em questão
E-ISSN 1808-5245
Revista da Faculdade de Biblioteconomia e Comunicação da UFRGS
v. 20, n.3 – Edição Especial 2014

Outliers na Lei do Elitismo

Dávilla Vieira Odízio da Silva
Graduanda; Universidade Federal de Rondônia;
davilla_jp@hotmail.com

Alexandre Masson Maroldi
Doutorando; Universidade Federal de São Carlos;
alexandremaroldi@yahoo.com.br

Luís Fernando Maia Lima
Doutor; Universidade Federal de Rondônia;
matematica.unir@gmail.com

Resumo: Visualizando os autores mais produtivos (elite) como *outliers* ou valores discrepantes, este trabalho propõe uma nova maneira para quantificar essa elite usando-se os conceitos da Análise Exploratória de Dados, onde há a detecção de *outliers*. A Lei do Elitismo ou raiz quadrada, proposta por Price, do total de autores apresenta os seguintes problemas: nem sempre fornece um número inteiro, e, além disso, há dificuldade também de se enquadrar o valor teórico encontrado com os dados reais. Já o método dos *outliers* permite a identificação unívoca dos autores que compõem a elite. Ao final do trabalho, concluímos que o método *outliers* pode até gerar valores fracionados para o valor da quantidade de produção que define a elite, todavia, fornece um valor único para a identificação dos autores que compõem a elite de uma determinada área do conhecimento.

Palavras-chave: *Outlier*. Análise Exploratória de Dados. Lei do Elitismo. Bibliometria. Estatística.

1 Introdução

Ao compulsarmos a literatura acerca do que é a ciência, vemos que Schwartzman (1979[1], p. 6 apud CARVALHO, 2011, p. 5) define que a ciência é "[...] um conjunto de conhecimentos a respeito das coisas, conhecimentos que se desenvolvem, acumulam, se transformam e se reestruturam em função de uma lógica própria de uma organização do conhecimento – seu *logos*".

Nesse aspecto acumulativo do conhecimento, Vanti (2012) assevera que a ciência, para atingir o nível de desenvolvimento e de credibilidade que apresenta nos dias de hoje, buscou nos números os alicerces para as suas teorias. A preocupação com a mensuração na ciência vem de tempos remotos. Teve início na Antiguidade, quando os filósofos, matemáticos e astrônomos tentavam dimensionar a distância entre as estrelas para, a partir daí, estabelecer as medidas de tempo em dias, meses, anos e em horas e minutos.

Em 1842 Joseph Rodes Buchaman criou o termo psicometria como um meio para medir a alma dos objetos. Com o passar do tempo, com os novos descobrimentos sobre as variações do comportamento humano, o termo psicometria passou a ser visto como uma disciplina métrica interessada na mensuração psicológica, com duas vertentes: uma teórica e outra prática. Já por volta dos anos de 1930, surge o termo econometria como a aplicação de técnicas estatísticas a economia como prova de validação das teorias econômicas, bem como de seus prognósticos e tendências (GORBEA PORTAL, 2005). Podemos então perceber que o homem sempre buscou na matemática e/ou na estatística as explicações para os vários fenômenos sociais que o inquietavam.

De fato, não foram somente a Psicologia e a Economia que buscaram nos números perspectivas quantitativas para suprir a necessidade de mensurar seu desenvolvimento, também os cientistas da informação são outro exemplo dessa busca. Para Mugnaini, Carvalho e Campanatti-Ostiz (2006, p. 316) "A amplitude da ciência produzida em um país pode ser apontada pela mensuração de sua produção bibliográfica e a representação deste tipo de dado é um dos papéis da Ciência da Informação", que se desenvolve como área interdisciplinar voltada para organização e estruturação dos documentos científicos, bem como registro, acesso, recuperação e uso da informação (MOSTAFA, 2004).

Um dos primeiros autores a pensar no crescimento vertiginoso das publicações e procurar mensurá-la foi Price (1976), que relata a passagem da "pequena ciência" para "grande ciência" ao explicar que o crescimento da produção científica pode ser avaliado pela quantificação de dados.

Price (1976, p. 3) detectou ainda que o "[...] crescimento da ciência é surpreendentemente rápido [...]" e exponencial, identificando que os resultados

dessa atividade quadruplicam a cada geração, e a literatura científica dobra num período de 10 a 15 anos.

Derek de Solla Price também deixou em seu legado o fato de perceber que a ciência se fundamenta em saberes acumulados, ou seja, a cada nova geração de cientistas, maior será o potencial de produção e de especialização de saberes, necessitando de mais pesquisas e consequentemente mais sistemas de controle do que é produzido, o que resultará em mais gastos para os órgãos financiadores e para o Estado (AGUIAR, 2011).

Outro importante estudo de Price (1976) foi a Lei do Elitismo. Para esse autor, a distribuição da produtividade dos autores numa coordenada cartesiana é tão inclinada que o inspirou a propor a Lei do Elitismo. Segundo esta Lei, a raiz quadrada do total de autores representaria a elite da área estudada, sendo creditada a ela a metade de todas as contribuições. A Lei do Elitismo apresenta aplicações e repercussões imediatamente eficazes para a política científica de um país (URBIZAGÁSTEGUI ALVARADO, 2009a; BRAGA, 1974).

