Outlier(s) nos cálculos bibliométricos: primeiras aproximações

Luís Fernando Maia Lima[*]

Alexandre Masson Maroldi[**]

Dávilla Vieira Odizio da Silva[***]

Resumo Este artigo objetiva ilustrar a análise de outlier(s) aplicada aos cálculos bibliométricos. O Outlier é uma observação, elevada ou reduzida, em relação ao conjunto de dados. São três as causas possíveis para ocorrência de outlier: registro errado, o outlier provir de outra população, ou a medição do outlier ser correta, mas representar um evento raro. Assim, o outlier pode potencialmente produzir impacto nas abordagens de estudos bibliométricos. É importante identificar o outlier para a condução de análises alternativas, com ou sem a presença do outlier.

Palavras-chaves Outlier, Bibliometria, Estatística, Cientometria, Métrica científica.

Outlier(s) in bibliometric calculations: preliminary approach

Abstract This article aims to illustrate the outlier analyses applied to bibliometrics analysis. Outlier is an observation that is unusually large or small relative to the data set. There are three possible causes to outlier: incorrect value; the outlier comes from another population, or that the measurement is correct, but represents a rare event. So, outlier can potentially have a deeper impact on the bibliometric study. It is important to identify outlier to at least to perform alternative analysis with or without the outlier.

Keywords Outlier, Bibliometric, Statistic, Scientometric, Metrology of scientific communication.

[*] Bacharel e Doutor em Engenharia Civil. Docente da Fundação Universidade Federal de Rondônia. Núcleo de Ciências Sociais Aplicadas; Departamento de Economia; Campus José Ribeiro Filho; BR 364 Km 9,5 sentido Acre; Porto Velho, Rondônia - CEP 76801-059. Telefone: (69) 2182-2107. E-mail: matematica.unir@gmail.com

[**] Bacharel e Doutor em Engenharia Civil. Docente da Fundação Universidade Federal de Rondônia. Núcleo de Ciências Sociais Aplicadas; Departamento de Economia; Campus José Ribeiro Filho; BR 364 Km 9,5 sentido Acre; Porto Velho, Rondônia - CEP 76801-059. Telefone: (69) 2182-2107. E-mail: matematica.unir@gmail.com

[***] Fundação Universidade Federal de Rondônia; Discente de Biblioteconomia; Núcleo de Ciências Sociais Aplicadas; Departamento de Biblioteconomia; Campus José Ribeiro Filho; BR 364 Km 9,5 sentido Acre; Porto Velho, Rondônia; CEP 76801-059. Telefone: (69) 2182-2107. E-mail: davillaodizio@gmail.com

Introdução

A bibliometria vem se apresentando como uma importante ferramenta para diagnosticar o estado das ciências por meio da produção da literatura científica em um determinado nível de especialização (MACIAS-CHAPULA, 1998).

Para Rostaing (1996, p.17), a bibliometria consiste na "aplicação de métodos estatísticos ou matemáticos sobre o conjunto de referências bibliográficas". Esse contexto é corroborado por Spinak (1998) ao considerar que a bibliometria utiliza-se de análises estatísticas para, entre outras coisas, mensurar a complexidade de um determinado meio científico por instrumentos quantitativos da produção de documentos elencados em suas bibliografias.

Atualmente a bibliometria vem sendo considerada um campo de estudos interdisciplinar, estendendo-se a vários domínios científicos, entre eles a matemática, as ciências sociais, a engenharia e a estatística. (GLÄNZEL, 2003). Para Vinkler (2010) os aspectos bibliométricos são quantificáveis, via métodos estatísticos, para se detectar as características ou fenômenos para a ciência e relacionados à ciência. Assim, cabe ao bibliometrista dominar as áreas que nutrem de aportes a bibliometria, ou seja, o mesmo deverá possuir conhecimentos dos "principais recursos estatísticos para a produção de indicadores" (SILVA; HAYASHI; HAYASHI, 2011, p. 125) para conduzir a resultados mais consistentes dos dados coletados.

Pesquisadores de diversas áreas utilizam a bibliometria para produzir indicadores e desvelar o atual estágio de desenvolvimento de uma determinada ciência, o que é importante para verificar nos cálculos estatísticos e bibliométricos a possibilidade de ocorrência de outliers[1], pois os mesmos podem conter informações adicionais sobre o objeto de estudo ou mascarar resultados, induzindo a generalizações incorretas.

