Avaliação de bolsas de produtividade em pesquisa do CNPq e medidas bibliométricas: correlações para todas as grandes áreas

Jacques Wainer

Professor titular do Instituto de Computação, Universidade Estadual de Campinas (UNICAMP)

Paula Vieira

Aluna de graduação em Ciência da Computação, Universidade Estadual de Campinas (UNICAMP)

Este trabalho estuda as correlações entre decisões tomadas no fim de 2009, sobre renovação ou não de bolsas de produtividade em pesquisa do CNPq e medidas bibliométricas. Para cada nível da bolsa e para cada subárea, calculamos a correlação da decisão em subir o pesquisador de nível, mantê-lo no nível original ou rebaixá-lo, com várias medidas bibliométricas, como produção total (artigos, conferências, livros e capítulos de livros), produção nos últimos 5 anos, produção indexada no Web of Science, citações recebidas por artigo, citações recebidas por artigo escrito nos últimos 5 anos, índice H, etc. Os dados de citações foram extraídos tanto do Google Scholar como do Web of Science. As correlações de cada subárea são agrupadas em cada uma das 8 grandes áreas do CNPq (Ciências Agrícolas, Ciências Biológicas, Ciências Exatas, Ciências Humanas, Ciências da Saúde, Ciências Sociais, Engenharia e Artes). Indicamos quais são as métricas bibliométricas com maior correlação, com as decisões do CNPq para cada nível e para cada uma das grandes áreas. Discutimos algumas grandes áreas nas quais parece haver uma maior coerência, através dos vários níveis da bolsa entre as métricas mais correlacionadas com as decisões.

Palavras chave: *Bibliometria; Avaliação por pares; CNPq; Bolsa de produtividade.*

Evaluation of the CNPq research scholarships and bibliometric measures: correlations for all large areas

This work analyses the correlation between decisions regarding the CNPq research scholarships (taken at the end of 2009) and some bibliometric measures. For all levels, and for all scientific subareas we calculated the correlation of the decision to raise, keep or lower the researcher's evaluation level and many bibliometric measures such as: total production (journal and conference papers, book an book chapters), production in the last 5 years, total production and last 5 years production that is indexed in the Web of Science, citations per article received, citations per article published in the last 5 years, H index and so on. We used both Google Scholar and Web of Science to obtain the citation data. The correlations are aggregated for all subareas of each of the CNPq large subject areas (Agriculture Sciences, Biology Sciences, Exact Sciences, Human Sciences, Health Sciences, Social Sciences, Engineering and Arts). We show which bibliometric measure has the higher correlation to the scholarship decisions, for each level and for each large area. We discuss some of the cases where there seems to be a coherent theme regarding the most highly correlated measure across all levels of the scholarship.

Keywords: *Bibliometrics; Peer evaluation; CNPq; Scholarship.*

Recebido em 10.09.2012 Aceito em 30.04.2013

1 Introdução

A avaliação de ciência e, em particular, de cientistas, leva em conta múltiplas dimensões. Dentre estas dimensões, estão produção e impacto. Medidas de produção de um cientista incluem o número total de artigos publicados em toda a carreira do pesquisador[1] ou em um período fixo de tempo e, adicionalmente, podem ser ponderadas por fatores que indicam

1 Usaremos o termo "pesquisador" no masculino, associado ao pronome "ele", quando for o caso, mas, obviamente, o leitor não dever assumir que todos os pesquisadores são do gênero masculino.

"qualidade" do veículo de publicação. Assim, medidas que contam apenas artigos publicados em revistas indexadas pela Thomson Reuters (antiga ISI) ou em revistas Qualis A, estão tentando ponderar a produção com um fator que indica a "qualidade" do veículo (mas não do artigo). Medidas de impacto tentam avaliar o quanto a produção do cientista teve consequências para a sua área de pesquisa (ou, talvez, em outras Ciências ou mesmo a sociedade em geral). A medida mais tradicional de impacto é o número de citações. Há algumas medidas que combinam produção com impacto (ou melhor, citações), tais como citações médias por artigo ou índice h (HIRSCH, 2005).

Avaliação de cientistas é tradicionalmente feita por pares. A avaliação da Bolsa de Produtividade em Pesquisa do CNPq, financiamento de pesquisa, bancas de concursos, prêmios, etc., são atividades pelas quais os cientistas são avaliados pelos seus pares e, na maioria dos casos, a avaliação é competitiva. Acredita-se que, na avaliação por pares, outros aspectos que não apenas a produção e o impacto podem ser levados em consideração. Por exemplo, os pares/avaliadores podem avaliar o impacto das pesquisas, usando não apenas citações. Os avaliadores podem, também, ter uma noção do "potencial" futuro da pesquisa do cientista, o que, obviamente, não é capturado pelas medidas que olham apenas para o passado do pesquisador (os artigos escritos e as citações recebidas). Por outro lado, a avaliação por pares pode levar em consideração aspectos que a maioria consideraria como espúrios, como, por exemplo, "quão afinado é o cientista com as visões da banca de avaliação" ou "quão grato o cientista ficará com uma boa avaliação".

O objetivo desta pesquisa é verificar quão correlacionada é a avaliação por pares para a concessão de bolsas de produtividade do CNPq ,com medidas quantitativas tradicionais de produção, produção recente, impacto total, impacto da produção recente, etc.

1.1 CNPq e a bolsa de produtividade

O CNPq é um dos principais órgãos brasileiros de financiamento da pesquisa científica. Entre as formas de financiamento do CNPq, inclui-se a Bolsa de Produtividade em Pesquisa, que é atribuída a pesquisadores de todas as áreas, baseado não só na qualidade de um projeto submetido, mas principalmente na "qualidade" do pesquisador. Como a Bolsa de Produtividade em Pesquisa inclui um forte componente da avaliação da "qualidade" do pesquisador (em contraste com outras formas de financiamento que incluem, também, a qualidade e adequação da proposta aos objetivos específicos do financiamento), ela se torna uma boa ferramenta para entender como é feita a avaliação de pesquisadores no Brasil.