Desta forma, estudar a Lei do Elitismo torna-se parte integrante da quantificação da produtividade científica, que é normalmente mensurada em termos de trabalhos publicados (URBIZAGÁSTEGUI ALVARADO, 2009a), dos quais se originam os cálculos necessários para a medição desta produtividade dos autores.

Contudo, Ravichandra Rao (1986), corroborado por Nicholls (1988), pondera que a alegação de Price é baseada em poucas evidências, e recomenda testes estatísticos para validar a Lei do elitismo. Uma primeira questão central é que o resultado numérico da raiz quadrada do total de autores pode gerar um número não inteiro, ou seja, caso haja um total de 216 autores, então a raiz quadrada é 14,7 autores, surgindo as seguintes situações (NICHOLLS, 1988): o bibliometrista deverá arredondar o valor encontrado de 14,7 autores para mais ou para menos, e o valor teórico obtido (14,7 autores) pode não ser facilmente relacionado aos dados reais.

Considerando-se a distribuição assimétrica da Lei de Lotka (em forma de "J" invertido), que serve de base ao raciocínio de Price, há a ocorrência de valores extremos, que são *outliers* ou valores discrepantes (DE BELLIS, 2009), pois apresentam comportamentos que estão bem distantes (fora do padrão geral) dos demais dados (TRIOLA, 2008).

Portanto, em um conjunto de dados, pode ocorrer a presença de *outliers*, que são valores que apresentam comportamento distinto dos outros dados e que também se denominam de extremos, observações estranhas, discordantes, discrepantes, exteriores (HOAGLIN; MOSTELLER; TUKEY, 1992).

Diante dessas dificuldades de adequar o valor teórico obtido pelo método de Price aos dados reais, e assumindo como hipótese que *outliers* podem ser um indicador de autores profícuos, cabe então a seguinte questão: como os *outliers* podem contribuir na Lei do Elitismo?

2 A Lei do Elitismo nos estudos bibliométricos: revisão de literatura

Atualmente os estudos bibliométricos podem ser considerados indicadores da produção científica e são necessários para se compreender a dinâmica e a evolução da ciência. Nesse sentido, os indicadores bibliométricos vêm sendo utilizados como instrumentos para análise da atividade científica e das suas relações com o desenvolvimento econômico e social, bem como para se obter as características de um determinado campo do conhecimento.

O emprego de indicadores bibliométricos apresenta uma série de vantagens para a avaliação científica por se tratarem de dados verificáveis, reproduzíveis, que apresentam resultados objetivos, numéricos, além de poderem ser aplicados a um grande volume de dados (IGAMI, 2011).

A bibliometria possui três importantes leis que contribuíram para ao avanço desse saber: Lei de Lotka (mede a produtividade dos pesquisadores); Lei de Bradford (Lei de dispersão do conhecimento) e Lei de Zipf (modelo de distribuição e frequência de palavras de um texto).

A primeira lei, a de Lotka, foi formulada em 1926 a partir de um estudo sobre a produtividade de cientistas, calculada com base na contagem de autores presentes no *Chemical Abstracts*, onde, posteriormente a esses estudos, se estabeleceram os fundamentos da lei do quadrado inverso, que afirma que o número de autores que fazem **n** contribuições em um determinado campo científico é aproximadamente $1/\mathbf{n}^2$ daqueles que fazem uma só contribuição, e que a proporção daqueles que fazem uma única contribuição é de mais ou menos 60%

(URBIZAGÁSTEGUI ALVARADO, 2002).

A partir de seus estudos, Price (1976) aperfeiçoou a Lei de Lotka e formulou a Lei do Elitismo, que conclui que 1/3 da literatura é produzida por menos de 1/10 dos autores mais produtivos, levando a uma média de 3,5 documentos por autor, e mantém o número de que 60% dos autores produz um único documento. Mas, em essência, a Lei do Elitismo considera que se "n" representa o número total de contribuintes numa disciplina, então "√n" representaria a elite da área estudada.

Desde quando foi formulada, a Lei do Elitismo vem sendo utilizada como objeto de estudos em várias ciências.

Braga (1973) procurou na área de Ciência da Informação relacionar os documentos citados em revisões da literatura à frente de pesquisa dos periódicos, no período de 1966 a 1970, sendo que o critério utilizado pela autora para a determinação da frente de pesquisa foi o de que esses autores deveriam corresponder a em torno de 10% do total de autores. Assim, para a frente de pesquisa, Braga encontrou uma elite correspondente a 10% do total de autores, e uma elite da revisão da literatura correspondente a 8,7% do total de autores.

Para a área de Botânica, Queiroz (1975) analisou o período de 1971 a 1972 e encontrou 599 autores, a elite calculada seria, então, de 24 autores. Todavia, seus dados apresentavam 20 ou 37 autores como a elite. Ressalta-se que Queiroz (1975), em seus estudos, demonstrou dificuldades de adequar o valor teórico da elite ao critério da raiz quadrada, notadamente quanto à produtividade dos autores da elite, pois o valor teórico da produtividade seria de 50% de todos os trabalhos, e mesmo que a elite fosse de 37 autores, a produtividade destes seria somente 22% do total de contribuições. Assim, o autor asseverou que a elite na área de Botânica apresentava baixa produtividade.