Para se identificar outliers, deve ser realizada a Análise Exploratória de Dados (AED), a qual utiliza meios gráficos e resumos numéricos para investigar um conjunto de dados com o objetivo de compreender as características fundamentais e obter a maior quantidade possível de informação destes dados e as relações entre os mesmos. Em seguida, o bibliometrista realiza a AED, com objetivo de compreender e calcular (TRIOLA, 2008; MOORE, 2000; BUSSAB; MORETTIN, 2002):

- medidas do centro de uma distribuição (média, mediana e moda), para indicar um valor representativo da localização do meio do conjunto de dados;

- medidas de variação (desvio padrão), para mensurar de quanto os valores dos dados variam entre eles;

- forma (gráfico) da distribuição de frequência (histogramas, diagrama de ramos e folhas, *boxplot* ou diagrama de caixa), estas ilustrações auxiliam na visualização de assimetria, curtose e possíveis outliers);

[1] Optamos em utilizar o termo *outlier* por considerarmos que o mesmo encontra-se totalmente ratificado na Estatística.

- outliers: valores amostrais (muito elevados ou muito reduzidos, ou seja, valores discrepantes) que diferem significativamente dos outros dados;

- variação ao longo do tempo, como as séries temporais, ou seja, dados cujas características que alteram-se no curso temporal.

Compreendido que os cálculos bibliométricos dependem da estatística, e que na AED há o interesse em outliers, e apoiado que nas análises dos dados da ciência as interpretações são mais ricas e interessantes quanto maior é o número de métodos distintos empregados, pois estes conduzem a esclarecimentos complementares (CALLON; COURTIAL; PENAN, 1995), nossa pesquisa pode ser representada pelas seguintes questões: como o(s) outlier(s) pode(m) influenciar os indicadores bibliométricos? Há informações adicionais contidas no(s) outlier(s) por ventura existente(s) nos cálculos bibliométricos?

Outliers: definição, histórico e fundamentos

O outlier é o valor significativamente distinto de um conjunto de observações, ou valor discrepante em relação aos restantes dos dados observados (BARNETT; LEWIS, 1994; MOORE, 2000; DOWLING; CLARK, 2005; MARTINS, 2005; LEVINE et al, 2008; TRIOLA, 2008).

O interesse nos estudos de outliers é um dos mais antigos em estatística. A literatura na área, desde os séculos XVIII e XIX mostra que já havia preocupação com os outliers (HAWKINS, 1980). Assim, o conceito de outlier atraiu a atenção dos analistas a partir das primeiras tentativas de interpretação de um conjunto de dados (BARNETT; LEWIS, 1994).

Os primeiros esboços consideravam o outlier como erros ou enganos que afetavam os dados, sendo então o outlier "retificado" aos dados ou excluído completamente da análise (BARNETT; LEWIS, 1994).

Em 1778, Daniel Bernoulli, um dos grandes expoentes das ciências exatas, criticou a prática difundida pelos astrônomos da época de alijar os outliers dos cálculos estatísticos. Então, nos primórdios do advento de cálculos estatísticos, havia uma espécie de aversão aos outliers, e estes desconsiderados da análise (HAWKINS, 1980).

Inicialmente, o critério de detecção de outliers era a opinião do analista se os dados eram de fato, próximos uns dos outros. Ao longo do tempo, este critério subjetivo para detectar outliers foi amenizado por propostas de critérios objetivos, começando em 1852 com Pierce. Outras propostas foram as de Chauvenet em 1866, Stone em 1868, Goodwin em 1913, Irwin em 1925, Thompson em 1935, Pearson e Chandra Sekar em 1936 (HAWKINS, 1980).

Modernamente, há diversas abordagens para outliers, entre elas (BARNETT; LEWIS, 1994):

- "rejeição" do mesmo (perda completa da informação);

- diminuir sua influência nos cálculos estatísticos via "acomodação" do outlier. Por exemplo, para o cálculo do centro da distribuição usar a mediana ou a média aparada (as quais são denominadas procedimentos robustos);

- identificar o outlier como principal aspecto na análise dos dados, e possíveis explicações para a ocorrência do mesmo;

- métodos *ad hoc* para identificar outliers via AED.