O CNPq é organizado em comitês assessores (CA) para cada subárea do conhecimento. Os CAs são compostos de pesquisadores reconhecidos daquela subárea, indicados por um conselho deliberativo do CNPq, que seleciona os membros através de consultas às entidades

cientificas, à comunidade cientifica, entre outros. Os CAs se reúnem periodicamente para avaliar as propostas submetidas e indicar os selecionados. Uma das funções de um CA é atribuir os níveis da bolsa de produtividade para os pesquisadores que submeteram propostas.

Os CAs são organizados em diferentes grandes áreas, que será a unidade de análise desta pesquisa. As grandes áreas científicas do CNPq e seus CAs correspondentes são:

- Ciências Agrárias (CAG), que incluem as seguintes subáreas (ou CAs): recursos florestais e engenharia florestal, agronomia, recursos pesqueiros e engenharia de pesca, medicina veterinária, zootecnia, engenharia agrícola, ciência e tecnologia de alimentos;
- Ciências Biológicas (CB): bioquímica, fisiologia, imunologia, microbiologia, zoologia, ecologia, biotecnologia, botânica, biofísica, morfologia, parasitologia, genética e farmacologia;
- Ciências Exatas e da Terra (CE): física, química, geociências, astronomia, ciência da computação, oceanografia, matemática, probabilidade e estatística;
- Ciências Humanas (CH): história, psicologia, filosofia, educação, antropologia, sociologia, geografia, arqueologia;
- Ciências da Saúde (CS): odontologia, medicina, educação física, enfermagem, farmácia, saúde coletiva, fisioterapia e terapia ocupacional, fonoaudiologia e nutrição;
- Ciências Sociais Aplicadas (CSA): economia, arquitetura e urbanismo, direito, administração, planejamento urbano e regional, serviço social, ciência política, comunicação, meio ambiente e agrárias, demografia, ciência da informação, desenho industrial, economia doméstica, turismo, museologia e teologia;
- Engenharias (EN): engenharia elétrica, engenharia civil, engenharia biomédica, engenharia química, engenharia mecânica, engenharia de materiais e metalúrgica, engenharia aeroespacial, engenharia de transportes, engenharia nuclear, engenharia sanitária, engenharia de produção, engenharia de minas, engenharia/tecnologia/gestão, engenharia naval e oceânica; e
- Linguística, Letras e Artes (LLA): letras, linguística e artes.

Na nossa análise, deixamos de fora a grande área "outros" e sua única subárea "multidisciplinar".

A Bolsa de Produtividade em Pesquisa é organizada em níveis, em ordem crescente: 2, 1D, 1C, 1B, 1A, sendo que os últimos quatro níveis são coletivamente chamados de "níveis 1". Cada nível provê uma complementação salarial crescente, mas há um grande salto entre os níveis 2 e 1D. Há, também, outras vantagens em ser um pesquisador de nível 1. Por exemplo, há chamadas para o financiamento de projetos que exigem que o pesquisador responsável seja nível 1. Além disso, apenas pesquisadores nível 1 podem ser membros dos CAs e apenas eles participam das consultas do conselho deliberativo à comunidade científica.

As Bolsas de Produtividade em Pesquisa tem duração de 3 a 4 anos. Ao final do período da bolsa, o pesquisador submete novamente um pedido. Este pedido é avaliado e, se aprovado, o pesquisador recebe um nível de avaliação e a bolsa correspondente. Na grande maioria dos casos (mas não todos), as avaliações finais são uma de quatro alternativas: o pesquisador perde a bolsa e passa a não ter mais uma classificação, o pesquisador perde um nível (ou seja, passa para o nível imediatamente inferior ao que ele tinha), o pesquisador mantém o seu nível anterior ou o pesquisador sobe um nível. Em poucas subáreas, o pesquisador pode subir ou descer mais de um nível em uma avaliação. Finalmente, o número de bolsas de produtividade é fixo, por subárea. Assim, um CA normalmente não tem liberdade de atribuir a um pesquisador uma bolsa do nível 1C, a não ser que outro pesquisador tenha perdido sua bolsa 1C, nesta mesma avaliação. Há, também, regras do CNPq sobre quanto tempo, após a obtenção do doutorado, um pesquisador se torna elegível para receber as bolsas de nível 2 e 1. Portanto, há restrições em atribuir bolsas de níveis mais altos a pesquisadores mais jovens (do ponto de vista de idade científica, não idade cronológica).

Infelizmente, o nível de bolsa de pesquisa do CNPq não pode ser usado diretamente como uma medida da "qualidade" atribuída pelos pares a um pesquisador. A razão principal é que há uma histerese entre os vários níveis. Vamos supor que o pesquisador X tenha uma bolsa 1C e que não tenha produzido muito, desde então. Na sua próxima renovação, ele deve cair para o nível 1D, mas são poucas as subáreas que vão reduzir o pesquisador para o nível 2 ou mesmo deixar de conceder-lhe a bolsa. Assim, este pesquisador terá métricas baixas para os últimos anos, mas, mesmo assim, ele ainda terá uma bolsa 1D. Vamos considerar o pesquisador Y, recém-formado, com altíssima produção e um alto número de citações. Pelas regras do CNPq, este pesquisador só pode pedir uma bolsa de produtividade nível 2 após 3 anos do doutoramento, e Y não poderá receber uma bolsa de nível 1 antes de 8 anos do doutoramento. Assim, o candidato Y tem uma produção e citações em curto prazo muito maiores que as de X, mas, mesmo assim, tem uma classificação abaixo da de X.

O nível da bolsa do CNPq não pode ser usado como uma medida direta da qualidade que o CNPq atribui ao pesquisador, já as variações (isto é, pesquisadores que subiram, foram renovados, ou caíram nas suas

avaliações) podem. Um pesquisador X, que na avaliação de 2009 subiu de 1D para 1C, claramente pode ser considerado melhor que o pesquisador Y que, na mesma avaliação, teve sua bolsa 1D renovada. Ambos podem ser considerados melhores que o pesquisador Z que teve a sua bolsa 1D rebaixada para 2. Mas um pesquisador W que tem uma bolsa 1D e não foi avaliado em 2009, não pode ser comparado com X, Y ou Z. Assim, apenas as variações em um mesmo ano e nível podem ser usadas como medida da qualidade do pesquisador na avaliação do CA corresponde. Nesta pesquisa, usaremos a comparação entre a variação entre níveis de bolsa de produtividade e as medidas bibliométricas destes pesquisadores. Em um mesmo nível (por exemplo, 1C), os pesquisadores que tiveram suas bolsas reduzidas são coletivamente considerados "piores" que aqueles que mantiveram a bolsa e estes são coletivamente "piores" que os que subiram para 1B (ou 1A se for o caso). Esta ordem de qualidade decidida pelo CA correspondente será correlacionada com a ordem, de menor para maior, de várias medidas bibliométricas destes mesmos pesquisadores.