Na área de Química, para o período de análise de 1964 a 1973, Carvalho (1975) determinou que a frente de pesquisa fosse constituída pelos pesquisadores que contribuíram com 3 ou mais trabalhos, que correspondiam a 11% do total de autores, e para o grupo de elite adotou a Lei do Elitismo. O valor teórico obtido foi de 57,5 autores, enquanto os dados reais conduziram a uma elite de 55 autores.

Para o período de 1968 a 1973, uma amostra da produção de artigos dos docentes do Instituto de Ciências Biológicas da Universidade Federal de Minas

Gerais foi pesquisada por Carvalho (1976), que quantificou que a frente de pesquisa e o grupo de elite seriam compostos de 36 de um total de 1516 autores.

Na área de Microbiologia, Imunologia e Parasitologia, Sá (1976) estudou durante o ano de 1971 o total de 229 títulos de periódicos e encontrou que a elite seria dada pela raiz quadrada do total de autores, mas também verificou que 10% do total de autores correspondiam a em torno de 33% do total de trabalhos publicados.

A aplicação da Lei do Elitismo para a área de siderurgia brasileira (GUSMÃO, 1978), no período de 1960 a 1972, produziu uma elite de 27 autores, de um total de 726 autores. Mas vale observar que Gusmão (1978) não informa qual critério utilizou para excluir da elite o 28º e o 29º autores, que produziram a mesma quantidade de trabalhos que o 27º autor.

Já Bomeny (1978) aplicou a Lei de Price ao arquivo privado do ex-presidente Getúlio Vargas, gerado no período de 1930 a 1939, encontrando um total de 356 autores, e a elite como 19 autores, sendo que, neste caso específico, houve a adequação do valor teórico de Price aos dados reais.

Novamente na área de Ciência da Informação, Rodrigues (1982) estudou as citações de 62 dissertações de mestrado defendidas no Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), cobrindo o período de 1972 a 1979, tendo usado como critério para o grupo de elite a raiz quadrada de Price, que obteve, para periódicos e livros/folhetos, um valor real menor do que o valor teórico encontrado, e para as reuniões o valor real determinado foi ligeiramente superior ao valor teórico. Menciona-se que para a determinação da frente de pesquisa foi adotado o valor teórico de 10% do total de autores; sendo que o grupo de elite estaria contido no grupo da frente de pesquisa.

A produtividade de autores sobre plantas medicinais do Peru foi objeto de estudo de Urbizagástegui Alvarado e Lane-Urbizagástegui (2007), que coletaram dados de 1238 autores, cobrindo o intervalo de tempo de 1913 a 2005. A elite teórica dada pela Lei de Price foi de 35,2 autores, enquanto os dados reais forneceram 36 autores que produziram seis ou mais artigos sobre o tema.

A Lei do Elitismo foi utilizada por Urbizagástegui Alvarado (2009b) na área de produtividade de autores (ou Lei de Lotka) cobrindo os anos de 1922 a 2003, onde, de um universo de 203 autores, a raiz quadrada fornece 14,25 autores. A elite

foi então considerada como composta de 15 pesquisadores.

No estudo de quatro revistas sobre Ciência da Informação na Colômbia, publicadas no intervalo de 1978 a 2009, Restrepo Arango e Urbizagástegui Alvarado (2010) encontraram 555 autores, e afirmaram que os autores mais produtivos seriam 11, os quais contribuíram com sete ou mais artigos; todavia, os autores (2010) não informam qual critério utilizado para identificar estes autores mais prolíficos.

Visando explorar as métricas científicas de produtividade em 16 cursos de doutorado na área de Botânica no Brasil, Amarante (2011) identificou 330 docentes que contribuíram entre 1960 até 2010 com 14.757 artigos. A raiz quadrada de 330 autores é 18,2 autores, o que permitiu a Amarante (2011) identificar como a sua elite 18 professores.

3 A Lei do Elitismo e as lacunas em sua aplicação

Após a proposta de Price nos idos de 1960, seguiram diversos trabalhos visando ratificar o critério da raiz quadrada como quantificador da elite, bem como confirmar as relações entre a Lei de Price e a Lei de Lotka.

Procurando provar as conexões da Lei de Lotka com a Lei de Price, Allison et al. (1976), chegaram à Equação 1:

Equação 1: $R\ (\%) = 81{,}2/(\sqrt{nmáx})$

"R (%)" representa a porcentagem da elite em relação ao total de autores; "nmáx" representa a maior contribuição individual de trabalhos.

Se "nmáx" for 100 trabalhos, então "R (%)" é igual a 8,1%, ou seja, a elite representa 8,1% do total de autores ("n"). Caso "nmáx" seja 500 trabalhos, "R (%)" resulta em 3,6%, assim, a elite é constituída por 3,6% do total de autores ("n").

É imperioso reforçar que "R (%)" não depende do total de autores ("n") que publicam no campo do conhecimento, mas fornece a porcentagem da elite em relação ao total de autores.

Uma observação final e relevante é que Allison et al. (1976) apontaram que uma hipótese plausível poderia ser obtida por um estudo estatístico detalhado dos poucos autores profícuos.

Já Glänzel e Schubert (1985), utilizando a denominada teoria dos valores

extremos (estatística de extremos), propõem uma distribuição de frequência a qual denominam distribuição de Price. Mas os autores apontam e reforçam que "√n" ("n" total de autores) nem sempre fornece um número inteiro, havendo, pois, necessidade de arredondar o valor obtido, seja para mais ou para menos.