E, quais as possíveis causas para o aparecimento dos outliers? De acordo com McClave, Benson e Sincich (2001) o outlier:

- pode originar-se de um registro errado, neste caso, deve-se retirar ou se possível, retificar o valor do outlier do conjunto de dados (MARTINS, 2005).

- provém de uma população distinta à observada, por exemplo, estudando peso de judocas, pode-se registrar que praticamente todos são pesos penas, exceto um ou dois classificados como pesos pesados. Este um ou dois pesos pesados são então outliers, fenômeno denominado de "contaminação" (BARNETT; LEWIS, 1994), pois estes pesos pesados pertencem a uma distribuição que é distinta dos outros dados que pertencem a um grupo homogêneo (os pesos leves).

- é um valor correto, todavia um evento raro, representando a inevitável variabilidade dos dados. Para se lidar com outlier é necessário realizar a "acomodação" do outlier nas análises; conduzir os cálculos com a presença dos outliers, e estes têm os mesmos pesos que os outros dados; e, identificar os outliers o que é primordial, pois eles revelam importantes características, informações novas, modelos de comportamento ou atitudes especiais de um grupo que é heterogêneo em relação aos demais dados (BARNETT; LEWIS, 1994). Esta última abordagem é reforçada por Triola (2008, p. 97) por ser "importante, em geral, investigar mais profundamente o conjunto de dados para identificar quaisquer características notáveis, especialmente aquelas que possam afetar fortemente os resultados e conclusões". Havendo os outliers estes exercem importante função nos cálculos de inferência estatística. Ou seja, uma vez detectado o outlier deve-se encontrar um motivo que pode ser o erro ou a natureza especial desta observação (MCCLAVE; BENSON; SINCICH, 2001; MOORE, 2000).

Nesse contexto observamos a importância dos outliers para os cálculos bibliométricos, seja ou por um grupo homogêneo e outro heterogêneo ou porque há outliers com baixa probabilidade de ocorrência, mas que representam informações adicionais nas análises.

E de que forma os outliers podem influenciar os cálculos bibliométricos? Os outliers podem mascarar drasticamente os cálculos da média (medida de centro), desvio padrão (medida de variação) e também a forma do histograma (BUSSAB; MORETTIN, 2002; MOORE, 2000; LEVINE et al, 2008; TRIOLA, 2008). É importante salientar que esta influência dos outliers na descrição dos dados também se expande naturalmente às inferências estatísticas. Para o caso de dados multivariados, os outliers podem aparecer, por exemplo, em análise de regressão e séries temporais (BARNETT; LEWIS, 1994), as quais também são usadas nas análises bibliométricas. Há então intrínseca interação dos outliers com as outras características objetos da AED (medidas de centro, variação, histograma e variação temporal).

Para a detecção de outliers via AED, realizam-se usualmente cálculos envolvendo: a) média e desvio padrão (LEVINE et al., 2008; MARTINS; DOMINGUES, 2011); b) quartis e intervalo interquartil (TRIOLA, 2008). Em tais cálculos, há os denominados outliers potenciais e outliers extremos. Visualmente podemos suspeitar da existência de outliers pelos gráficos de ramo e folha ou por meio de *boxplot* ou diagrama de caixa. Havendo os outliers, devem-se estudar os seus efeitos na distribuição dos dados via construção gráfica e resumos estatísticos com e sem a presença dos outliers (TRIOLA, 2008).

Outliers: algumas aplicações

Encontramos em Araújo (2010) explanação de diversas aplicações dos outliers, quando identificados: detecção de fraudes, e de invasões em sistemas computacionais e aplicações (monitoramento) médicas; diagnóstico de falhas; e, gerenciamento de empréstimos financeiros. Para ilustrar o caso de detecção de fraudes: genericamente um correntista possui um determinado "padrão" de saques de sua conta corrente mensalmente. Pode ocorrer então uma retirada de valor elevado (o outlier), então cabe à instituição bancária entrar em contato com este cliente para verificar se ele reconhece esta retirada atípica (outlier) ou, se houve fraude, bloquear imediatamente a conta corrente do sujeito.