Há várias medidas bibliométricas que podem ser relevantes nas avaliações dos CAs do CNPq. Obviamente, não podemos testar todas e, portanto, escolhas foram feitas. Abaixo, as medidas que usamos, neste artigo, e suas abreviações:

a) IDADE: idade científica do pesquisador, isto é, o número de anos desde a obtenção do doutorado. Dado obtido do Currículo Lattes dos pesquisadores;

b) PRODTOTAL: produção total do pesquisador durante sua carreira, isto é, o número total de artigos (revistas, artigos de conferências, livros e capítulos de livros) listados pelo pesquisador no seu Currículo Lattes;

c) PRODWOS: produção total do pesquisador registrada no Web of Science (WOS), isto é, apenas a produção em revistas indexadas pelo WOS;

d) PROD5: produção do pesquisador nos últimos cinco anos (segundo o Lattes);

e) PROD5WOS: produção nos últimos cinco anos registrada no WOS;

f) CITWOS: total de citações recebidas pelo pesquisador na sua carreira, segundo o WOS;

g) CITSCH: total de citações recebidas pelo pesquisador na sua carreira, segundo o Google Scholar;

h) CIT5WOS: total de citações recebidas pelo pesquisador nos artigos publicados nos últimos cinco anos, segundo o WOS;

i) CIT5SCH: total de citações recebidas pelo pesquisador nos artigos publicados nos últimos cinco anos, segundo o Google Scholar;

j) CITARTSCH: citações por artigo, usando todas as publicações do Lattes e Scholar (PRODTOTAL/CITSCH);

k) CITARTWOS: citações recebidas por artigo indexado no WOS (PRODWOS/CITWOS);

l) CITART5SCH: citações por artigo publicado nos últimos cinco anos (PROD5/CIT5SCH);

m) CITART5WOS: citações recebidas por artigo indexado no WOS nos últimos cinco anos (PROD5WOS/CIT5WOS);

n) HWOS: índice h do pesquisador segundo os dados do WOS;

o)HSCH: índice h do pesquisador segundo os dados do Google Scholar.

Estas medidas podem ser classificadas em algumas famílias. Nós chamaremos a medida de IDADE uma métrica *histórica*. Como as decisões dos CAs são altamente correlacionadas com esta métrica, então estes CAs estão privilegiando os pesquisadores mais antigos da subárea. A segunda família, de *produção,* inclui as medidas PRODTOTAL e PRODWOS e avaliam de diferentes formas a produção total da carreira do pesquisador. A terceira família inclui as medidas CITWOS e CITSCH e são medidas essencialmente de *impacto acumulado* da carreira do pesquisador. As medidas PROD5 e PROD5WOS são medidas de *produtividade* ou *produção recente*. As medidas de CIT5WOS e CIT5SCH medidas de *impacto recente* ou o equivalente da produtividade para a medida de impacto total. As medidas CITARTSCH e CITARTWOS são medidas de *eficiência* científica por toda a carreira, enquanto que as medidas de CITART5SCH e CITART5WOS são medidas de *eficiência recente*. Finalmente, as medidas HWOS e HSCH são medidas *híbridas,* que combinam informações de produção e impacto acumulado.

1.2 Nível de análise

Há três alternativas para a análise das correlações entre as avaliações dos CAs e as várias métricas discutidas acima. A análise pode ser para cada grande área ou cada subárea do CNPq, para cada nível ou uma combinação de nível e subárea. A análise por grande área combina as várias medidas de correlação para todos os cinco níveis e para todas as subáreas daquela grande área, em uma única correlação; a análise por subárea combina as várias correlações para os diferentes níveis dentro da mesma subárea; finalmente, a análise por grande área e nível combina as correlações de cada subárea da grande área, mas mantém os níveis separados. Para este tipo de análise, teríamos uma medida de correlação para cada nível dentro da grande área, mas nenhuma distinção entre as várias subáreas.

Neste artigo, optamos pelo terceiro tipo de análise. Há duas razões para tal decisão: por que juntar as várias subáreas e por que não agrupar os níveis? As razões para agregar através das subáreas é que nos parece

razoável que diferentes CAs, dentro de uma mesma grande área, tenham critérios similares para a avaliação dos seus pesquisadores e, portanto, a agregação dessas correlações seria mais robusta a pequenos desvios dentro de um ou outro CA. A razão para não agregar os níveis é que é razoável que os diferentes grandes áreas tenham critérios diferentes para atribuir bolsas de nível 2, 1D e 1A, por exemplo. É razoável que um CA use a bolsa nível 2 para incentivar os pesquisadores da subárea a atingir um determinado nível de produção e, portanto, as decisões deste CA para as bolsas de nível 2 estariam fortemente correlacionadas com métricas de produção. Para o nível 1D, dado o salto nos privilégios associados a este nível, o CA pode se basear mais em medidas de produtividade e impacto recente. No outro extremo, um CA pode querer premiar um pesquisador com a bolsa 1A e, portanto, métricas de impacto total ou métricas históricas podem ser mais apropriadas para este nível. Assim, é possível que a correlação de uma métrica ou outra varie entre níveis. Mas é importante lembrar que não faremos a análise para cada CA e, sim, a análise agregada para todos os CA da grande área.

1.3 Pesquisas relacionadas

Há várias publicações sobre a correlação de diferentes medidas bibliométricas com a avaliação por pares e algumas que relacionam especificamente medidas bibliométricas com as bolsas do CNPq. Nenhuma das pesquisas cobre tantas áreas das ciências como esta.