Todavia, a Lei do Elitismo apresenta alguns problemas para a sua aplicação, segundo Nicholls (1988): independentemente do valor teórico advindo de "√k" ("k" total de autores), há o problema de se adequar esse número aos dados reais. Ou seja, qualquer que seja o valor teórico de "√k", mesmo que arredondado (seja para mais ou para menos), ele obrigatoriamente se encontra dentro de umas das faixas da distribuição de frequência, portanto, gerando o problema de escolher quais de fato são os autores mais profícuos.

Nicholls (1988) utiliza em sua pesquisa os dados originados dos estudos de Dresden sobre a produtividade de matemáticos americanos, reunindo um total de 278 autores. A raiz quadrada deste valor, de acordo com a Lei do Elitismo, fornece 16,7 autores como a elite. Assim, o primeiro problema que emerge é o arredondamento para menos (16 autores) ou para mais (17 autores). Mas, ao consultar os dados reais, surge o problema de adequação do valor teórico: há 15 autores (portanto abaixo de 16 autores) que produziram mais de 13 trabalhos por autor; por outro lado, há 20 autores (logo acima de 17 autores) que produziram mais de 12 trabalhos por autor.

Logo, onde está a elite? Nos primeiros 15 autores? Nos primeiros 20 autores? E qual critério complementar deve ser usado para caracterizar o 17º autor profícuo, visto que este produziu o mesmo que os autores classificados entre o 16º e o 20º lugar em número de publicações?

O mesmo problema foi encontrado nos estudos de Urbizagástegui Alvarado (2009a), pois há um total de 376 autores que escreveram sobre a produtividade de autores; logo, a elite, segundo a raiz quadrada do total de autores, é 19,4. A elite é 19 ou 20 autores? Ao confrontar com seus dados reais, refletindo as dificuldades do uso do critério da raiz quadrada, "[...] a cifra mais próxima a essa quantidade é de 17 autores", o que de fato reflete as dificuldades quando se utiliza o cálculo de Price (URBIZAGÁSTEGUI ALVARADO, 2009a, p. 74).

Esta questão da arbitrariedade de impor um limite para o tamanho da elite é

também reforçada por Vinkler (2010), que aponta que esse limite deve depender do objetivo da avaliação e do tamanho e características do conjunto de dados analisados. Outro detalhe apontado por Vinkler (2010) é que a Lei de Price pode superestimar o tamanho da elite para grandes conjuntos de dados.

É interessante também citar a proposta de Matsas (2009), que utiliza o fator de impacto normalizado (NIF). O NIF leva em conta três fatores (para um único trabalho): o número de autores que publicaram o trabalho; o total de referências utilizadas no trabalho; o total de vezes em que este trabalho é então citado pelos pares. Para se calcular então o NIF de um pesquisador, deve-se efetuar toda a somatória de sua produção com base nesses três fatores.

Entretanto, o método apresenta limitações. Dentre elas, Matsas (2009) aponta a necessidade de se aplicar o método somente para pesquisadores seniores; além disso, Marques (2009) assevera que o método não se aplica às ciências humanas, em virtude da maioria das publicações neste campo do saber dar-se em livros e capítulos de livros.

Por sua vez, Christovão (1979) propôs o critério denominado "filtro de qualidade", no qual se parte da literatura informal (cartas, conversas, comunicações e congressos), e o que é produzido começa a ser "filtrado" até transformar-se na literatura superformal (revisões da literatura, serviços de indexação e resumos; e bibliografia de bibliografias). A hipótese de Christovão (1979) é que a frente de pesquisa encontra-se em uma parte do conjunto "filtrado".

Em seu estudo, Christovão (1979) trabalhou na área de Ciência da Informação, no período de 1969 a 1977, utilizando como fonte de dados as publicações "Library and Information Science Abstracts" (LISA) e o "Annual Review of Information Science and Technology" (ARIST). Ao filtrar as comunicações do LISA até o ARIST, o autor conclui que, de 277 autores em comum nas duas fontes de dados, somente 133 constituiriam a frente de pesquisa, e finaliza dizendo que seriam necessários alguns testes com o método proposto, no tocante a interação dos autores com coautores.

4 O uso de *outliers* nos cálculos bibliométricos da elite

A comunicação pode ser entendida como um mecanismo que permite a troca de ideias, podendo variar de interesse de acordo com as características do grupo a qual pertence. E é esse processo de troca de ideias e informação entre os grupos de interesse de um mesmo conhecimento que é o manancial de onde surgem os conhecimentos científicos, que depois de registrados se transformarão em informações científicas (LE COADIC, 1996).

Tal mecanismo de circulação do conhecimento registrado entre os membros de uma mesma comunidade científica é denominado comunicação científica, que, de acordo com Garvey (1979), são as atividades associadas à produção, à disseminação e ao uso da informação, do momento em que o cientista concebe uma ideia para pesquisar até o processo final, ou seja, o momento em que o resultado final é aceito como constituinte do estoque final do conhecimento científico.

Assim, os estudos relacionados à produção científica dos autores de um determinado ramo do conhecimento geram uma distribuição assimétrica cujo gráfico apresenta-se em forma de "J" invertido, em virtude da elevada variabilidade dos dados, os quais comprometem a utilização de estatísticas elementares tais como a média aritmética e o desvio padrão (DE BELLIS, 2009) como indicadores bibliométricos.

