UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

FAHAD KALIL

Sobre estatística de dados bibliométricos em grupos de pesquisadores: universalidade e avaliação

Dissertação apresentada como requisito parcial
para a obtenção do grau de
Mestre em Ciência da Computação

Prof. Dr. José Palazzo Moreira de Oliveira
Orientador

Prof. Dr. Roberto da Silva
Co-orientador

Porto Alegre, fevereiro de 2012

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Kalil, Fahad

Sobre estatística de dados bibliométricos em grupos de pesquisadores: universalidade e avaliação / Fahad Kalil. – Porto Alegre: PPGC da UFRGS, 2012.

55 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Informática, Porto Alegre, BR–RS, 2012. Orientador: José Palazzo Moreira de Oliveira; Co-orientador: Roberto da Silva.

1. Análise de pesquisadores. 2. Índice h. 3. Bibliometria. I. de Oliveira, José Palazzo Moreira. II. da Silva, Roberto. III. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Carlos Alexandre Netto
Pró-Reitor de Coordenação Acadêmica: Prof. Rui Vicente Oppermann
Pró-Reitor de Pós-Graduação: Prof. Aldo Bolten Lucion
Diretor Pró-Tempore do Instituto de Informática: Prof. Luís da Cunha Lamb
Coordenador do PPGC: Prof. Álvaro Freitas Moreira
Bibliotecária-chefe do Instituto de Informática: Beatriz Regina Bastos Haro

“ You’ve got the future on your side
You’re gonna be fine now
I know whatever you decide
You’re gonna shine. ”
(DREAM THEATER : THE ANSWER LIES WITHIN)

AGRADECIMENTOS

Meus sinceros agradecimentos a todos que contribuíram para o desenvolvimento desse trabalho. Em especial, gostaria de agradecer ao meu co-orientador, Prof. Dr. Roberto da Silva, que foi fundamental na elaboração dessa dissertação por sua capacidade ímpar de construir conhecimento e pelo perfil de um verdadeiro cientista. Sua disposição, comentários e sugestões engrandeceram em muito a pesquisa realizada, bem como me fizeram ter uma visão diferenciada de como um trabalho deve ser construído com ênfase na excelência do problema, da análise e dos resultados obtidos.

Agradeço também ao meu orientador, Prof. Dr. José Palazzo Moreira de Oliveira, por me proporcionar a oportunidade de realizar o Mestrado Acadêmico em uma universidade de nível internacional como a UFRGS e por me incluir em discussões e reuniões durante o Programa que me fizeram crescer como pesquisador, além de poder conviver com alguém que possui características de um grande empreendedor em prol da ciência.

Agradeço ao Prof. Me. Cristiano Roberto Cervi por me apresentar ao mundo da pesquisa, incentivar trabalhos inovadores e mostrar a importância da realização do mestrado. Suas conversas e suporte no meu primeiro ano de mestrado foram muito úteis. Deixo aqui um sincero agradecimento a Prof. Dr. Carina Friedrich Dorneles que, por ter participado da banca em meu trabalho de conclusão de curso, me fez acreditar que o trabalho ali iniciado poderia render bons frutos caso expandido e aprofundado.

À UFRGS e ao Instituto de Informática pela infraestrutura disponibilizada, bem como ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro. Aos professores pelos ensinamentos transmitidos através das disciplinas cursadas e aos funcionários que se mostraram sempre atenciosos e prestativos na resolução dos problemas e dúvidas.

Aos colegas de laboratório do Grupo de Sistemas de Informação e Modelagem Computacional que me receberam muito bem, sendo companheiros tanto para descontração como para tratar de questões importantes durante o período do curso. Estendo também um agradecimento aos colegas do vôlei, que me proporcionou uma ótima forma de integração e alívio do estresse do dia a dia de um pesquisador.

