são duas vezes melhores que aqueles do *Proceedings*?

Quando afirmamos que um documento individual do *Transactions* é melhor (no sentido de citações) que um documento individual do *Proceedings*, precisamos fazer uma pergunta não sobre médias, mas particularmente uma pergunta sobre probabilidades: qual é a probabilidade que nós estejamos errados? Qual é a probabilidade que um documento selecionado aleatoriamente do *Proceedings* tenha pelo menos tantas citações quanto um documento selecionado aleatoriamente do *Transactions*?

Este é um cálculo elementar, e a resposta é 62%. Isto significa que nós estamos errados 62% do tempo, e um documento selecionado aleatoriamente do *Proceedings* será exatamente tão bom quanto (ou melhor que) um documento selecionado aleatoriamente do *Transactions* – apesar que o fator de impacto do *Proceedings* seja apenas metade daquele do *Transactions*! Estamos mais frequentemente errados do que certos. A *maioria* das pessoas acha isso surpreendente, mas é uma consequência da distribuição altamente enviesada e a lacuna estreita de tempo usada para computar o fator de impacto (que é a razão para a alta porcentagem de documentos não citados)[12]. Isto mostra o valor do pensamento estatístico preciso ao invés da observação intuitiva.

Este é comportamento típico para periódicos, e não há nada especial sobre as escolhas destes dois periódicos. (Por exemplo, o *Journal of the AMS* durante o mesmo período tem um fator de impacto 2,63 – seis vezes aquele do *Proceedings*.

[12] A distribuição de obliquidade combinada com a estreita janela (usando apenas periódicos de um ano como a fonte de citações e cinco anos como o objetivo) significa que um grande número de artigos tem nenhuma ou muito poucas citações. Isto torna intuitivamente óbvio que artigos escolhidos aleatoriamente são freqüentemente equivalentes.
O fato de que muitos artigos não têm quaisquer citações (ou só algumas) também é uma conseqüência do longo tempo de citações para a matemática – artigos geralmente levam muitos anos para acumular citações. Se escolhermos períodos de tempo mais longos tanto para periódicos de fonte quanto anos alvos, então as contagens de citações aumentam substancialmente e fica mais fácil distinguir os periódicos pelo comportamento das citações. Esta é a abordagem usada em (STRINGER *et al.*, 2008) para analisar as citações. Elas mostram que para períodos de tempo suficientemente longos, a distribuição das contagens de citações para artigos individuais parece ser log-normal. Isto fornece um mecanismo para confrontar dois periódicos comparando-se as distribuições, e é certamente mais sofisticado do que usar fatores de impacto. Porém, novamente considera apenas citações e nada mais.

Contudo, um artigo selecionado aleatoriamente do *Proceedings* é pelo menos tão bom quanto um artigo do Journal, no sentido de citações, 32% do tempo).

Deste modo, enquanto é incorreto dizer que o fator de impacto não dá informações alguma sobre documentos individuais em um periódico, as informações são surpreendentemente vagas e podem estar dramaticamente equivocadas.

Segue que os tipos de cálculos realizados nos três exemplos acima – usando o fator de impacto como um substituto para contagens reais de citações para documentos individuais—tenha pouca base racional. Fazer afirmações que são incorretas mais da metade do tempo (ou um terço do tempo) seguramente não é um bom modo para realizar uma avaliação.

Uma vez que se percebe que não faz sentido substituir o fator de impacto por contagens de citações de artigo individual, segue que não faz sentido usar o fator de impacto para avaliar os autores desses artigos, os programas nos quais eles trabalham, e (mais certamente) as disciplinas que eles representam. O fator de impacto e médias em geral são muito rudimentares para fazer comparações sensatas deste tipo sem mais informações.

É claro, classificar pessoas não é a mesma coisa que classificar os seus documentos. Mas se você quiser classificar os documentos de uma pessoa usando apenas citações para medir a qualidade de um documento particular, você deve começar contando as citações daquele documento. O fator de impacto do periódico no qual o documento aparece não é um substituto confiável.

## Classificando cientistas

Enquanto o fator de impacto tem sido a melhor estatística conhecida baseada em citações, há outras estatísticas mais recentes que estão agora efetivamente promovidas. Aqui está uma pequena amostra de três destas estatísticas designadas para classificar os indivíduos.

*Índice h*: O índice *h* de um cientista é o maior *n* para o qual ele/ela publicou *n* artigos, cada um com pelo menos *n* citações.

> Esta é a mais popular das estatísticas mencionadas aqui. Foi proposta por J. E. Hirsch (HIRSCH, 2006) para medir "a produção científica de um pesquisador" ao focalizar sobre a conclusão sofisticada e perspicaz da distribuição de citações de uma pessoa. O objetivo era substituir um único número por contagens de publicações e distribuições de citações.