Araújo (2010) também investigou a detecção de outliers em redes complexas, não apenas em nós periféricos, como em nós centrais, sendo os outliers identificados classificados como rotuladores de comunidade semissupervisionada, onde os nós outliers centrais são bons propagadores de informação e os nós outliers periféricos situam-se em torno da borda da comunidade.

Os outliers podem indicar também alguma utilidade não esperada em tratamento industrial, ou sucesso surpreendente em certa cultura vegetal, ou de período de gestação de mulher que pode ter cometido suposto adultério e estar grávida: de fato o marido é o pai, mas o período de gravidez é um evento raro ou pode ter ocorrido que o marido não seja o pai, se de fato houve um deslize da esposa, daí o período de gestação incomum, pois ocorreu a concepção mais tarde (BARNETT; LEWIS, 1994).

Por sua vez, Pinheiro et al. (2009) cita o outlier como medida para detectar fraudes no imposto de renda. Os autores apresentam exemplo numérico com dados de telefones fixos em cada Estado do Brasil em 2001, sendo os cálculos realizados com e sem a presença dos outliers.

Bussab e Morettin (2002) ilustram que entre as 30 cidades com maiores populações, São Paulo e Rio de Janeiro são considerados outliers. Levando em conta todas as cidades, a média da população é 145,4 mil habitantes por cidade. Mas excluindo-se São Paulo e Rio de Janeiro, a média torna-se 100,6 mil habitantes por cidade (uma redução de 30,8% em relação a todos os dados). Assim, reitera-se a influência dramática dos outliers sobre a média (dada pela redução já mencionada), além de neste caso os outliers indicarem cidades com elevadas concentrações de pessoas (São Paulo e Rio de Janeiro). Outra pesquisa similar envolvendo população é dada por Guimarães (2008).

Os outliers também foram utilizados nas estatísticas de segurança pública na cidade de Belém,PA, para identificação dos bairros mais violentos com relação ao atentado pudor, atos obscenos e estupros (RAMOS et al, 2008), para violência contra mulheres (PAMPLONA et al, 2008), e mapear bairros com valores extremos de acidentes de trânsito (PEREIRA et al, 2008).

Para as 50 empresas high-tech, McClave e Benson e Sincich (2001) calcularam os seus gastos percentuais em pesquisa e desenvolvimento. Constataram que duas empresas eram outliers, cuja informação relevante era que se tratava de empresas jovens e em franco crescimento.

Levin e Rubin (1998) analisaram os resultados econômicos de 224 companhias citadas por um jornal em fevereiro de 1990. Dentre estas empresas, duas foram outliers, e na busca pelos motivos, uma delas havia recebido elevados valores pela venda de suas operações, enquanto outra apresentava custos exorbitantes por haver encerrado suas operações. Os autores recomendaram a retirada destes dois valores extremos dos demais dados.

Triola (2008) informa que um tipo de distribuição assimétrica à direita ocorre quando se estuda a renda anual dos indivíduos. A distribuição é justamente assim, pois "há umas poucas pessoas que ganham milhões de dólares por ano" (TRIOLA, 2008, p. 70). Estas "poucas pessoas" provavelmente serão outliers, ou seja, neste caso, as pessoas mais ricas do conjunto de dados estudados.

Lima, Maroldi e Silva (2012a) trabalharam uma análise estatística dos grupos de pesquisa em desenvolvimento regional existentes no diretório de grupos de pesquisa do Conselho Nacional de Pesquisa e Desenvolvimento. Na parte relativa à produção bibliográfica do primeiro líder, quantidade de pesquisadores nos grupos, e número de estudantes por grupo, Lima, Maroldi e Silva (2012a) utilizaram outliers para separar os líderes que se destacavam dos demais em termos de publicação, bem como para destacar os grupos que possuíam elevado número de pesquisadores ou estudantes por grupo de pesquisa.