À minha família por proporcionar toda estrutura, emocional e financeira, para que conseguisse realizar o mestrado da melhor forma possível, bem como a motivação para buscar qualificações cada vez melhores. Agradeço à minha namorada, Natalia Zancan, pela compreensão, paciência e amor durante esse período, no qual ficamos mais tempo longe do que próximos. Agradeço a minha irmã, Samara, pelos bons momentos compartilhados na capital gaúcha.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CG	Cumulative Gain (Ganho Acumulado)
DCG	Discounted Cumulative Gain
GS	Google Scholar
ISI	Institute for Scientific Information
NDCG	Normalized Discounted Cumulative Gain
PRD	Physics Review D
WoS	ISI Web of Science
WWW	World Wide Web

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Agências de fomento à pesquisa, centros de pesquisas, universidades e a comunidade científica de uma forma geral buscam incessantemente pelo aperfeiçoamento e aumento da qualidade da produção científica de seus pesquisadores. Logo, faz-se necessário que sejam providas ferramentas e métodos eficazes para obtenção de avaliações coerentes. Vários métodos têm sido propostos ao longo dos anos e diferentes formas de avaliação vêm sendo empregadas em órgãos reguladores, como a agência brasileira de pós-graduação CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), todavia algumas questões ainda foram pouco exploradas. Com o advento do índice h (*h-index*) de Hirsch, que une fundamentalmente quantidade com qualidade, pois avalia um conjunto de artigos de acordo com sua distribuição de citações, uma série de estudos com as mais variadas estatísticas têm sido propostos. Contudo, não há na literatura atual, por exemplo, uma expressão analítica para distribuição de índices h de um grupo de pesquisadores, nem a verificação da existência de universalidade desta distribuição para diferentes grupos e bases de dados. Este trabalho aborda, num primeiro momento, um estudo sobre a distribuição de índice h e de citações em três áreas científicas distintas: física, biologia e ciência da computação, que possuem diferentes práticas de publicações e métodos de pesquisa. O trabalho ainda propõe uma nova métrica para construção de *rankings* baseado no índice h sucessivo de segunda ordem, por nós denominada $s - index$, o qual torna possível a comparação entre grupos de pesquisadores de tamanhos diferentes, assim proporcionando, por exemplo, uma comparação em termos de produtividade de grupos com grande quantidade de pesquisadores e grupos menores, mas que ainda assim reflita seus potenciais de produção. Por fim, é realizado estudo da correlação entre o $s - index$ desenvolvido no trabalho e a classificação de cursos de mestrado e doutorado recomendados e reconhecidos realizada pela CAPES, usando dados de pesquisadores de programas de pós-graduação em física e biologia. As abordagens apresentadas podem ser usadas na classificação de grupos de pesquisadores, a partir de uma visão quantitativa, tentando eliminar, assim, métodos qualitativos de avaliação de difícil generalização e replicação.

Palavras-chave: Análise de pesquisadores, índice h, bibliometria.

About statistics on bibliometric data of researchers' groups: universality and evaluation

ABSTRACT

Research financing agencies, research centers, universities and the scientific community are frequently seeking for improvement and enhancement on the quality of researchers' work. Therefore, it is necessary to provide optimized tools and methods to get consistent evaluations. Several methods have been proposed over the years and different forms of evaluation are used by agencies such as Coordination for the Improvement of Higher Level Personnel (Capes), although some issues have been overlooked. With the invention of the h-index (Hirsch), that binds quantity with quality by measuring a set of papers according to their citation distribution, many studies have been considered with several different statistical methods. In the current literature, it is not found an analytical expression for h-index distribution over a researchers' group, nor the proof of universality in this kind of distribution for different groups and databases. This master thesis discusses primarily a study about h-index distribution and citations in three distinct scientific fields: physics, biology and computer science, which has different publication and research practices. Also, it is proposed a novel metric for ranking based on successive h-index, named as s-index, which makes possible to compare researchers' groups with different sizes, providing for example, a comparison in terms of productivity on higher and smaller groups of researchers, reflecting their skills on scientific production. A correlation study is conducted in order to compare the s-index, developed in this thesis, with the classification of post-graduation courses performed by Capes, using data from post-graduation researchers in physics and biology. The approaches presented can be used to classify researchers' groups through a quantitative view, by eliminating some qualitative evaluations that are hardly generalizable and replicable.