*Índice m*: O índice *m* de um cientista é o índice *h* dividido pelo número de anos desde o seu primeiro documento.

Este também foi proposto por Hirsch no documento acima. A intenção é compensar os cientistas mais jovens porque eles não tiveram tempo para publicar documentos ou conquistar muitas citações.

*Índice g*: O índice *g* de um cientista é o maior *n* para o qual os *n* documentos mais citados tenham um total de pelo menos $n^2$ citações.

Este foi proposto por Leo Egghe em 2006 (EGGHE, 2006). O índice *h* não leva em conta o fato que alguns documentos no topo *n* podem ter contagens de citações extraordinariamente altas. O índice *g* é designado para compensar por isto.

Há mais índices – muitos mais deles – incluindo variantes desses acima que levam em conta a idade dos documentos ou o número de autores (BATISTA *et al*., 2005; BATISTA *et al*.,2006; SIDIROPOULS *et al*., 2006).

Em seu documento definindo o índice h, Hirsch escreveu que ele propôs o índice h como "um índice facilmente computável, o qual dá uma estimativa da importância, significação, e grande impacto das contribuições cumulativas da pesquisa de um cientista" (HIRSCH, 2005, p. 5). Ele continuou acrescentando que "este índice pode fornecer um parâmetro útil para comparar diferentes indivíduos competindo para o mesmo recurso quando um critério de avaliação importante for o empreendimento científico."

Nenhuma destas afirmações é apoiada por evidência convincente. Para apoiar a sua reivindicação que o índice h mede a importância e significação da pesquisa cumulativa de um cientista, Hirsch analisa o índice h para uma coleção de vencedores do prêmio Nobel (e, separadamente, membros da Academia Nacional). Ele demonstra que as pessoas nestes grupos geralmente têm altos índices h. Pode-se concluir que é provável que um cientista tenha um alto índice h, visto que o cientista seja um ganhador do Prêmio Nobel. Mas sem informações adicionais, sabemos muito pouco sobre a probabilidade de alguém se tornar um ganhador do Prêmio Nobel ou um membro da Academia Nacional, visto que eles têm um alto índice h. Este é o tipo de informação que se deseja para estabelecer a validade do índice h.

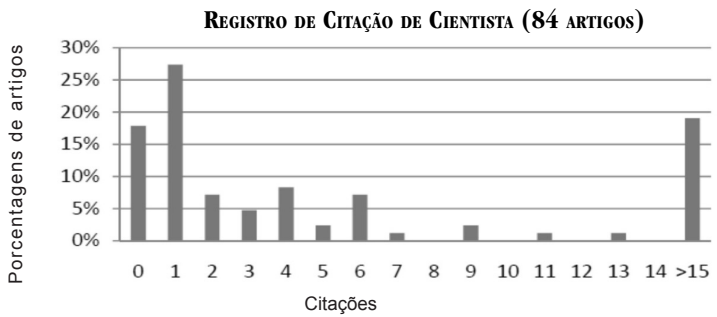
No seu artigo, Hirsch também alega que se pode usar o índice h para comparar dois cientistas:

> Afirmo que dois indivíduos com h semelhante são comparáveis em termos do seu total impacto científico, até mesmo se o seu número total de documentos ou o seu número total de citações for muito diferente. De modo inverso, que entre dois indivíduos (da mesma idade científica) com semelhante número total de documentos ou de contagem total de citações e valor h muito diferente, aquele com o h mais alto é provável de ser o cientista mais realizado (HIRSCH, 2005, p. 1).

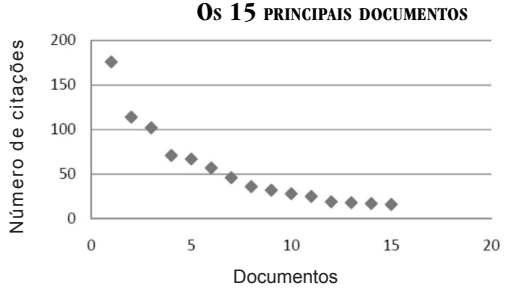
Estas afirmações parecem ser refutadas pelo bom senso. (Pense em dois cientistas, cada um com 10 documentos com 10 citações, mas um com uns 90 documentos adicionais com 9 citações cada; ou suponha que se tenha exatamente 10 documentos de 10 citações e o outro exatamente 10 documentos de 100 cada. Alguém os consideraria equivalentes?)[13].