Portanto, a questão da coexistência de muitos autores com pouca produção e poucos autores com muita produção conduz ao que De Bellis (2009) chama de natureza bipolar dos dados. Neste caso, a denominada cauda da distribuição (a que contém os autores prolíficos) deve receber um tratamento diferenciado, podendo ser encarada como ocorrência de eventos raros ou não usuais, os quais De Bellis (2009) esclarece se tratarem de valores extremos ou *outliers.*

Diante dos problemas da aplicação da Lei de Price, surge como alternativa para os cálculos bibliométricos o uso dos *outliers*, que são valores discrepantes ou valores muito afastados das demais observações coletadas, que, uma vez identificados, podem revelar informações importantes (TRIOLA, 2008; BORNMANN et al., 2008; BUSSAB; MORETTIN, 2013; LIMA; MAROLDI; SILVA, 2013).

No trabalho de Lima, Maroldi e Silva (2013), os autores procuraram expor que os *outliers* podem influir ou representar detalhes relevantes nos cálculos bibliométricos, principalmente no momento discussão dos dados, uma vez que atualmente os **"**[...] estudos bibliométricos são mais complexos do que apenas um levantamento estatístico puro" (FERREIRA, 2010, p. 2), necessitando de cálculos mais apurados.

Este ponto de vista da influência de *outliers* na bibliometria é reforçado por Glänzel e Moed (2013), Glänzel (2013) e Prathap (2014), que alertam que, em geral, os *outliers* são valores que merecem atenção, pois podem servir como indicadores suplementares nos cálculos bibliométricos; no nosso trabalho, utilizaremos os *outliers* como indicadores da detecção da elite de produção científica.

Dentre as diversas possibilidades de aparecimento de um *outlier*, a primeira pode ser o registro errado de uma observação (neste caso, exclui-se automaticamente o *outlier* dos cálculos); a segunda pode indicar a coleta de um dado pertencente à outra população diferente da população de interesse; terceira, o *outlier* é um valor correto, mas de probabilidade de ocorrência muita baixa (McCLAVE; BENSON; SINCICH, 2009), como no caso da determinação de uma elite, ou seja, poucos autores (baixa probabilidade) produzem a maior parte da produção científica. Já Bornmann et al. (2008) utilizam os *outliers* para identificação de periódicos altamente citados, bem como os muito poucos citados.

Para a detecção e cálculo dos *outliers*, utiliza-se a Análise Exploratória de Dados (AED), técnica introduzida em 1977 por John Tukey. A AED objetiva extrair dos dados coletados o máximo possível de informações, incluindo os *outliers* (TRIOLA, 2008; BUSSAB; MORETTIN, 2013; LIMA; MAROLDI; SILVA, 2013).

Como estamos interessados nos valores de *outliers* que caracterizam uma elite, e tendo visualizado que valores extremos de produção científica podem estar associados aos *Outliers* Superiores Extremos (O.S.E.), que representam elevada produção científica, sugerimos aos bibliometristas, durante os cálculos da Lei do Elitismo, a Equação 1, com uso do primeiro quartil (sigla Q1; 25% das observações abaixo dele) e terceiro quartil (sigla Q3; 75% das observações abaixo dele). Em termos matemáticos (TRIOLA, 2008):

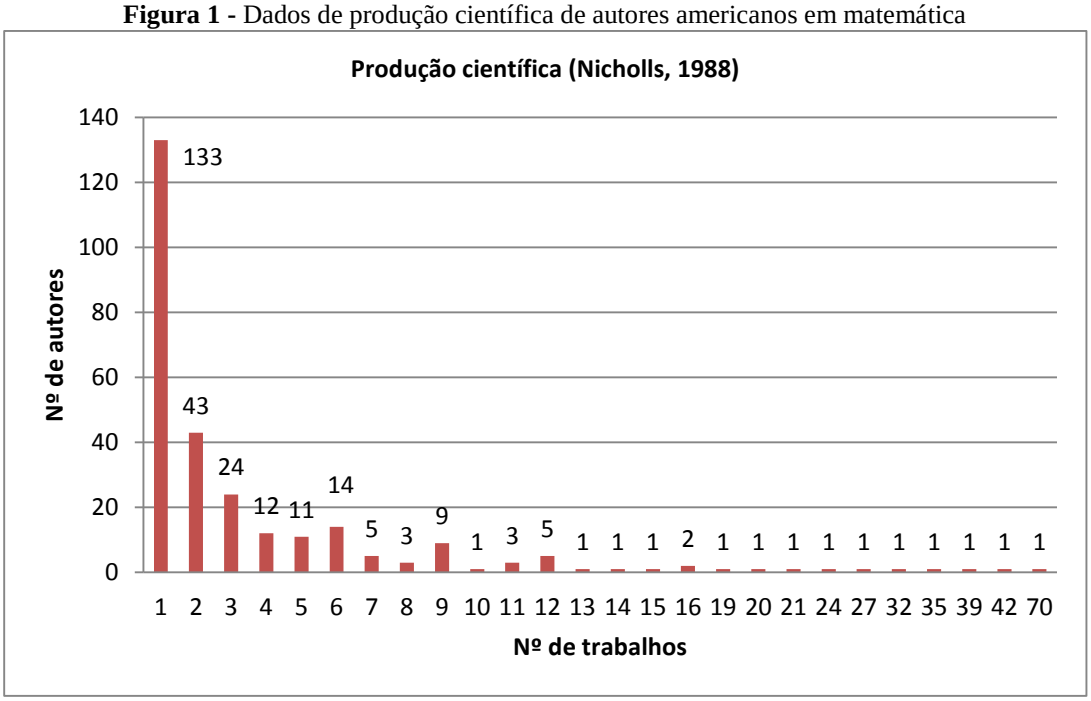
Equação 2: O.S.E. > Q3 + 3,0*(Q3 – Q1)