Oliveira *et al.* (2012) descrevem o perfil dos pesquisadores de Medicina Clínica que têm Bolsa de Produtividade em Pesquisa e comparam algumas métricas, tais como produção total, produção nos últimos cinco anos, número de alunos orientados, total de citações recebidas, número de citações por ano, índices h e índice m (HIRSCH, 2005). Eles não estudam a correlação, mas apenas comparam as medidas bibliométricas entre os grupos correspondentes aos níveis do CNPq. Note que, diferentemente desta pesquisa, eles comparam pesquisadores nos vários níveis de bolsa e não pesquisadores que mudaram de nível num ano. De forma geral, eles só encontram diferenças significativas entre as medidas nos extremos da escala - entre os pesquisadores com bolsa 1A e 1B e os com bolsa nível 2.

Junior *et al*. (2010) estudam as diferenças de produtividades (artigos por ano) entre os pesquisadores bolsistas em Medicina, segundo as suas especialidades. Nenhuma comparação entre os níveis é feita no artigo. Em uma linha similar, Santos, Candido e Kuppens (2010) analisam o conjunto dos bolsistas de Química, estudando a distribuição de bolsas por região e estado, gênero, número de orientados, artigos publicados e índice h. Eles não comparam os resultados para os diferentes níveis. Cavalcante *et al.* (2008) fazem o mesmo tipo de análise para os pesquisadores de Odontologia, assim como Barata e Goldbaum (2003) para Saúde Pública e Mendes *et al.* (2010) para Medicina.

Há uma série de artigos publicados que mostram a correlação entre medidas bibliométricas e a avaliação por pares, mas não para pesquisadores brasileiros. Nesta linha, o resultado da avaliação por pares é considerado o "padrão ouro" ou o "resultado correto" e mostra-se que uma ou outra métrica tem uma correlação alta ou, pelo menos, importante, com a avaliação por pares.

A pesquisa feita por Raan (2006) faz uma comparação entre índice h e outros indicadores bibliométricos, como produção total, citações totais e citações por artigo, e a avaliação por pares, para 147 *grupos* de pesquisa (e não pesquisadores) em Química da Holanda. A pesquisa mostra que, para os "grupos bons," aqueles com as melhores avaliações pelos pares, há uma forte correlação com o índice h e o total de citações recebidas, mas esta correlação diminui para os grupos menos claramente "bons", isto é, grupos que não são tão bem avaliados.

Rinia *et al.* (1998) comparam indicadores bibliométricos e avaliação por pares para 56 grupos de pesquisa holandeses em Física da Matéria Condensada. Os resultados mostram que há correlações significativas entre várias métricas e as avaliações por pares, mas que essas correlações são diferentes para os grupos que fazem pesquisa aplicada e os que fazem pesquisa básica. Relevante para nossa pesquisa é que as maiores correlações (ro de Spearman) que eles encontram são de 0.68 para um grupo e 0.57 para outro.

Aksnes e Taxt (2004) estudam 34 grupos de pesquisa em Matemática e Ciências Naturais de uma mesma universidade da Noruega e mostram que as correlações entre a avaliação por pares e as diferentes medidas bibliométricas são positivas, mas baixas. E, diferente dos outros artigos acima, eles concluem que a principal razão para as discrepâncias são os "erros" ou as limitações dos processos de avaliação por pares e que grandes diferenças entre as duas avaliações devem ser analisadas com mais cuidado, pois pode ter havido erro na avaliação por pares.

2 Métodos e dados

Em 2011, obtivemos do CNPq a lista de todos os pesquisadores bolsistas, com o nível da bolsa e a subárea do pesquisador. Apenas os pesquisadores cujas bolsas foram renovadas ou modificadas em 2010 (e que, portanto, foram avaliados no segundo semestre de 2009), foram analisados, como discutido acima. Obtivemos os currículos Lattes de todos estes pesquisadores. O busca do Lattes foi feita em julho de 2011. Na época, o site do Lattes não permitia buscas automáticas e, portanto, achamos os currículos indiretamente, procurando pelo nome e a palavra "pesquisador" no site de buscas "http://search.yahoo.com/". Do currículo Lattes, extraímos, através de um programa, as seguintes informações:

a) ano de doutorado;

b) produção total: a quantidade de artigos em revistas, conferências, livros, e capítulos de livros;

c) produção nos últimos cinco anos: o total de produção desde 2006; e

d) o título das publicações.

Em setembro de 2011, coletamos, no Google Scholar, o número total de citações recebidas para cada uma das publicações do pesquisador listadas no seu Lattes. A busca no Scholar foi feita de forma automatizada, de três formas: pelo último nome (excetuando os nomes "Junior", "Neto" e afins), mais as iniciais do nome completo, pelo primeiro e último nome e pelo nome completo. As páginas retornadas pelo Google Scholar com as três alternativas de busca foram salvas para processamento posterior. Este processamento consistiu em verificar quais publicações da página tinha o mesmo título que alguma publicação listada no Lattes do pesquisador e em coletar o número de citações desta publicação. Esta correspondência entre o título da página retornada pelo Scholar e do Lattes não levou em consideração letras acentuadas, brancos ou diferenças entre letras maiúsculas ou minúsculas. Em poucos casos, a página retornada pelo Scholar retornava mais de uma referência para a mesma publicação. Neste caso, o programa escolhia a primeira referência e coletava o número de citações correspondentes.

Em outubro de 2011, coletamos os dados do *Web of Science* novamente, de forma automática, através da busca no site pelo último nome do pesquisador e a inicial do primeiro nome. De novo, foi feita a correspondência entre os títulos das publicações retornadas pelo WOS e os títulos das publicações listadas no Lattes do pesquisador. Com número de citações para cada uma das publicações do pesquisador (pelo Scholar e pelo WOS), calculamos as 15 medidas listadas acima.

2.1 Medida de correlação

A análise estatística dos dados foi feita, usando a correlação não-paramétrica gama de Kruskal (ou gama de Goldman e Kruskal) (SHESKIN, 2003). A medida de correlação não-paramétrica mais comum é a correlação de Spearman (ou *ro* de Spearman). Mas, como Sheskin (2003) explica, para situações nas quais há muitos casos de empate entre as duas medidas, o gama de Kruskal é mais indicado. No nosso caso, há muitos empates, pois, por exemplo, todos os bolsistas de um nível, que mantiveram suas bolsas, estão empatados entre si, assim como todos os bolsistas daquele nível que subiram de nível.