Hirsch exalta as virtudes do índice h alegando que "o h é preferível a outros critérios de um único número geralmente usados para avaliar a produção

[13] Para ilustrar quantas informações se perdem ao usar apenas o índice h, aqui está um exemplo da vida real de um ilustre matemático em meio de carreira que publicou 84 documentos de pesquisa. A distribuição de citações *se parece com o seguinte*:




Observe que apenas abaixo de 20% das publicações tem 15 ou mais citações. A distribuição de contagens de citações reais para estes 15 *documentos é*:




Na análise de Hirsch, porém, todas estas informações são jogadas fora. Só se lembra que o índice h é 15, significando que os 15 documentos principais têm 15 ou mais citações.

científica de um pesquisador (...)" (HIRSCH, 2005, p. 1), mas ele nem define "preferível" nem explica por que se *deseja* achar "critérios de um único número".

Enquanto houve alguma crítica desta abordagem, houve pouca análise séria. Muito da análise consiste em mostrar "validade convergente", isto é, o índice h correlacionar-se bem com outras métricas de Publicações / citações, tais como o número de documentos publicados ou o número total de citações. Esta correlação é comum, visto que todas estas variáveis são funções do mesmo fenômeno básico – publicações. Em um documento digno de nota sobre o índice h (LEHMANN *et al*., 2006) os autores realizam uma análise mais cuidadosa e eles demonstram que o índice h (na verdade, o índice m) não é tão "bom" quanto meramente considerando o número médio de citações por documento. Até mesmo aqui, porém, os autores não definem adequadamente o que o termo "bom" significa. Quando o paradigma estatístico clássico é aplicado (LEHMANN *et al*., 2006), o índice h prova ser menos confiável que outras medidas.

Várias variantes do índice h foram desenvolvidas para comparar a qualidade de pesquisadores não apenas dentro de uma disciplina mas para todas as disciplinas também (BATISTA *et al*., 2006; MOLINARI, MOLINARI, 2008). Outros alegam que o índice h pode ser usado para comparar institutos e departamentos (KINNEY, 2007). Estas são frequentemente tentativas impressionantemente ingênuas para capturar um registro complexo de citações com um único número. De fato, a vantagem primária destes novos índices sobre simples histogramas de contagens de citações é que os índices descartam quase todos os detalhes dos registros de citações, e isto torna possível classificar qualquer um dos dois cientistas. Até mesmo simples exemplos, porém, mostram que a informações descartadas são necessárias para entender um registro de pesquisa. Entender realmente devia ser o objetivo ao avaliar a pesquisa, não meramente garantir que quaisquer duas pessoas sejam comparáveis

Em alguns casos, órgãos nacionais de avaliação estão coletando o índice h ou uma de suas variantes como parte seus dados. Este é um uso indevido dos dados. Infelizmente, ter um único número para classificar cada cientista é uma noção atraente – uma que pode divulgar mais amplamente a um público que geralmente interpreta mal o uso adequado de raciocínio estatístico em ambientes muito mais simples.

## O SIGNIFICADO DAS CITAÇÕES

Aqueles que promovem estatísticas de citações como a medida predominante da qualidade da pesquisa não respondem a pergunta essencial: o que significam citações? Eles reúnem grandes quantidades de dados sobre contagens de citações, processam os dados para derivar estatísticas, e então afirmam que o processo de avaliação resultante é "objetivo." Contudo é a interpretação das estatísticas que leva à avaliação, e a *interpretação* confia no *significado* das citações, o que é bastante subjetivo.

Na literatura que promove esta abordagem, é surpreendentemente difícil achar afirmações claras sobre o significado das citações.

> O conceito por trás do índice de citações é fundamentalmente simples. Reconhecendo-se que o valor das informações é determinado por aqueles que as usam, qual a melhor maneira para medir a qualidade do trabalho do que se medindo o impacto que ela causa na comunidade como um todo. A mais ampla população possível dentro da comunidade erudita (i.e. qualquer um que usa ou cita o material fonte) determina a influência ou impacto da idéia e seu criador em nosso corpo de conhecimento (THOMSON: HISTORY).

> Embora quantificar a qualidade de cientistas individuais seja difícil, a visão geral é que é melhor publicar mais do que uma quantidade menor e que a contagem de citações de um documento (relativo aos hábitos das citações no campo) é uma medida útil de sua qualidade (LEHMAN *et al*., 2006, p. 1003).

> A frequência das citações reflete o valor de um periódico e o uso feito dele (...) (GARFIELD, 1972, p. 535).

> Quando um médico (a) ou um pesquisador (a) biomédico (a) cita um artigo de periódico, isto indica que o periódico citado o (a) influenciou de alguma maneira (GARFIELD, 1987, p. 7).

> As citações são um reconhecimento de dívida intelectual (THOMSON: FIFTY YEARS).

Os termos relevantes são "qualidade", "valor", "influência", e "dívida intelectual." O termo "impacto" se tornou a palavra genérica usada para atribuir significado a citações – um termo que surgiu primeiramente em um pequeno documento escrito em 1955 por Eugene Garfield para promover a idéia de criar um índice de citações. Ele escreveu:

> Assim, no caso de um artigo muito importante, o índice de citações tem um valor quantitativo, porque ele pode ajudar o historiador a medir a influência do artigo – isto é, seu 'fator de impacto' (GARFIELD, 1955, p. 3).