O uso de *outliers* muda o foco de cálculo da elite, pois os quartis (Q1 e Q3) são associados aos trabalhos produzidos, e assim o O.S.E. também se refere aos trabalhos produzidos. O valor encontrado para o O.S.E. pode até ser um número não inteiro, mas como indica a equação 2, utiliza-se o primeiro número inteiro maior que o resultado encontrado.

Identificado o O.S.E., então, o número de autores que compõem a elite é obrigatoriamente um número inteiro, sem necessidade de se adaptar o resultado teórico aos dados reais. E aqui se encontra a principal diferença dos métodos, a Lei do Elitismo utiliza o total de autores, gerando um valor teórico de autores em geral não inteiro, e provoca a necessidade de se adaptar o valor teórico aos dados observados. Já os *outliers* utilizam no cálculo os trabalhos produzidos, que podem até gerar um número fracionário, mas, quando arredondado para mais, gera um número inteiro de autores que se encaixa nos dados coletados.

Nicholls (1988) apresenta dados coletados por Dresden de autores americanos da área da matemática, ilustrados na Figura 1.

Figura 1 - Dados de produção científica de autores americanos em matemática



Fonte: Elaborada pelos autores com base em Nicholls (1988).

Portanto, a Figura 1 ilustra que 133 autores produziram somente um trabalho;

43 autores contribuíram com dois trabalhos; já 24 autores escreveram três trabalhos, e assim sucessivamente, sendo que o valor máximo encontrado corresponde a um único autor o qual produziu 70 trabalhos. O total de autores na área são 278.

Com as ideias dos *outliers* e utilizando os achados de Nicholls (1988), e ainda seguindo a metodologia de cálculo dos quartis proposto por Triola (2008), a localização do primeiro quartil (Q1) é dada por (278/4) ou 69,5; portanto, o primeiro quartil (Q1) corresponde aos trabalhos produzidos pelo 70º autor, número esse que é igual a um (Q1 = 1 trabalho).

A localização do terceiro quartil (Q3) é dada por (3*278/4) ou 208,5. Assim, o terceiro quartil (Q3) corresponde aos trabalhos publicados pelo 209º autor, o que corresponde a quatro trabalhos (Q3 = 4 trabalhos).

Dessa forma, utilizando a equação 2, verifica-se que o O.S.E. (e, portanto, a elite) é quem produz mais de 13 trabalhos, de outra maneira, a elite é quem produz exatamente 14 ou mais trabalhos. E quem produz 14 ou mais trabalhos? São 14 autores.

Percebe-se que o cálculo da elite via *outliers* permite a caracterização única do número de autores mais profícuos, sem necessidade de arredondamentos (em relação ao total de autores) nem utilização de critérios complementares para adequar o valor teórico obtido aos dados reais. Ademais, o cálculo envolvendo a determinação dos *outliers* leva em conta a forma da distribuição dos dados.

5 Considerações finais

Observamos que há um grande interesse da comunidade científica das mais diversas áreas do conhecimento nos estudos relacionados à Lei do Elitismo de Price, porém, muitas vezes, durante os cálculos, quando o bibliometrista se depara com números fracionados, o mesmo deve fazer a opção de arredondar para mais ou para menos seus dados, podendo, com frequência, enviesar os achados de sua pesquisa.

Com o uso da equação para o cálculo de *outliers* extremos que advém da Análise Exploratória de Dados, mostramos a aplicabilidade do método para a determinação da elite de qualquer campo do conhecimento, fornecendo outro critério além da raiz quadrada de Price.

Para o cálculo da produtividade de autores americanos em matemática, há um total de 278 autores, assim, a Lei de Price fornece 16,7 autores como a elite (NICHOLLS, 1988). Vimos na Lei do Elitismo que, em alguns cálculos, para detectar a elite de uma determinada área do conhecimento, os números deverão ser arredondados para cima ou para baixo e, nesses casos, o bibliometrista deverá adaptar os números obtidos aos dados coletados, ou seja, a elite são 15 (equivalentes a 5,4% do total de autores) ou 20 autores (correspondentes a 7,2% de todos os autores)? Já o critério dos *outliers* gerou uma elite de exatamente (detalhe) 14 autores (representam 5,0% do total de autores).

Portanto, o critério de Price, além de fornecer geralmente um valor não inteiro para a elite, ainda possui o problema de adaptação aos dados reais. Nesse caso, é necessário o uso de critérios complementares para determinar a elite, todavia, nem sempre esses critérios complementares são completamente enunciados e/ou elucidados nos trabalhos, não havendo, em alguns casos, nem mesmo menção sobre qual critério complementar foi adotado.