Dado o par (X1,Y1) e (X2,Y2), onde *X* indica se o bolsista teve seu nível reduzido, mantido ou aumentado, e *Y* é uma medida bibliométrica (por exemplo PROD), diremos que este par de bolsistas são *concordantes* se X1 > X2 e Y1 > Y2 , ou X1 < X2 e Y1 < Y2, isto é, se o bolsista 1 teve seu nível aumentado enquanto o bolsista 2 manteve o seu nível (X1>X2) e a produção total do bolsista 1 é maior que a do bolsista 2 (Y1> Y2), ou vice e versa. O par é *discordante* se X1 > X2 e Y1 < Y2 ou se X1 < X2 e Y1 > Y2. Finalmente, os casos são *empatados* se X1 = X2 ou Y1 = Y2.

O gama de Kruskal é calculado pela fórmula:

$$G = \frac{C - D}{C + D}$$

onde *C* é o número de pares concordantes e *D* o número de pares discordantes. Como toda medida de correlação, o gama de Kruskal varia de 1 (total concordância ou D=0) a -1 (total discordância ou C=0). Além disto, definimos que, quando *C*+*D*=0, a correlação é 0.

2.2 Análise dos resultados

A análise dos resultados é um pouco complexa, pois as diferentes correlações têm que ser combinadas entre si. Como mencionamos acima, teremos que combinar todas as correlações das várias subáreas de uma grande área em uma só correlação, para cada nível. Para combinar correlações, usaremos uma metodologia desenvolvida na área de meta-análise, que estuda como combinar resultados de diferentes experimentos (normalmente em Medicina) em um único resultado. Em particular, usaremos a técnica de modelo de efeitos fixos (*fixed effect model*) de Hedges e Vevea (1998) (veja FIELD, 2001), para um resumo das alternativas em relação à combinação de medidas de correlações em meta-análises. A técnica envolve os seguintes passos:

converter as medidas de correlação em uma medida normalizada, através da conversão de Fischer de *r* para *z*:

$$z_i = \frac{1}{2} \log_e \frac{1 - r_i}{1 + r_1}$$

combinar as várias medidas em uma média ponderada:

$$z = \frac{\sum w_i z_i}{\sum w_i}$$

onde os pesos são:

$$w_i = n_i - 3$$

calcular o erro padrão da medida z como sendo:

$$SE(z) = \sqrt{\frac{1}{\sum (n_i - 3)}}$$

computar a significância de z, assumindo que é uma variável com distribuição normal e com desvio padrão SE(z); e

converter a medida *z* para uma medida de correlação através da fórmula:

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

que é o inverso da fórmula de Fischer para converter *r* em *z*.

É preciso notar que esta técnica de combinar correlações foi desenvolvida para a tradicional correlação linear de Pearson. Não encontramos nenhuma fonte que afirme que esta técnica pode ou não pode ser usada para o gama de Kruskal. Em particular, a transformação de Fisher de *r* para *z* pode não ser válida para o gama. Por exemplo, a fórmula não é definida para $r = 1$, porque tal correlação (linear) seria extremamente rara. Mas o gama de Kruskal pode facilmente assumir o valor 1, basta que não haja nenhum par discordante. Para tratar da situação onde o gama é 1 ou -1, convertemos estes valores para 0.95 e -0.95, respectivamente.

3 Resultados

A Tabela 1 lista o número de pesquisadores incluídos nesta análise, por grande área e a avaliação qualitativa (se caíram, mantiveram ou subiram de nível). A Tabela 2 lista o numero de pesquisadores por grande área, para as quais nosso procedimento automático de busca de citações não funcionou. Note que os números são muito maiores para o WOS, indicando que, para alguns dos pesquisadores, o WOS não tinha publicações associadas ao nome. Para estes pesquisadores, agimos com cautela e consideramos que não tínhamos os dados corretos da sua produção indexada pelo WOS e, consequentemente, suas citações. Assim, tais pesquisadores não foram incluídos nos cálculos das métricas que usam o dado (PRODWOS, CITWOS, etc.). Note, também, que os casos sem dados de publicações indexadas pelos WOS são uma proporção significativa dos pesquisadores de CH e CSA e, portanto, as correlações com essas métricas para pesquisadores dessas grandes áreas devem ser tomados com cuidado.

Tabela 1 - Número de bolsistas que perderam a bolsa, caíram, mantiveram e subiram de nível, na avaliação de 2009, por grande área

Grande área	Perderam	Caíram	Mantiveram	Subiram	Total
CAG	72	118	391	115	696
CB	62	96	487	112	757
CE	93	117	583	163	956
CH	37	44	299	82	462
CS	72	115	275	99	561
CSA	49	55	180	54	338
EN	71	95	342	123	631
LLA	15	20	132	13	180

Fonte: Dados da pesquisa.

Tabela 2 - Número de bolsistas para os quais nosso procedimento automático de busca de citações não funcionou, para o Scholar e para o WOS

Grande área	Scholar	WOS
CAG	25	114
CB	34	99
CE	91	137
CH	17	340
CS	8	62
CSA	12	223
EN	24	67
LLA	13	146

Fonte: Dados da pesquisa.

Finalmente, a Tabela 3 reporta os principais resultados desta pesquisa. Para cada grande área e para cada nível de bolsa do CNPq, a tabela lista qual a métrica que tem a maior correlação com as decisões de aumento, manutenção ou rebaixamento da bolsa naquele nível e, ainda, o valor da correlação. As linhas faltantes na tabela indicam que nenhuma métrica teve uma correlação significativa com 90% de confiança, para aquele nível.