Está bastante claro que aqui, como em outro lugar, o termo "fator de impacto" é destinado para sugerir que o documento citante foi "baseado" no trabalho do citado – que as citações são o mecanismo pelo qual a pesquisa se propaga.

Há uma rica literatura sobre o significado real de citações que sugere que as citações são mais complicadas do que estas declarações nos levam a acreditar. Por exemplo, em seu documento de 1983 sobre avaliar pesquisa, Martin e Irvine escrevem:

> Implícito em todos estes problemas com o uso das citações como uma medida de qualidade está a nossa ignorância das razões *por que* os autores citam fragmentos específicos de trabalho e não outros. Os problemas descritos acima... Simples análise de citações pressupõe um modelo muito racional de referência dada, no qual as citações são mantidas para refletir a apreciação primariamente científica de trabalho antecedente de alta qualidade ou importância, e citadores potenciais todos têm a mesma chance para citar documentos específicos (...) (MARTIN, IRVINE, 1983, p. 69).

Em seu documento de 1988 sobre o significado das citações, Cozzens (1989) afirma que as citações são o resultado de dois sistemas implícitos na conduta da publicação científica, um sistema de "recompensa" e o outro "retórico." O primeiro tipo tem o significado mais freqüentemente associado com uma citação – um reconhecimento que o documento citante tem "dívida intelectual" para os citados. Porém, o segundo tem um significando bastante diferente – uma referência a um documento antecedente que explica algum resultado, talvez não um resultado do autor citado em absoluto. Tais citações retóricas são meramente um modo para realizar uma conversação científica, não estabelecer compromisso intelectual. Claro que, em alguns casos, uma citação pode ter os dois significados.

Cozzens faz a observação que a *maioria* das citações é retórica. Isto é confirmado pela experiência dos matemáticos mais praticantes. (No banco de dados de citações de Revisões de Matemática, por exemplo, quase 30% de mais que 3 milhões de citações são para livros e *não* para artigos de pesquisa em periódicos). Por que isto é importante? Porque de modo diferente das citações de "recompensa", as quais tendem a referir-se a documentos embrionários, a escolha de qual documento para citar retoricamente depende de muitos fatores – o prestígio do autor citado (o "efeito halo"), o relacionamento dos autores citantes e citados, a disponibilidade do periódico (os periódicos de acesso aberto são mais prováveis de ser citados?), a conveniência de referenciar vários resultados

de um único documento, e assim sucessivamente. Poucos destes fatores estão relacionados diretamente à "qualidade" do documento citado.

Até mesmo quando as citações forem citações de "recompensa", elas podem refletir uma variedade de motivos, inclusive "moeda corrente, crédito negativo, Informações operacionais, persuasão, crédito positivo, alerta ao leitor, e consenso social" (BROOKS, 1996). Na maioria dos casos, as citações foram motivadas por mais que um destes. Alguns resultados dignos de nota podem sofrer o efeito "obliteração", imediatamente sendo incorporados no trabalho de outros, o qual serve então como a base para citações adicionais.

Outras citações não são recompensas para pesquisa importante, mas especialmente advertências sobre resultados ou pensamentos falhos. O presente relatório fornece muitos exemplos de tais citações de "advertência." A sociologia das citações é um assunto complexo – algo que está além do escopo deste relatório. Porém, até mesmo esta discussão superficial mostra que o significado das citações não é simples e que as estatísticas baseadas em citações não são quase tão "objetivas" quanto os proponentes afirmam.

Alguns poderiam argumentar que o significado das citações é secundário porque as estatísticas baseadas em citações estão muito correlacionadas com alguma outra medida de qualidade de pesquisa (tais como a revisão por pares). Por exemplo, o relatório de *Evidence* mencionado anteriormente argumenta que as estatísticas de citações podem (e devem) substituir outras formas de avaliação por causa desta correlação:

> A *Evidence* tem debatido que as técnicas de bibliometria podem criar indicadores de qualidade de pesquisa que são congruentes com a "percepção do pesquisador (EVIDENCE REPORT, 2007, p. 9).

A conclusão parece ser que as estatísticas baseadas em citações, independente do seu significado exato, devem substituir outros métodos de avaliação, porque elas geralmente concordam com eles. Com exceção da circularidade deste argumento, a falácia de tal raciocínio é fácil de ver.