Outliers e bibliometria

No campo da biologia como na área da ciência da informação, De Bellis (2009) assevera que os dados podem demonstrar características de assimetria, em virtude do aparecimento de valores extraordinários, fora dos valores "padrões", os quais elevam a variabilidade dos dados sugerindo alguma causa substancial. Nestas distribuições assimétricas há uma coexistência de valores baixos, mas com grande probabilidade de ocorrência (por exemplo, para determinado campo de conhecimento na publicação de somente um artigo, que é um valor baixo, há muitos autores, ou seja, probabilidade elevada); e, um núcleo de valores elevados, mas com baixa probabilidade (no mesmo campo de conhecimento, para publicar grande quantidade de artigos, o qual é um valor elevado, há poucos autores que conseguem isto, ou seja, há reduzida probabilidade de ocorrência).

De Bellis (2009) ainda salienta que as questões dos modelos assimétricos apresentam um desafio aos cientistas da informação, pois estes utilizam as estatísticas descritiva e inferencial,

via média aritmética e desvio padrão, o que pode prejudicar conclusões e testes de significância estatísticos, caso a assimetria (e também a natureza dos outliers) não seja levada em consideração.

Então, para este padrão de dados assimétricos, o qual De Bellis (2009) denomina de natureza "bipolar" dos dados, a cauda da distribuição deve receber um tratamento diferente, justamente por representar valores não usuais ou extremos (outliers). Deve-se então estar atento para outras possibilidades de ocorrências de outliers nos cálculos bibliométricos.

Para Portal (2005) os métodos quantitativos nas métricas científicas são:

- comportamento matemático de regularidades (Leis de Lotka, Price, Bradford, Zipf, etc). No caso destas leis, De Bellis (2009) reitera o caráter assimétrico dos dados, e existe, portanto, a possibilidade de outliers.

- ao fluxo de informação documental (produção e comunicação científica). Vislumbra-se a aplicação dos outliers aos denominados indicadores específicos (VINKLER, 2010; ANDRÉS, 2009), como por exemplo: autores/artigo; citação/artigo; autor/citação; citação/pesquisadores; artigos/pesquisadores; artigos/autores; citação/autor; autor/pesquisadores, citação/periódico, pois nestes indicadores bibliométricos há cálculos de média e desvio padrão, e, neste caso é importante a identificação dos outliers e seus possíveis efeitos nos indicadores.

- prognóstico, simulação e quantificação (cadeias de Markov, análise de regressão, modelos multivariados e outros). Nos modelos multivariados (análise de regressão e séries temporais) podem ocorrer outliers.

Para Sanfelice (2007) que estudou uma amostra de 182 empresas transnacionais que apresentavam maior dispêndio em pesquisa e desenvolvimento no Brasil, China e Índia, sendo o foco do trabalho a atividade científica de tais empresas nos três países existente na base de dados da *Science Citation Index Expanded*. Inicialmente foi retirada da amostra selecionada a empresa indiana Tata, que publicava nos três países, mas com enorme quantidade de publicações na Índia e que iria falsear a conclusão do trabalho.

Outro aspecto que Sanfelice (2007) detecta é que das cinco empresas que mais publicam artigos na China, a empresa Microsoft é considera outlier, sendo uma possível causa o elevado aporte de recursos financeiros na China. Por sua vez, na Índia, o outlier é a empresa IBM. Portanto, o outlier aqui foi tratado como informação adicional.

Já Martins (2010) procurou avaliar a qualidade de conferências científicas, aplicando seu método proposto a conferências na área de Ciência da Computação e retirou dos dados originais os outliers, que significavam eventos com baixo impacto entre os pesquisadores da área ou eventos que versavam sobre temas específicos. Estes outliers poderiam conduzir a interpretação incorreta dos dados da qualidade das conferências e assim também influir em novas classificações das conferências. Neste caso, o outlier depois de identificado foi excluído dos cálculos.

Lima, Maroldi e Silva (2012b) pesquisaram 77 dissertações de mestrado em Biologia da Fundação Universidade Federal de Rondônia, visando verificar o número de citações em língua inglesa contidas nas referências de cada dissertação, detectaram pelo menos um outlier. Este

outlier (calculado via quartis e intervalo interquartil) significava uma dissertação em que o autor havia citado elevado número de publicações em língua inglesa nas suas referências. Aqui, o outlier possui uma característica especial e foi apenas identificado.