Tabela 3 - Métrica com maior correlação com as decisões de aumento, manutenção ou rebaixamento da bolsa de produtividade, para cada grande área e para cada nível da bolsa

Grande área	Nível	Métrica	Correlação
CAG	2	citwos	0.32
CAG	1D	citwos	0.61
CAG	1C	cit5sch	0.49
CAG	1B	prod5wos	0.56
CAG	1A	cit5wos	0.84
CB	2	prod5wos	0.52
CB	1D	prod5wos	0.59
CB	1C	citwos	0.51
CB	1B	prodwos	0.77
CE	2	citwos	0.54
CE	1D	cit5sch	0.46
CE	1C	cit5sch	0.44
CE	1B	citwos	0.94
CE	1A	cit5sch	0.88
CH	2	hwos	0.69
CH	1D	prod5	0.41
CH	1C	cit5sch	0.48
CS	2	cit5wos	0.49
CS	1D	prod5	0.49
CS	1C	cit5wos	0.56
CS	1B	cit5sch	0.57
CS	1A	prodtotal	0.55
CSA	2	prod5wos	0.95
CSA	1D	citart5sch	0.42
CSA	1C	hsch	0.72
EN	2	prod5wos	0.53
EN	1D	cit5wos	0.78
EN	1C	prod5wos	0.68
EN	1B	prod5wos	0.71
LLA	2	citart5sch	0.20
LLA	1D	idade	0.52
LLA	1B	prodtotal	0.72
LLA	1A	prod5	0.86

Obs.: Linhas faltantes na tabela indicam que nenhuma métrica teve correlação significativamente diferente de 0 para aquele nível. Os significados das siglas estão listados no texto.

Fonte: Dados da pesquisa.

A grande área de Ciências Agrícolas (CAG) parece usar uma combinação de impacto acumulado e impacto recente nas suas decisões. As correlações mais altas para o nível 1D e 1A parecem indicar que há uma maior consenso e uniformidade sobre o que é considerado como importante para a avaliação do pesquisador nestes níveis, no entendimento coletivo da área. As Ciências Biológicas (CB) parecem se basear mais em produção e produção recente de artigos indexados no WOS. Não há correlação significativa para o nível 1A. Este é um fenômeno que se repete em outras 4 grandes áreas.

As Ciências Exatas (CE) se baseiam apenas em impacto acumulado e impacto recente. Esta é a grande área que apresenta maior uniformidade através dos níveis – todas as métricas são baseadas em citações. A grande área de Ciências Humanas (CH) apresenta uma grande variação nas métricas, mais fortemente correlacionadas com as decisões: índice h, produção recente e citações recentes (pelo Scholar), enquanto que para os níveis 1B e 1A, não há correlação significativa. Lembramos que CH é uma das grandes áreas nas quais a nossa coleta de dados no WOS não retornou resultados na maioria dos casos e, portanto, nenhuma métrica baseada nos dados do WOS deveria ter alta correlação.

Ciências da Saúde (CS) apresentam uma grande variação entre os níveis, alternando entre produção e citação, mas quase todos eles recentes. Ciências Sociais Aplicadas (CSA) apresenta a maior variação entre as métricas através dos vários níveis. Discutiremos, nas conclusões, o que tal variação pode indicar.

Engenharias (EN) se baseiam em métricas recentes, na sua grande maioria, produção qualificada (indexada no WOS). Finalmente, a grande área de Linguística, Letras e Artes (LLA) também alterna citações, produção e mesmo idade científica, como métrica mais correlacionada com as decisões de bolsas de produtividade.

4 Discussão e conclusão

Há duas formas de encarar esta pesquisa. A primeira, que enfatizamos neste artigo, é que estamos analisando os critérios que cada grande área usa para avaliar seus pesquisadores, na esperança que os CAs correspondentes verifiquem se suas decisões de avaliação correspondem a métricas objetivas e, ainda, se estas métricas correspondem as suas visões do que significa a qualidade de um cientista. A segunda forma é considerar que a avaliação do CNPq é a correta avaliação por pares e, assim, este artigo mostra que existe correlação entre a avaliação por pares e métricas bibliométricas, como outros trabalhos já mostraram (RAAN, 2006; RINIA *et al.*, 1998; AKSNES; TAXT, 2004).

Acreditamos que é mais produtivo encarar o artigo pelo primeiro enfoque. As grandes áreas devem ponderar se as práticas que estão seguindo para avaliar seus pesquisadores estão de acordo com os seus objetivos para as bolsas de produtividade. Obviamente, não existe um

objetivo "certo" ou único para a atribuição de bolsas e, portanto, não deve existir uma única métrica que deve ser seguida. No entanto, nos parece que há duas grandes vertentes para explicar quais são os objetivos de uma bolsa de produtividade, que, se não são contraditórios entre si, são pelo menos diferentes. O primeiro objetivo possível é de **premiar** cientistas de qualidade. O segundo objetivo possível é **incentivar** a produção de qualidade dos cientistas brasileiros. Há uma diferença importante entre estes dois objetivos: se o objetivo é premiar os cientistas pela qualidade e importância do seu trabalho, então, a história passada do pesquisador é o fator mais importante. Caso o objetivo seja incentivar a produção de qualidade e relevância, então, o futuro deste pesquisador é mais importante que seu passado. É claro que, nesta segunda alternativa, o passado é importante, mas apenas como ferramenta para prever o futuro do pesquisador – na falta de melhores dados acredita-se que o pesquisador, no futuro, terá os mesmos resultados (do ponto de vista de produção científica) que no passado ou, pelo menos, no passado recente. Além do mais, as duas vertentes de objetivos têm impacto muito diferente no agraciado. Quem recebe um prêmio, não precisa mais continuar fazendo o que fez para ganhar o prêmio, apenas quando outra pessoa tiver acumulado história suficiente que o prêmio será transferido. Quem recebe um incentivo, deve manter pelo menos a mesma produção que fez com que recebesse o incentivo, pois, senão, ele será retirado.

Caso o objetivo seja premiar, então, alguns aspectos da produção são mais relevantes que outros, na avaliação dos cientistas. A visão mais comum é que a qualidade de um cientista é melhor avaliada pelo seu impacto nas ciências e este impacto é melhor avaliado (de forma geral) pelo número citações recebidas pelo cientista. Obviamente, há visões divergentes desta, por exemplo, que a qualidade de um cientista é melhor avaliada pelo seu impacto na vida da população em geral[2]. Estamos cientes que, mesmo dentro da visão mais comum, a qualidade de um cientista não é *proporcional* ao número de citações recebidas, já que o número de citações recebidas por um cientista depende da área (PODLUBNY, 2005) e de outros fenômenos (BORNMANN; DANIEL, 2008). Mas, adotaremos a posição que a qualidade do cientista é pelo menos uma função do número de citações. Assim, se o objetivo é premiar a qualidade do cientista, então, as métricas de impacto (CITWOS, CITSCH) deveriam ter uma alta correlação com as decisões de atribuição de bolsas de produtividade. Uma visão talvez mais ingênua da qualidade de um cientista é usar as métricas de produção (PRODTOTAL e PRODWOS) como medida de qualidade. Parece-nos que decisões de atribuição de bolsa com alta correlação com métricas de produção, estão, na verdade, premiando o *esforço* do pesquisador, premiando o fato dele ter "feito seu trabalho", mas não necessariamente ter causado um impacto na sua área cientifica.