## USANDO PRUDENTEMENTE AS ESTATÍSTICAS

A excessiva confiança cuidadosa sobre a métrica objetiva (estatística) para avaliar a pesquisa não é um fenômeno novo nem isolado. É descrito eloquentemente no livro popular de 2001, *Damned lies and statistics* (Malditas

mentiras e estatísticas), escrito pelo sociólogo Joel Best:

> Há culturas em que as pessoas acreditam que alguns objetos têm poderes mágicos; os antropólogos chamam estes objetos de fetiches. Em nossa sociedade, as estatísticas são um tipo de fetiche. Tendemos a considerar as estatísticas como se elas fossem mágicas, como se elas fossem mais do que meros números. Nós as tratamos como representações poderosas da verdade; agimos como se elas destilassem a complexidade e confusão da realidade em simples fatos. Usamos as estatísticas para converter problemas sociais complicados em estimativas mais facilmente compreendidas, porcentagens, e índices. As estatísticas dirigem nossa preocupação; elas nos mostram sobre o que devemos nos preocupar e o quanto devemos nos preocupar. De certo modo, o problema social se torna a estatística e, porque tratamos as estatísticas como verdadeiras e incontrovertíveis, elas alcançam um tipo de fetiche, controle mágico sobre como vemos os problemas sociais. "Pensamos em estatísticas como fatos que descobrimos e não números que criamos" (BEST, 2001, p.160).

Esta convicção mística na magia das estatísticas de citações pode ser encontrada por toda a documentação para exercícios de avaliação de pesquisa, tanto nacional quanto institucional. Também pode ser encontrada no trabalho daqueles que promovem o índice h e suas variantes.

Esta atitude também é evidente em recentes tentativas para aperfeiçoar o fator de impacto usando algoritmos matemáticos mais sofisticados, incluindo algoritmos de classificação por página, para analisar citações. (BERGSTROM, 2007; STRINGER *et al*., 2008). Seus proponentes fazem alegações sobre sua eficácia que são injustificadas pela análise e difíceis de avaliar. Por estarem baseadas em cálculos mais complicados, as (geralmente escondidas) suposições atrás deles não são fáceis para a maioria das pessoas discernir[14]. Nós temos a intenção de tratar os

[14] Bergstrom (2007) usa um algoritmo de classificação de página para dar a cada citação um peso, e então computa um "fator de impacto" usando as médias ponderadas para citações. Algoritmos de classificação de página têm mérito porque eles levam em conta o "valor" das citações. Por outro lado, a complexidade deles pode ser perigosa porque os resultados finais são mais difíceis de entender. Neste caso, todas as "autocitações" são descartadas – isto é, todas as citações de artigos em um determinado periódico J para artigos publicados em J durante os cinco anos anteriores estão descartadas. Estas não são "autocitações" em qualquer sentido normal da palavra, e um olhar de relance em alguns dados do banco de dados de Citações da Math Reviews sugere que isto descarta aproximadamente um terço de todas as citações. O algoritmo em Stringer *et al* (2008) é interessante, em parte porque tenta tratar as discrepantes escalas de tempo para as citações como também a questão de comparar documentos selecionados

números e classificações com respeito – como verdades ao invés de criações.

A pesquisa não é a primeira atividade financiada publicamente para vir sob escrutínio, e durante as décadas passadas as pessoas têm tentado realizar avaliações quantitativas de desempenho de todas as coisas dos sistemas educacionais (escolas) para assistência de saúde (hospitais e até mesmo cirurgiões individuais). Em alguns casos, os estatísticos intrometeram-se para aconselhar aqueles fazendo a medição sobre métrica sensata e o uso adequado da estatística. Se uma pessoa consultar com médicos ao praticar medicina, seguramente a pessoa deve consultar com (e atender ao conselho de) estatísticos ao praticar estatísticas. Dois excelentes exemplos podem ser encontrados em Bird (2005) e Goldstein e Spiegelhalter (1996). Enquanto cada um deles lida com avaliação de desempenho de coisas exceto pesquisa – o monitoramento de desempenho do setor público no primeiro e assistência de saúde/educação no segundo – cada um deles estabelece discernimento sobre o uso sensato das estatísticas em avaliar pesquisa.

O documento de Goldstein e Spiegelhalter em particular trata do uso de Tabelas da Liga (classificações) baseado em números simples (por exemplo, realizações estudantis ou resultados médicos), e é particularmente relevante para avaliar pesquisa ao classificar periódicos, documentos, ou autores que usam estatísticas de citações. Em seu documento, os autores esboçam uma estrutura de três partes para qualquer avaliação de desempenho:

## Dados

> Nenhuma quantidade de manobra estatística exagerada superará insuficiências básicas ou na *conveniência* ou na *integridade* dos dados coletados (GOLDSTEIN, SPIEGELHALTER, 1996, p. 389).