Em outra pesquisa, Maroldi, Lima e Souza (2012) estudaram os autores que compunham o grupo de elite da Psicologia Escolar por meio da Lei do Elitismo de Derek de Solla Price, e no mesmo trabalho apresentaram proposta de um método empírico baseado nos outliers (uso dos quartis e intervalo interquartil) para detectar o grupo de elite, onde o outlier significa o grupo de autores que se destacaram com elevada publicação de trabalhos em relação aos outros autores. Maroldi, Lima e Souza (2012) recomendam ou reiteram outros estudos com o método empírico dos outliers na detecção de grupo de elite.

Esta associação de outliers com o estudo do grupo de elite de determinada área de conhecimento pode ser visualizada mediante a lei de Lotka, que em geral apresenta a forma de um "J" invertido: muitos autores publicam não mais que um artigo, e somente um pequeno número de autores publicam quantidade considerável de artigos (CALLON; COURTIAL; PENAN, 1995; DE BELLIS, 2009; VINKLER, 2010). Portanto, a este grupo reduzido que publica bastante associamos os outliers.

Andrés (2009) comenta que quando no conjunto de dados obtidos encontram-se somente valores unitários por classe na extremidade superior da distribuição (ou seja, em cada classe há somente um autor publicando elevados números de artigos, que culmina no denominado grupo de autores prolíficos), estes valores devem ser excluídos dos dados para não superestimar os resultados obtidos. Sem dúvidas, tais valores excluídos tratam-se dos outliers. Este aspecto de exclusão dos dados de elevadas publicações com autoria unitária (outliers) é ilustrado por Andrés (2009) com base nos dados de trabalho publicado por Pulgarín e Gil-Leiva em 2004. Neste caso, a exclusão de valores unitários é a sua natureza subjetiva, pois na hipótese de na última classe (o de maior valor) haver frequência ou ocorrência distinta da unidade, então o critério de remoção do valor extremo fornecido por Andrés (2009) não se aplica, e deve-se então levar em consideração todos os dados na análise. Mas se há um critério objetivo (para detectar os valores extremos), então os dados discrepantes podem ser excluídos.

Nos estudos de rede de colaboração entre pesquisadores (ANDRÉS, 2009), ou entre redes de citação entre as publicações periódicas, ou entre rede de relação entre palavras chaves (CALLON; COURTIAL, PENAN, 1995), redes cibersociais ou cibersociedades (SANTIAGO, 2003), há possibilidade dos outliers serem aplicados e identificados com ideias similares empregadas por Araújo (2010), onde os outliers centrais são bons propagadores de informação e os nós outliers periféricos situam-se na borda da comunidade analisada.

Também há possibilidade de aplicar outliers para os diagramas estratégicos, que são representações de estrutura de um campo de informação, visto que estes diagramas estratégicos são similares aos gráficos de análise de regressão (CALLON; COURTIAL; PENAN, 1995).

Na área de cibermetria, há o termo diâmetro Web definido como a distância máxima para alcançar um determinado documento (BERROCAL; FIGUEROLA; ZAZO, 2003), onde domínios que possuem o valor de diâmetro mais alto refletem zonas de sua estrutura que são mais difíceis de alcançar, e nos seus dados coletados, os autores encontraram que os diâmetros

Web são similares entre os diversos domínios, salvo algumas exceções. Justamente estas exceções podem ser os outliers.

Nos estudos bibliométricos para rankeamento (classificação) de periódicos, podemos encontrar os mais produtivos, Andrés (2009) ilustra o caso da pesquisa em neurociência no período de 2004 a 2008, onde identificamos entre os mais produtivos, a presença de um outlier, no caso o periódico *Neuroscience Research*.

É interessante observar que Andrés (2009) também nos apresenta uma tabela de evolução temporal de 1981 a 2005, dividida a cada cinco anos, com base no trabalho publicado no *Journal of Economic Psychology* de autoria de Kirchler e Hölzl em 2006, onde aparece o conceito de média aparada (que pode "acomodar" ou diminuir os efeitos de eventuais outliers) nos temas: páginas por artigo, referências por artigo e autores por artigo.

No caso hipotético de produção de autores de distintas instituições, De Bellis (2009) reforça a questão da assimetria (com presença de outlier), pois há um núcleo central com muitas universidades com baixa produção, e outro núcleo, disperso e com poucas instituições, mas que são altamente produtivas.