Keywords: researchers' analysis, h-index, bibliometrics.

1 INTRODUÇÃO

Com o advento da *World Wide Web* (WWW), o compartilhamento de dados e a disseminação de conteúdos tornaram-se muito mais rápidos e dinâmicos, fazendo com que as comunidades científicas, através de seus pesquisadores, aumentassem a quantidade de trabalhos publicados, bem como o alcance por esses atingido. Barreto (2007) observou que houve uma modificação estrutural, na relação tempo e espaço, no fluxo de informação científica provocada pela comunicação eletrônica. Sendo assim, o modelo atual de pesquisa nas universidades baseia-se na consulta de estoques de periódicos e artigos eletrônicos, ou portais, permitindo uma nova forma de acesso, por meio da qual os usuários compartilham de um mesmo material concorrentemente, promovendo uma interoperabilidade e também uma utilização de forma rápida sem o incômodo do deslocamento físico à biblioteca (COSTA, 2007).

Neste contexto, no qual a geração de conhecimento é quase instantânea, as agências de fomento à pesquisa, centros de pesquisa, universidades, bem como outros participantes da comunidade científica, buscam constantemente o aperfeiçoamento de seu quadro de pesquisadores e da qualidade das publicações. Para tanto, é necessário adotar medidas para avaliação dos pesquisadores no intuito de canalizar recursos para grupos de comprovada competência em áreas específicas do conhecimento estimulando o desenvolvimento da ciência e tecnologia.

Diversos estudos abordando o problema de avaliação de pesquisadores têm sido conduzidos, porém não existem formas gerais, unificadas e absolutas de avaliar pesquisadores, pois essa análise depende de diversos fatores que muitas vezes partem de uma visão subjetiva. Algumas das formas mais simples de avaliação baseiam-se na análise de citações, co-citações, acoplamento bibliográfico e fator de impacto, e são estudadas normalmente dentro da área de bibliometria. As variáveis usualmente utilizadas para medir produtividade são: número total de publicações e suas citações (ANASTASIADIS; ALBUQUERQUE; ALBUQUERQUE, 2009). Outro foco na avaliação está no estudo e aplicações de métricas baseadas no índice h proposto por J. Hirsch e que tem como objetivo quantificar a importância da produção científica e o impacto do trabalho realizado por um cientista durante sua vida. Essa métrica funciona reduzindo a complexidade da distribuição de dados para quantificar a importância de toda produção em pesquisa de um cientista em uma única medida (HIRSCH, 2005). Em trabalhos recentes, alguns autores propuseram variações do índice h para medir não apenas pesquisadores de forma isolada, mas também grupos de pesquisadores (ex.: pesquisadores de um mesmo instituto). Schubert (2007) criou um conceito chamado índice h sucessivo de segunda ordem, por meio do qual se calcula, a partir do h de cada autor de um grupo, um novo *h* (chamado h_2) representativo do índice do h desse grupo. Esse conceito já havia sido definido superficialmente em Prathap (2006) para apoiar a classificação ou avaliação de institutos de pesquisa.

Esta dissertação tem com objetivo explorar de forma quantitativa dois diferentes estudos em bibliometria: (i) fazer uma análise detalhada sobre a distribuição de índice h, analisando sua universalidade sobre diferentes bases de dados; (ii) propiciar um método para construção de *ranking* de pesquisadores baseado numa reformulação adequada de um índice da literatura (chamado índice h sucessivo de segunda ordem), tornando possível a comparação de grupos de diferentes tamanhos e com isso, computar as correlações entre os resultados gerados e os resultados presentes em avaliações efetuadas por agências nacionais de fomento.

Estes objetivos vão ao encontro de uma tendência mundial elucidada por Julia Lane, na Revista Nature. A pesquisadora aponta a necessidade de métricas melhores e mais científicas que as atuais sobre o desempenho de pesquisadores para evitar o risco de "decidir por financiamentos equivocados ou afastar bons cientistas" (LANE, 2010).