2 Não conhecemos uma citação específica que defende esta posição, mas acreditamos que esta visão é que esta por trás de posições que "o papel da ciência é resolver os problemas da população".

De qualquer forma, todas as métricas apropriadas para prêmios não decrescem com o tempo, isto é, uma vez que o pesquisador tem um número de citações que fazem com que ele mereça o prêmio, este número nunca decrescerá e, portanto, o pesquisador não precisa fazer mais nada. O prêmio (neste caso, a bolsa) só será transferido para outro pesquisador que alcançar um número maior de citações, mas, enquanto isto não acontece, o primeiro pesquisador manterá sua bolsa mesmo que nunca mais receba nenhuma citação. O mesmo vale para total de publicações.

Caso o objetivo seja incentivar a produção de qualidade dos cientistas brasileiros, então, provavelmente, o aspecto mais relevante para avaliar um pesquisador é a sua produção recente. É claro que existem fatores que modificam a produtividade de um pesquisador, como disponibilidade de alunos, existência de financiamento para a pesquisa, situação pessoal, estágio na carreira profissional, etc. Mas, na falta de outras informações, provavelmente a produtividade no futuro próximo será a mesma que a produtividade no passado recente. Assim, nos parece que, se o objetivo é incentivar a produção de qualidade, deve-se privilegiar os pesquisadores com métricas de produção ou impacto ou eficiência *recente*. E, na direção que qualidade é mais bem medida por impacto que por produção, então, provavelmente, uma alta correlação com eficiência ou impacto recente seria mais desejável.

Há as métricas que, nos parece, não contribuem para nenhum destes dois objetivos gerais. Embora o índice h tenha ganhado certa proeminência recentemente, não é totalmente claro qual objetivo seria servido por tomar decisões altamente correlacionadas com o índice h. O índice h não decresce com o tempo e, portanto, há um componente de prêmio em usá-lo como medida para atribuir ou não bolsas, mas não é claro porque tal medida seria melhor que, por exemplo, total de citações recebidas. Não surpreendentemente, há apenas dois casos nos quais o índice h é a métrica com maior correlação com as decisões sobre bolsa (CH nível 2 e CSA nível 1C). Outra família de métricas, que não sabemos como interpretar, são as métricas de eficiência (citações médias por artigo). Não surpreendentemente, métricas desta família não aparecem como as mais altamente correlacionadas com as decisões em nenhum dos casos.

Finalmente, a métrica histórica (IDADE) aparece em apenas um caso na Tabela 3 (LLA nível 1D). Decisões fortemente correlacionadas com a idade científica do pesquisador estão privilegiando (mais do que qualquer outra medida) os pesquisadores mais velhos. Claramente, tais decisões estão mais na linha de prêmio: pesquisadores mais velhos provavelmente tiveram mais alunos, prestaram mais serviços para a sua comunidade científica, etc., e, provavelmente, estão sendo premiados por isso. Mas o impacto nos pesquisadores mais novos pode ser bastante negativo - se nenhuma outra métrica de produção é relevante para receber ou melhorar a sua bolsa de produtividade, então, um pesquisador mais jovem só deve esperar "o tempo passar".

No contexto das hipóteses levantadas acima, passaremos a analisar os resultados das métricas mais fortemente correlacionadas com as avaliações de cada uma das grandes áreas.

A grande área de Ciências Agrícolas (CAG) tem uma preferência por prêmio (citações), no começo de carreira científica (níveis 2 e 1D), e passa, nos níveis seguintes, para uma direção de incentivo. Esta política acabaria permitindo que a entrada no sistema de bolsas (níveis 2 e 1D) se dê no ritmo do pesquisador, ou seja, o pesquisador tem o tempo que precisar para acumular as citações necessárias no conjunto de sua obra. Mas, uma vez nos níveis mais altos, o pesquisador deve se pautar pela sua produtividade de artigos em revistas, já que a métrica se baseia no WOS e em citações. Esta parece ser uma política interessante, que respeita os diferentes ritmos dos pesquisadores juniores. Mas, acreditamos que a política seria ainda melhor se para aos níveis superiores houvesse uma maior correlação com métricas de impacto recente.

A segunda grande área com uma política mais simples de interpretar é a área de Engenharias (EN). Produção recente em revistas indexadas é a métrica importante. Há uma forte correlação com citações recentes no nível 1D, que é um pouco difícil de interpretar, mas, de modo geral, a área preza produtividade em revistas. A ênfase em produção recente mostra que a área encara bolsas como incentivo e não prêmio. O fato que as decisões relativas ao nível 1A não possuem correlação significativa com nenhuma métrica, que é verdade, também, para outras grandes áreas, é possivelmente derivado do fato que há muito pouca variação para o nível 1A, ou seja, a maioria das bolsas 1A se mantém. Na nossa amostra de 2010, houve sete mudanças no nível 1A para CAG, uma para CB, cinco para CE, duas para CH, 4 para CS, 0 para CSA, duas para EN e três para LLA. Ou seja, a não ser em poucas áreas, não há número suficiente de casos para obter qualquer correlação significativa. Mas o baixo número de variações nas bolsas 1A revela outro fenômeno. Os sete casos de mudanças no nível 1A para CAG representam 15% das bolsas 1A da grande área (dentre as que foram avaliadas, no final de 2009). Para CB 1%, para CE 5% para CH 5%, para CS 10%, para CSA 0%, para EN 4% e para LLA 14%. Compare com as proporções de bolsas 1D que variaram naquele ano: CAG 40%, CB 39%, CE 32%, CH 38%, CS 72%, CSA 51%, EN 42% e LLA 7%. Ou seja, há muito mais variação no nível 1D que no nível 1A. Os CAs não mostram muito interesse em fazer grandes modificações no nível 1A; uma vez que um pesquisador ascende ao nível 1A, ele passa a ser mais intocável que nos níveis anteriores. Isso pode indicar que o componente de prêmio é mais importante para este nível ou que as decisões sobre o nível 1A são mais "políticas" e que, portanto, os CAs se sentem com menos liberdade de tomar decisões que modifiquem o *status quo*.