De Bellis (2009) ainda argumenta que uma análise estatística encontraria sérias dificuldades em produzir conclusões gerais, justamente pelo fato das universidades altamente produtivas, tais como Harvard, Cambrigbe, Stanford ou MIT (ou seja os outliers) influírem nos cálculos padrões de média e desvio padrão, conduzindo a possíveis conclusões incompletas.

Considerações finais

Este trabalho procurou alertar aos pesquisadores da área de bibliometria a importância que deve ser dada aos outliers, pois estes podem apresentar características e informações adicionais. Em geral os trabalhos bibliométricos limitam-se somente ao cálculo da média aritmética, e não raro, ao desvio padrão, e poucas publicações procedem ao resumo dos cinco números: valores mínimo, primeiro quartil, mediana, terceiro quartil e máximo.

Portanto, caso o bibliometrista realize a AED com seus dados há a possibilidade de já identificar possível(is) outlier(s), e então decidir sobre: rejeitar qualquer valor que seja o outlier ou manter integralmente os valores dos dados; ou proceder a "acomodação" (diminuir efeitos) do outlier nas análises; ou pesquisar se há informação adicional contida nos outliers. É recomendável fazer diversas simulações com ou sem a presença dos outliers para que as conclusões do estudo sejam confiáveis.

Observamos que há possibilidade de aplicação dos outliers nas métricas científicas, como detecção de grupo de elite, maiores citações por trabalho (periódico, teses, dissertações), casos de exceção em diâmetro Web, entre outros.

Enfim, recomendamos que novas pesquisas sobre os efeitos dos outliers nas análises bibliométricas (e provavelmente nas outras métricas científicas) sejam realizadas, visto ser este tema pouco explorado pelos analistas.

Artigo recebido em 04/02/2013 e aprovado em 20/03/2013.

Referências

ANDRÉS, A. *Measuring academic research:* how to undertake a bibliometric study. Oxford: Chandos, 2009.

ARAÚJO, B. M. de. *Identificação de outliers em redes complexas baseado em caminhada aleatória.* 2010. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional)- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2010. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-06102010-141931/>. Acesso em: 02 dez. 2012.

BARBA, B. M. *Los indicadores bibliométricos:* fundamentos y aplicación al análisis de la ciencia. Gijón: Trea, 2003.

BARNETT, V.; LEWIS, T. *Outliers in statistical data.* New York: John Wiley & Sons, 1994.

BERROCAL, J. L. A.; FIGUEROLA, C. G.; ZAZO, A. F. *Cibermetría:* nuevas técnicas de estúdio aplicables al web. Gijón: Trea, 2003.

BUSSAB, W. de O.; MORETTIN, P. A. *Estatística básica.* 5. ed. São Paulo: Saraiva, 2002.

CALLON, M.; COURTIAL, J-P.; PENAN, H. *Cienciometria:* el studio cuantitativo de la actividade científica: de la bibliometría a la vigilância tecnológica. Gijón: Trea, 1995.

DE BELLIS, N. *Bibliometrics and citation analysis:* from the Science Citation Index to cybermetrics. Maryland: Scarecrow Press, 2009.

DOWLING, D.; CLARK, J. *Estatística aplicada.* 2. ed. São Paulo: Saraiva, 2005.

GUIMARÃES, P. R. B. *Métodos quantitativos estatísticos.* Curitiba: IESDE, 2008.

HAWKINS, D. M. *Identification of outliers.* London: Chapman and Hall, 1980.

LEVIN, R.; RUBIN, D. S. *Statistics for management.* 7. ed. New Jersey: Prentice Hall, 1998.

_____ et al. *Statistics for managers using Microsoft Excel.* 5. ed. New Jersey: Prentice Hall, 2008.

LIMA, L. F. M., MAROLDI, A. M., SILVA, D. V. O. Análise estatística dos grupos de pesquisa em desenvolvimento regional. In: JORNADA CIENTÍFICA CEDSA, 7., 2012. *Resumos...* Porto Velho: [s.n.]. 2012a. Cd-rom.

_____. Análise de citações em literatura inglesa nas dissertações do programa de mestrado em Biologia da Universidade Federal de Rondônia. In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 3., 2012, Gramado. *Resumos...* Gramado: [s.n.]. 2012b.