Esta dissertação está organizada da seguinte forma: No Capítulo 2 são apresentados os trabalhos relacionados e necessários à compreensão de termos e conceitos usados no restante da dissertação, divididos em Métricas de avaliação de pesquisadores e Distribuição de citações. No Capítulo 3 são apresentadas as formas de aquisição dos dados utilizados, nas bases de dados ISI Web of Science e Google Scholar, além do uso auxiliar da Plataforma Lattes do CNPq. O Capítulo 4 detalha os modelos e abordagens desenvolvidas e adotadas no trabalho, que são: a Formulação para distribuição de índices h para grupos de pessoas e uma Proposta para *ranking* de pesquisadores baseado no índice h sucessivo de segunda ordem. O Capítulo 5 descreve os experimentos envolvendo os modelos desenvolvidos e os resultados relevantes obtidos. O Capítulo 6 apresenta as considerações finais quanto ao trabalho desenvolvido nesta dissertação e mostra possíveis trabalhos futuros dentro da mesma área de pesquisa. O Apêndice I apresenta métodos para estimativa de parâmetros e, por fim, o Apêndice II apresenta métricas de avaliação de *rankings*.

2 TRABALHOS RELACIONADOS

Os trabalhos relacionados sintetizam o ponto de partida da pesquisa apresentada nesta dissertação, de forma a expor as diferentes técnicas de avaliação e análise de conjuntos de dados sobre pesquisadores de diferentes áreas. Os problemas de identificação e classificação de pesquisadores quanto à produção e qualidade é tarefa bastante subjetiva e pouco pragmática. Isto motiva cientistas preocupados com a qualidade e excelência acadêmica a buscarem formas novas e mais completas de avaliação que sejam, ao mesmo tempo, de simples adoção e menos complexas que métodos atuais. Como exemplo, podemos citar a avaliação de cursos de pós-graduação realizada pela CAPES (Coordenadoria de Aperfeiçoamento de Pessoal de Nível Superior do Ministério da Educação) ou por outras instituições, nas quais a coleta e análise de dados demanda muito tempo, trabalho manual e grandes deslocamentos de pesquisadores em reuniões presenciais.

2.1 Métricas de avaliação de pesquisadores

Medir a produção de pesquisadores é uma atividade que tem despertado grandes discussões nas comunidades científicas. Atualmente, quase todas as avaliações de competências (como aceitação de projetos de pesquisa, contratação de pesquisadores, concessão de recursos, entre outros) dependem de uma grande compreensão, por parte dos gestores, dos méritos científicos dos pesquisadores envolvidos (ALONSO et al., 2009). Portanto, a exigência cada vez maior por informações que levam a definição de investimentos e prioridades na ciência fez com que os órgãos envolvidos optassem por métodos quantitativos ao analisar a produção científica (MUGNAINI, 2006).

Um conceito bastante difundido na área é o fator de impacto, empregado por Eugene Garfield para suporte às análises do ISI (Institute for Scientific Information). Nesse contexto, consiste em "[...] dividir o número total de citações obtidas por um periódico em um ano qualquer pelo número de artigos publicados naquele ano" (RODRIGUES, 1981 apud ARAÚJO, 2006). Em sua evolução, além da análise de periódicos, o fator de impacto também foi expandido para análise individual de pesquisadores, sendo aplicada como a divisão do número de citações recebidas por um autor pelo número de trabalhos com no mínimo uma citação (SIMONS, 2008). Esse índice busca a classificação de autores em torno da relevância de seus trabalhos, priorizando aqueles que receberam muitas citações em muitos trabalhos publicados e não apenas um grande número de citações diluídas em uma quantidade vasta de trabalhos com pouca repercussão no meio científico (ARAÚJO, 2006).

Os principais conceitos utilizados nesta dissertação serão apresentados nas seções 2.1.1 e 2.1.2. Na Seção 2.1.3 também é descrito o índice g, criado por Egghe, que usa conceitos do índice h de forma a deixá-lo mais justo na sua forma de avaliação.