Esta é uma observação importante para avaliação de desempenho baseada em citações. Por exemplo, o fator de impacto está baseado em um subconjunto de dados, o qual inclui somente aqueles periódicos selecionados pelo *Thomson Scientific*. (Observamos que o próprio fator de impacto é a parte principal do critério de seleção). Alguns questionaram a integridade destes dados (ROSSNER

aleatoriamente em um periódico com aqueles de outro. Novamente, a complexidade dos algoritmos torna difícil para a maioria das pessoas avaliar os seus resultados. Uma hipótese digna de nota passou despercebida no documento na página 2: "Nossa primeira suposição é que os documentos publicados no periódico J têm uma distribuição normal de 'qualidade' (...) Isto parece contradizer a experiência comum".

*et al*., 2007). Outros mostram que outros conjuntos de dados poderiam ser mais completos (MEHO, YANG, 2007). Diversos grupos estimularam a idéia de usar o *Google Scholar* para implementar estatísticas baseadas em citações, como por exemplo o índice h, mas os dados contidos no *Google Scholar* são geralmente imprecisos (visto que coisas como nomes de autores são automaticamente extraídos de postagens pela web). As estatísticas de citações para cientistas individuais às vezes são difíceis de obter porque os autores não são exclusivamente identificados, e em alguns ambientes e certos países, isto pode ser um impedimento enorme para coletar dados de citações exatos. A coleta específica de dados que se usa para a análise de citações é geralmente negligenciada. É provável que se tire conclusões erradas de estatísticas baseada em dados errados.

## Análise estatística e apresentação

> Devemos prestar atenção específica à especificação de um *modelo* estatístico apropriado, a importância crucial da *incerteza* na apresentação de todos os resultados, técnicas para *ajuste* de resultados por fatores confundidores e finalmente a extensão a qual qualquer confiança pode ser colocada sobre *classificações* explícitas (GOLDSTEIN, SPIEGELHALTER, 1996, p. 390).

Como escrevemos anteriormente, na maioria dos casos nos quais as estatísticas de citações são usadas para classificar documentos, pessoas, e programas, nenhum modelo especial é especificado com antecedência. Ao invés disso, os próprios dados sugerem um modelo, o qual é geralmente vago. Um processo circular parece classificar mais alto os objetos porque eles são classificados mais altos (no banco de dados). Há freqüentemente atenção limitada à incerteza em *quaisquer* destas classificações, e pouca análise de como aquela incerteza (por exemplo, variações anuais no fator de impacto) afetaria as classificações. Finalmente, fatores confundidores (por exemplo, a disciplina específica, o tipo de artigos que um periódico publica, se um cientista particular é um experimentalista ou teórico) são freqüentemente ignorados em tais classificações, especialmente quando realizadas em avaliações nacionais de desempenho.

## Interpretação e impacto

> As comparações discutidas neste documento são de grande interesse público, e isto é claramente uma área onde a atenção cuidadosa para as limitações é

> tanto vital quanto provável de ser ignorada. Se resultados ajustados forem de alguma forma medidas válidas de 'qualidade' institucional é um assunto, enquanto os analistas também devem estar cientes do potencial efeito dos resultados em termos de mudanças comportamentais futuras por instituições e indivíduos que buscam melhorar a sua subsequente 'classificação' (GOLDSTEIN, SPIEGELHALTER, 1996, p. 390).

A avaliação de pesquisa *também* é de grande interesse público. Para um cientista individual, uma avaliação pode ter efeitos profundos e de longo prazo em sua carreira; para um departamento, ela pode mudar as perspectivas para sucesso num futuro distante; para disciplinas, uma coleção de avaliações pode fazer a diferença entre ter sucesso e se enfraquecer. Para uma tarefa tão importante, certamente se deve entender tanto a validade quanto as limitações das ferramentas que são usadas para realizá-la. Até que ponto as citações medem a qualidade da pesquisa? As contagens de citações parecem estar correlacionadas com a qualidade, e há uma compreensão intuitiva que artigos de alta qualidade são muito citados.

Mas conforme explicado acima, alguns artigos, especialmente em algumas disciplinas, são muito citados por razões diferentes da alta qualidade, e não segue que os artigos muito citados sejam necessariamente de alta qualidade. A interpretação exata de classificações baseada em estatísticas de citações precisa ser mais bem entendida. Além disso, se as estatísticas de citações exercem um papel central em avaliação de pesquisa, está claro que os autores, os editores, e até mesmo as editoras encontrarão modos para manipular o sistema para seu benefício (MACDONALD, KAM, 2007). As implicações em longo prazo disto não são evidentes e naturais.