A grande área de Ciências Exatas (CE) parece focar fortemente em citações. Não acreditamos que a variação entre citações calculadas pelo WOS ou pelo Scholar mostre um fenômeno relevante. A principal

diferença é que o Scholar conta citações feitas e recebidas por artigos que não apenas os publicados em revistas indexadas, que, no caso das subáreas dentro da CE, significa artigos publicados em conferências científicas. Finalmente, a área de Ciências Biológicas (CB) parece apresentar um padrão de incentivo para os níveis iniciais (2 e 1D), pois usa a produção indexada recente e de prêmio para os níveis mais altos (1C e 1B), embora a mudança de citações para produção total é um pouco difícil de explicar.

As outras grandes áreas (CH, CS, CSA e LLA) são mais difíceis de interpretar. As variações entre os vários níveis, talvez possam ser explicadas pela variação das correlações *dentro* da grande área. Um pressuposto desta pesquisa é que faz sentido agregar as correlações através de várias subáreas de uma grande área, porque acreditamos que as subáreas têm práticas de avaliação em comum. Talvez isto não seja verdade para a maioria destas cinco grandes áreas.

Finalmente, os valores numéricos de algumas das correlações são altos, principalmente se comparados com os resultados que Rinia *et al.* (1998) consideram altos (0.57 e 0.68). Mas não podemos comparar os resultados diretamente. Os valores do gama de Kruskal são sempre mais altos que os do ro de Spearman, já que o denominador do gama só contém os valores discordantes e concordantes, enquanto que o ro usa, também, os valores empatados. Infelizmente, a escolha da medida de correlação não é neutra. Fizemos a análise da tabela 3, usando o ro de Spearman em vez do gama e, de forma geral, um número maior das correlações passam a ser não-significativas e os valores numéricos das correlações são mais baixos. Em vez de 33 correlações significativas, usando o gama, teríamos apenas 20 usando o ro, e a maior correlação teria valor 0.60 (em vez de 0.94 no caso do gama). Isto era de se esperar, dado que o ro é necessariamente menor que o gama. Mas o uso do ro de Spearman modifica a métrica com maior correlação: das 20 correlações significativas usando o ro, cinco não correspondem à métrica obtida usando o gama. E, na maioria dos casos, estas métricas novas tornam a análise da grande área mais difícil de interpretar (como as áreas CB, CH, CS, CSA e LLA, no caso do gama).

Esta pesquisa tem dois objetivos gerais. O primeiro é mostrar que decisões relativas a Bolsas de Produtividade em Pesquisa possuem correlações (provavelmente implícitas) com medidas bibliométricas. Em alguns casos, a correlação é bastante alta. O segundo objetivo é induzir os diferentes CA a repensarem seus critérios de avaliação. A correlação entre as decisões e as métricas reflete a visão do CA do que deve ser um bom cientista na área? O CA tem claro o papel que ele espera da bolsa de produtividade; incentivo ou prêmio? E as decisões dos CAs estão refletindo estas visões?

Referências

AKSNES, D.; TAXT, R. Peer reviews and bibliometric indicators: a comparative study at a Norwegian university. *Research evaluation*, Beech Tree Publishing, v. 13, n. 1, p. 33-41, 2004.

BARATA, R.; GOLDBAUM, M. Perfil dos pesquisadores com bolsa de produtividade em pesquisa do CNPq da área de saúde coletiva. *Cadernos de Saúde Pública*, SciELO Public Health, v. 19, n. 6, p. 1863-1876, 2003.

BORNMANN, L.; DANIEL, H. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, Emerald Group Publishing Limited, v. 64, n. 1, p. 45-80, 2008.

CAVALCANTE, R. *et al.* Perfil dos pesquisadores da área de odontologia no Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). *Rev. bras. epidemiol*, v. 11, n. 1, p. 106-113, 2008.

FIELD, A. P. Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. *Psychol Methods*, v. 6, n. 2, p. 161-80, Jun. 2001.

HEDGES, L. V.; VEVEA, J. L. Fixed- and random-effects models in meta- analysis. *Psychological Methods*, v. 3, p. 486-504, 1998.

HIRSCH, J. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United states of America*, National Academy of Sciences, v. 102, n. 46, p. 16569, 2005.

JUNIOR, H. M. *et al.* Pesquisadores do CNPq na área de medicina: comparação das áreas de atuação. *Rev Assoc Med Bras*, SciELO Brasil, v. 56, n. 1, p. 478-83, 2010.

MENDES, P. *et al.* Perfil dos pesquisadores bolsistas de produtividade científica na medicina no CNPq, Brasil. *Revista Brasileira de Educação Medica*, SciELO Brasil, v. 34, n. 4, p. 535-41, 2010.

OLIVEIRA, E. *et al.* Comparison of brazilian researchers in clinical medicine: are criteria for ranking well-adjusted? *Scientometrics*, Springer, v.90, p. 429-443, 2012.

PODLUBNY, I. Comparison of scientific impact expressed by the number of citations in different fields of science. *Scientometrics*, Springer, v. 64, n. 1, p. 95-99, 2005.

RAAN, A. van. Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, Springer, v. 67, n. 3, p. 491-502, 2006.

RINIA, E. *et al.* Comparative analysis of a set of bibliometric indicators and central peer review criteria: evaluation of condensed matter physics in the Netherlands. *Research Policy*, Elsevier, v. 27, n. 1, p. 95-107, 1998.

SANTOS, N.; CANDIDO, L.; KUPPENS, C. Produtividade em pesquisa do CNPq: análise do perfil dos pesquisadores da química. *Quimica Nova*, SciELO Brasil, v. 33, n. 2, p. 489-95, 2010.

SHESKIN, D. J. *Handbook of parametric and nonparametric statistical procedures*. Chapman and Hall/CRC, 2003. cap. 32. ISBN 978-1-58488-440-8. Disponível em: <http://dx.doi.org/10.1201/9781420036268.ch32>. Acesso em: Set. 2012.