MACIAS-CHAPULA, C. A. O papel da infometria e da cientometria e sua perspectiva nacional e internacional. *Ciência da Informação,* v. 27, n. 2, p. 134-140, 1998.

MAROLDI, A. M.; LIMA, L. F. M., SOUZA, A. M. de L. Elitismo na psicologia escolar: um estudo a partir da produtividade de seus autores. In: SEMINÁRIO DE PSICOLOGIA: 50 anos de psicologia no Brasil: a produção do conhecimento e as desigualdades regionais, 2., 2012, Porto Velho. *Resumos...* Porto Velho: [s.n.]. 2012.

MARTINS, G. de A. *Estatística geral e aplicada.* 3. ed. São Paulo: Atlas, 2005.

_____; DOMINGUES, O. *Estatística geral e aplicada.* 4. ed. São Paulo: Atlas, 2011.

MARTINS, W. M. *Abordagens para avaliação automática da qualidade de conferências científicas:* um estudo de caso em ciência da computação. 2010. Dissertação (Mestrado em Ciências de Computação)- Universidade Federal de Minas Gerais, 2010. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/bitstream/handle/1843/SLSS-7WFQ2F/waistersilvamartins.pdf?sequence=1>. Acesso em: 02 dez. 2012.

MCCLAVE, J. T.; BENSON, P. G.; SINCICH, T. *Statistic for business and economics.* 8th ed. New Jersey: Prentice Hall, 2001.

MOORE, D. S. *A estatística básica e sua prática.* Rio de Janeiro: LTC, 2000.

PAMPLONA, V. M. S. et al. O perfil da vítima de crimes contra a mulher na região metropolitana de Belém. In: RAMOS, E. M. L. S.; ALMEIDA, S. dos S. de; ARAÚJO, A. dos R. (Org.). *Segurança pública:* uma abordagem estatística e computacional. Belém: EDUFPA, 2008. V. 2.

PEREIRA, V. S. de P. et al. Estudo estatístico dos acidentes de trânsito fatais, no município de Belém-PA, no ano de 2006. In: RAMOS, E. M. L. S.; ALMEIDA, S. dos S. de; ARAÚJO, A. dos R. (Org.). *Segurança pública:* uma abordagem estatística e computacional. Belém: EDUFPA, 2008. V. 1.

PINHEIRO, J. I. D. et al. *Estatística básica:* a arte de trabalhar com dados. Rio de Janeiro: Elsevier, 2009.

PORTAL, S. G. *Modelo teórico para el estudio métrico de la información documental.* Gijón: Trea, 2005.

RAMOS, E. M. L. S. et al. Atentado violento ao pudor, ato obsceno e estupro, ocorridos na região metropolitana de Belém. In: RAMOS, E. M. L. S.; ALMEIDA, S. dos S. de; ARAÚJO, A. dos R. (Org.). *Segurança pública:* uma abordagem estatística e computacional. Belém: EDUFPA, 2008. V. 1.

ROSTAING, H. *La bibliométrie et ses techniques.* Tolouse: Scienses de la société, 1996.

SANFELICE, V. *A atividade científica de empresas transnacionais instaladas no Brasil medida por meio de indicadores bibliométricos.* 2007. Trabalho de Conclusão de Curso (Graduação em Matématica Aplicada à Negócios)- Universidade de São Paulo, 2007.

Disponível em: <http://dcm.ffclrp.usp.br/man/upload/Sanfelice_V.pdf>. Acesso em: 15 dez. 2012.

SANTIAGO, L. G. *Extraer y visualizar información em Internet:* el web mining. Gijón: Trea, 2003.

SILVA, M. R. da; HAYASHI, C. R. M.; HAYASHI, M. C. P. I. Análise bibliométrica e cientométrica: desafios para os especialistas que atuam no campo. *InCID,* v. 2, n. 1, p. 110-129, jan./jun. 2011.

SPINAK, E. Indicadores cienciométricos. *Ciência da Informação,* v. 27, n. 2, p. 141-148, maio/ago.1998.

TRIOLA, M. F. *Introdução à estatística.* 10. ed. Rio de Janeiro: LTC, 2008.

VINKLER, P. *The evaluation of research by scientometric indicators.* Oxford: Chandos, 2010.