O artigo de Goldstein e Spiegelhalter é valioso para se ler na época atual porque torna claro que a confiança excessiva em estatísticas ingênuas em avaliação de pesquisa não é um problema isolado. Os governos, instituições, e indivíduos lutaram com problemas semelhantes no passado em outros contextos, e encontraram maneiras para entender melhor as ferramentas estatísticas e aumentá-las com outros meios de avaliação. Goldstein e Spiegelhalter terminam seu documento com uma declaração positiva de esperança:

> Finalmente, embora tenhamos sido geralmente críticos de muitas tentativas atuais para estabelecer julgamentos sobre instituições, não desejamos dar a impressão que acreditamos que todas estas comparações estejam necessariamente erradas. Parece-nos que a comparação de instituições e a tentativa de entender por que as instituições diferem é uma atividade

> extremamente importante e é realizada melhor em um espírito de colaboração ao invés de confrontação. É talvez o único método seguro para obter informações objetivamente baseadas que podem conduzir à compreensão e no final das contas resultarem em melhorias. "*O real problema com os procedimentos simplistas que começamos a criticar é que eles distraem tanto a atenção quanto os recursos deste objetivo mais conceituado*" (GOLDSTEIN, SPIEGELHALTER, 1996, p. 406).

Seria difícil encontrar um parecer melhor para expressar os objetivos que devem ser compartilhados por todas as pessoas envolvidas na avaliação de pesquisa.

## Referências


ADLER, Robert. The impact of impact factors. *IMS Bulletin*, Vol. 36, No. 5, p. 4, 2007. http://bulletin.imstat.org/pdf/36/5

AMIN, M.; MABE, M. Impact factor: use and abuse. *Perspectives in Publishing*, n. 1, Outubro, pp. 1-6, 2000. http://www.elsevier.com/framework_editors/pdfs/Perspectives1.pdf

BATISTA, Pablo Diniz; CAMPITELI, Monica Guimaraes; KINOUCHI, Osame; MARTINEZ, Alexandre Souto. Universal behavior of a research productivity index. *arXiv*: physics, v.1, pp. 1-5, 2005. arXiv:physics/0510142v1

BATISTA, Pablo Diniz; CAMPITELI, Monica Guimaraes; KINOUCHI, Osame. Is it possible to compare researchers with different scientific interests?. *Scientometrics*, vol 68, n. 1, pp. 179-189, 2006. http://dx.doi.org/10.1007/s11192-006-0090-4

BERGSTROM, Carl. Eigenfactor: masuring the value and presitige of scholarly journals. *College & Research Libraries News*, vol 68, n. 5, May 2007. http://www.ala.org/ala/acrl/acrlpubs/crlnews/backissues2007/may07/eigenfactor.cfm (Ver também: http://www.eigenfactor.org/methods.pdf .)

BEST, Joel. *Damned lies and statistics*: untangling the numbers from the media, politicians, and activists. University of California Press: Berkeley, 2001.

BIRD, Sheila; *et al*. Performance indicators: good, bad, and ugly; Report of a working party on performance monitoring in the public services. *J.R.Statist. Soc. A*, 168, Part 1, pp. 1-27, 2005. http://dx.doi.org/10.1111/j.1467-985X.2004.00333.x

BROOKS, Terrence. Evidence of complex citer motivations. *Journal of the American Society for Information Science*, vol 37, n.. 1, pp. 34-36, 1986. http://dx.doi.org/10.1002/asi.4630370106

CAREY, Alan L.; COWLING, Michael G.; TAYLOR, Peter G. Assessing research in the mathematical sciences. *Gazette of the Australian Math Society*. Vol. 34, n. 2, maio, pp. 84-89, 2007. http://www.austms.org.au/Publ/Gazette/2007/May07/084CommsCarey.pdf

COZZENS, Susan E. What do citations count? The rhetoric-first model. *Scientometrics*, Vol 15, Nos 5-6, pp. 437-447, 1989. http://dx.doi.org/10.1007/BF02017064

EGGHE, Leo. Theory and practice of the g-index. *Scientometrics*, vol. 69, n.1, pp. 131-152, 2006. http://dx.doi.org/10.1007/s11192-006-0144-7

EVIDENCE REPORT. *The use of bibliometrics to measure research quality in the UK higher education system*, 2007. [Um relatório produzido para a Research Policy Committee of Universities (Comissão Universitária de Política de Pesquisa), Reino Unido, por Evidence Ltda., uma empresa especializada em análises e interpretação de desempenho de pequisa. Evidence Ltda. tem "aliança estratégica" com a Thomson Scientific.] http://bookshop.universitiesuk.ac.uk/downloads/bibliometrics.pdf

EWING, John. Measuring journals. *Notices of the AMS*, vol. 53, n. 9, pp. 1049-1053, 2006. http://www.ams.org/notices/200609/comm-ewing.pdf

GARFIELD, Eugene. *Citation indexes for science*: A new dimension in documentation through association of ideas. Science, 122(3159), p.108-11, julho, 1955. http://garfield.library.upenn.edu/papers/science1955.pdf

_____. Citation analysis as a tool in journal evaluation. *Science*, 178 (4060), pp. 471-479, 1972. http://www.garfield.library.upenn.edu/essays/V1p527y1962-73.pdf