Já o método dos *outliers* pode até gerar valores fracionados para o valor da quantidade de produção que define a elite (devendo ser arredondado para mais), todavia, fornece um valor único para a identificação dos autores que compõem a elite.

Referências

AGUIAR, R. R. G. B. de. **Um olhar sobre a história**: características e tendências da produção científica na área de história do Brasil (1985 – 2009). 2011. 156 f. Dissertação (Mestrado)-Instituto Brasileiro de Ciência e Tecnologia, Rio de Janeiro, 2011.

ALLISON, P. D. et al. Lotka's law: a problem in its interpretation an application. **Social Studies of Science**, London, v. 6, n. 2, p. 269-276, 1976. Disponível em: <http://sss.sagepub.com/content/6/2/269.citation>. Acesso em: 31 jan. 2014.

AMARANTE, C. Professores/pesquisadores da Pós-Graduação em Botânica no Brasil: análise métricas de produtividade. **Tendências da Pesquisa Brasileira em Ciência da Informação**, João Pessoa, v. 4, n. 1, 2011. Disponível em: <http://inseer.ibict.br/ancib/index.php/tpbci/article/view/50/87>. Acesso em: 03 jul. 2014.

BOMENY, R. H. D. Estudo bibliométrico aplicado ao arquivo privado de Getúlio Vargas. **Ciência da Informação**, Rio de Janeiro, v. 7, n. 1, p. 37-42, 1978.

BORNMANN, L. et al. Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. **Ethics in Science and Environmental Politics**, Geneva, v. 8, n. 1, p. 93-102, 2008. Disponível em: <http://www.int-res.com/articles/esep2008/8/e008p093.pdf>. Acesso em: 07 fev. 2014.

BRAGA, G. M. Relações bibliométricas entre a frente de pesquisa (research front) e revisões da literatura: estudo aplicado à Ciência da Informação. **Ciência da Informação**, Rio de Janeiro, v. 2, n. 1, p. 9-26, 1973.

BRAGA, G. M. Informação, ciência, política científica: o pensamento de Derek de Solla Price. **Ciência da Informação**, Rio de Janeiro, v. 3, n. 2, p. 155-177, 1974.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística básica**. 8. ed. São Paulo: Saraiva, 2013.

CARVALHO, K. Revista científica e pesquisa: perspectiva histórica. In: PROBACIÓN, D. A. et al. (Org.). **Revistas científicas**: dos processos tradicionais as perspectivas alternativas de comunicação. Cotia, SP: Ateliê Editorial, 2011. p. 23-42.

CARVALHO, M. L. B. Estudos de citações da literatura produzida pelos professores do Instituto de Ciências Biológicas da UFMG. **Ciência da Informação**, Rio de Janeiro, v. 5, n. 1/2, p. 27-42, 1976.

CARVALHO, M. M. Análises bibliométricas da literatura de Química no Brasil. **Ciência da Informação**, Rio de Janeiro, v. 4, n. 2, p. 119-141, 1975.

CHRISTOVÃO, H. T. Da comunicação informal a comunicação formal: identificação da frente de pesquisa através de filtros de qualidade. **Ciência da Informação**, Rio de Janeiro, v. 8, n. 1, p. 3-36, 1979.

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

FERREIRA, A. G. C. Bibliometria na avaliação de periódicos científicos. **DataGramaZero**, Rio de Janeiro, v. 11, n. 3, jun. 2010.

GARVEY, W. D. **Communication**: the essence of science; facilitating information among librarians, scientists, engineers and students. Oxford: Pergamon, 1979.

GLÄNZEL, W.; High-end performance or outlier? Evaluating the tail of scientometric distributions. **Scientometrics**, Amsterdam, v. 97, n. 1, p. 13-23, 2013.

GLÄNZEL, W.; MOED, H. F. Thougts and facts on bibliometric indicators. **Scientometrics**, Amsterdam, v. 96, n. 1, p. 381-394, 2013.

GLÄNZEL, W.; SCHUBERT, A. Price distribution: an exact formulation of Price's "square root law". **Scientometrics**, Amsterdam, v. 7, n. 3-6, p. 211-219, 1985.

GORBEA PORTAL, S. **Modelo teórico para el estúdio métrico de la información documental**. Gijón: Ediciones Trea, 2005.

GUSMÃO, H. R. Análise da literatura brasileira de Siderurgia. **Ciência da Informação**, Rio de Janeiro, v. 7, n. 1, p. 25-35, 1978.

HOAGLIN, D. C.; MOSTELLER, F.; TUKEY, J. W. **Análise exploratória de dados**: técnicas robustas: um guia. Lisboa: Salamandra, 1992.

IGAMI, M. P. Z. **Elaboração de indicadores de produção científica com base na análise cientométrica das dissertações e teses do IPEN**. 2011. 179 f. Tese (Doutorado) - Instituto de Pesquisas Energéticas e Nucleares, São Paulo, 2011.

LE COADIC, Y. F. **A ciência da informação**. Brasília: Briquet de Lemos, 1996.

LIMA, L. F. M.; MAROLDI, A. M.; SILVA, D. V. O. da. Outlier(s) em cálculos bibliométricos: primeiras aproximações. **Liinc em Revista**, Rio de Janeiro, v. 9, n. 1, 2013. Disponível em: <http://revista.ibict.br/liinc/index.php/liinc/article/view/522/399>. Acesso em: 07 fev. 2014.