_____. Why are the impacts of the leading medical journals so similar and yet so different? *Current Comments* #2, p. 3, 12 de janeiro, 1987. http://www.garfield.library.upenn.edu/essays/v10p007y1987.pdf

_____. Long-term vs. short-term journal impact (part II). *The Scientist* 12(14):12-3, 6 de julho, 1998. http://garfield.library.upenn.edu/commentaries/tsv12(14)p12y19980706.pdf

_____. Agony and the ecstasy – the history and meaning of the journal

impact factor. Presented at the *International Congress on Peer Review and Bibliomedical Publication*, Chicago, 16 de setembro, 2005. http://garfield.library.upenn.edu/papers/jifchicago2005.pdf

GOLDSTEIN, Harvey; SPIEGELHALTER, David J. League tables and their limitations: Statistical issues in comparisons of institutional performance. *J. R. Statist. Soc. A*, 159, n. 3, pp 385-443, 1996. http://links.jstor.org/sici?sici=0964-1998%281996%29159%3A3%3C385%3ALTATLS%3E2.0.CO%3B2-5 http://dx.doi.org/10.2307/2983325

HALL, Peter. Measuring research performance in the mathematical sciences in Australian universities. *The Australian Mathematical Society Gazette*, vol. 34, n. 1, pp. 26-30, 2007. http://www.austms.org.au/Publ/Gazette/2007/Mar07/26HallMeasuring.pdf

HIRSCH, J. E. An index to quantify an individual's scientific research output. *Proc Natl Acad Sci USA*, vol. 102, n. 46, pp. 16569-16573, 2006. http://dx.doi.org/10.1073/pnas.0507655102

KINNEY, A. L. National scientific facilities and their science impact on nonbiomedical research. *Proc Natl Acad Sci USA*, vol. 104, n. 46, pp. 17943-17947, 2007. http://dx.doi.org/10.1073/pnas.0704416104

LEHMANN, Sune; JACKSON, Andrew D.; LAUTRUP, Benny E. Measures for measures, *Nature*, vol 444, n. 21, pp. 1003-1004, 2006. http://www.nature.com/nature/journal/v444/n7122/full/4441003a.html

MACDONALD, Stuart; KAM, Jacqueline. Aardvark et al.: quality journals and gamesmanship in management studies. *Journal of Information Science*, vol. 33, pp. 702-717, 2007. http://dx.doi.org/10.1177/0165551507077419

MARTIN, Ben R. The use of multiple indicators in the assessment of basic research, *Scientometrics*, vol 36, n. 3, pp. 343-362, 1996. http://dx.doi.org/10.1007/BF02129599

MARTIN, Ben R.; IRVINE, John. Assessing basic research. *Research Policy*, vol 12, pp. 61-90, 1983. http://dx.doi.org/10.1016/0048-7333(83)90005-7

MEHO, Lokman; YANG, Kiduk. Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, vol 58, n. 13, pp. 2105-2125, 2007. http://dx.doi.org/10.1002/asi.20677

MOLINARI, J. F., MOLINARI, A. A new methodology for ranking scientific

institutions. To appear in *Scientometrics*, 2008. http://imechanica.org/files/paper.pdf

MONASTERSKY, R. The number that's devouring science. *Chronicle Higher Ed*. vol. 52, n. 8, 2005. http://chronicle.com/free/v52/i08/08a01201.htm

ROSSNER, Mike; VAN EPPS, Heather; HILL, Emma. Show me the data. *Journal of Cell Biology*, vol 179, n. 6, 17 de dezembro, pp. 1091-1092, 2007. http://dx.doi.org/10.1083/jcb.200711140

SEGLEN, P. O. Why the impact factor for journals should not be used for evaluating research; *BMJ*, 314:497, 15 de fevereiro, 1997. http://www.bmj.com/cgi/content/full/314/7079/497

SIDIROPOULOS, Antonis; KATSAROS, Dimitrios; MANOLOPOULOS, Yannis. Generalized h-index for disclosing latent facts in citation networks. *V1, arXiv:cs*, 2006. arXiv:cs/0607066v1 [cs.DL]

STRINGER, M. J.; SALES-PARDO, M.; NUNES AMARAL, L. A. Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE* 3(2): e1683, 2008 http://dx.doi.org/10.1371/journal.pone.0001683

THOMSON: JOURNAL CITATION REPORTS. 2007. http://scientific.thomson.com/products/jcr/

THOMSON: SELECTION. 2007. http://scientific.thomson.com/free/essays/selectionofmaterial/journalselection/

THOMSON: IMPACT FACTOR. http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/

THOMSON: HISTORY. http://scientific.thomson.com/free/essays/citationindexing/history/

THOMSON: FIFTY YEARS. http://scientific.thomson.com/free/essays/citationindexing/50y-citationindexing/