MARQUES, F. Leve-me ao seu líder: método criado por professor da Unesp movimenta o debate sobre avaliação da produção acadêmica. **Pesquisa Fapesp**, São Paulo, v. 156, n. 2, p. 32-34. 2009. Disponível em: <http://www.revistapesquisa.fapesp.br/wp-content/uploads/2009/02/32-34_156.pdf>. Acesso em: 31 jan. 2014.

MATSAS, G. E. A. What are scientific leaders? The introduction of a normalized impact factor. Brazilian Journal of Physics, São Paulo, v. 42, n. 5-6, p. 319-322, 2012. Disponível em: <http://www.arxiv.org/pdf/0809.0290v2.pdf>. Acesso em: 31 jan. 2014.

McCLAVE, J. T.; BENSON, P. G.; SINCICH, T. **Estatística para economia e administração**. 10. ed. São Paulo: Pearson, 2009.

MOSTAFA, S. P. As ciências da informação. **São Paulo em Perspectiva**, São Paulo, v. 8, n. 4, p. 22-27, 2004.

MUGNAINI, R.; CARVALHO, T. de; CAMPANATTI-OSTIZ, H. Indicadores de produção científica: uma discussão conceitual. In: POBLACIÓN, D. A.; WITTER, G. P.; SILVA, J. F. M. da. **Comunicação e produção científica**: contexto, indicadores e avaliação. São Paulo: Angellara, 2006. p. 315-340.

NICHOLLS, P. T. Price's square root law: empirical validity and relation to Lotka's law. **Information Processing & Management**, Elmsford, v. 24, n. 4, p. 469-477, 1988.

PRATHAP, G. Single parameter indices and bibliometric outliers. **Scientometrics**, Amsterdam, v. 101, n. 3, p. 1781-1787, dez. 2014.

PRICE, D. J. S. **O desenvolvimento da ciência**. Rio de Janeiro: LTC, 1976.

QUEIROZ, S. S. Bibliografia brasileira de Botânica, 1971-1972. **Ciência da Informação**, Rio de Janeiro, v. 4, n. 1, p. 55-66, 1975.

RAVICHANDRA RAO, I. K. **Métodos quantitativos em biblioteconomia e ciência da informação**. Brasília: Associação dos Bibliotecários do Distrito Federal, 1986.

RESTREPO ARANGO, C.; URBIZAGÁSTEGUI ALVARADO, R. La productividad de los autores em la ciência de la información colombiana. **Ciência da Informação**, Brasília, v. 39, n. 3, p. 9-22, 2010.

RODRIGUES, M. P. L. Citações nas dissertações de mestrado em Ciência da Informação. **Ciência da Informação**, Brasília, v. 11, n. 1, p. 35-61, 1982.

SÁ, E. S. Participação dos pesquisadores de Microbiologia, Imunologia e Parasitologia (MIP) na literatura científica internacional. **Ciência da Informação**, Rio de Janeiro, v. 5, n. 1/2, p. 43-69, 1976.

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

URBIZAGÁSTEGUI ALVARADO, R. Elitismo na literatura sobre a produtividade dos autores. **Ciência da Informação**, Brasília, v. 38, n. 2, p. 69-79, maio/ago. 2009a.

URBIZAGÁSTEGUI ALVARADO, R. A frente de pesquisa na literatura sobre a produtividade dos autores. **Encontros Bibli**, Florianópolis, v. 14, n. 28, p. 38-56, 2009b.

URBIZAGÁSTEGUI ALVARADO, R. A Lei de Lotka na bibliometria brasileira. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 14-20, maio/ago. 2002.

URBIZAGÁSTEGUI ALVARADO, R.; LANE-URBIZAGÁSTEGUI, S. Productividad de los autores de literatura sobre plantas medicinales del Perú. **Revista ACB**, Florianópolis, v. 12, n. 2, p. 235-253, 2007.

VANTI, N. A cientometria revisitada à luz da expansão da ciência, da tecnologia e da inovação. **PontodeAcesso**, Bahia, v. 5, n. 3, p. 5-31, 2012.Disponível em: <http://www.portalseer.ufba.br/index.php/revistaici/article/view/5679/4099>.

Acesso em: 07 fev. 2014.

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

Outliers in Price's Law

Abstract: Looking at the most productive subset of authors as outliers, this study proposes the use of Exploratory Data Analysis (EDA) to detect outliers, and thus the productive elite. Price's Law or the square root criteria proposed by Price has two major problems: first, the square root of the total number of authors is not always an integer value, creating the need to employ truncate or rounded values. Second, the cutoff point is not easy to select, since the theoretical value rarely fits the observed data. Besides, the detection of outliers from EDA finds a unique way to show the elite. In the end of the study we conclude that outliers may give non integer values for author production, but offers only one value to identify the elite set in a given area of knowledge.

Keywords: Outlier. Exploratory Data Analysis. Price's Law. Bibliometrics. Statistics.

---

[1] SCHWARTZMAN, S. Pesquisa e política. **Literatura Econômica**, Rio de Janeiro, v. 1, n.2, p. 121-124, 1979.

Recebido: 23/07/2014
Aceito: 24/11/2014