2.1.1 Índice h

Em 2005, Hirsch apresentou uma nova forma de quantificar a importância da produção científica e o impacto do trabalho realizado por um cientista durante sua vida. O índice criado por Hirsch chama-se índice h e funciona reduzindo a complexidade da distribuição de dados para quantificar a importância das pesquisas de um cientista em uma única medida. De acordo com sua definição, esse índice revela que: "Um cientista possui um índice *h* se *h* de seus artigos possuem pelo menos *h* citações cada e os outros artigos (N_p - *h*) possuem um número menor ou igual a h de citações cada, onde $0 \leq h \leq N_p$" (HIRSCH, 2005).

Hirsch argumenta que um índice h com alta magnitude identifica aqueles pesquisadores que produzem de forma consistente uma quantidade de bons trabalhos durante um período significativo de tempo, ao invés de pesquisadores que escrevem grandes artigos, altamente citados, durante um curto periodo e depois ficam, de certa forma, estagnados em termos de produção científica. A simplicidade do índice h justifica a sua popularidade, pois embora outras métricas sejam mais eficazes, elas possuem menor aceitação por não propiciarem uma ligação clara com o que está sendo analisado (FRANCESCHINI; MAISANO, 2010). Portanto, o índice h pode ser visto como uma alternativa ao fator de impacto, pois abrange uma fatia de tempo maior e pode ser usado tanto para avaliação individual como de grupos (ex.: instituições), sempre levando em conta suas limitações (BORNMANN; MARX, 2011).

O uso do índice h, ao invés de outras abordagens que medem a produção científica de pesquisadores, é sustentado em algumas vantagens listadas por Rousseau (2008), como: (i) apresenta-se como um índice melhor que o total de número de publicações ou número total de citações de forma isolada; (ii) incentiva trabalhos de alta qualidade (ou pelo menos de alta visibilidade); (iii) pode ser aplicado a vários níveis de agregação, como no h_2; (iv) é um indicador robusto à medida que ruídos e erros nos dados utilizados causam poucos efeitos no resultado; (v) os autores devem ter uma boa distribuição de citações entre duas publicações, pois o índice h pouco se altera apenas por uma ou duas publicações muito citadas. Como todo índice, também possui algumas desvantagens principais: (i) é uma medida dependente da área de atuação, pois diferentes áreas possuem diferentes práticas relacionadas a citações e produtividade; (ii) pode ser influenciada por autocitações (ex.: um autor que cita muito suas outras publicações); (iii) número de co-autores pode influenciar as citações recebidas; (iv) pode trazer dificuldades na coleta de informações a respeito de todas as publicações de um autor para determinar um índice h livre de erros e também existem problemas de homografia[1] (SCHREIBER, 2008). Na Figura 2.1, é possível observar de forma gráfica como é estabelecido o índice h, seguindo o conceito já abordado, segundo o qual no ponto de encontro do valor de citação e número de papers, tem-se o equilíbrio na forma de "Citações = Papers = Índice h". Hirsch destaca também que existe uma constante de proporcionalidade entre o número total de citações e o índice h de um pesquisador. Essa constante é definida como $N_{c,tot} = ah^2$, onde h é o índice h do pesquisador e o a varia num intervalo entre 3 e 5, valores obtidos empiricamente no seu estudo (HIRSCH, 2005).

[1] Palavras que têm grafia igual e significação diferente.
Ex.: publicações de autores com o mesmo nome e com atuação na mesma área.

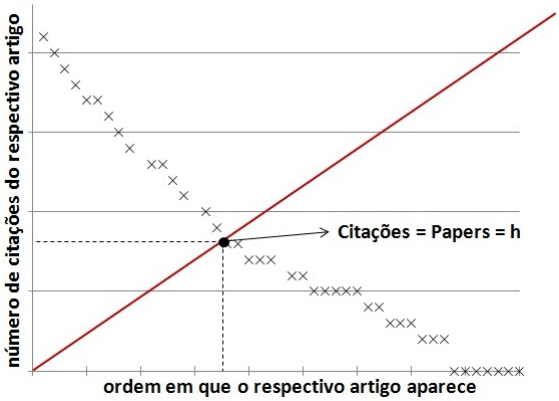

Figura 2.1: Gráfico representando o conceito do índice h

2.1.2 Índice h sucessivo de segunda ordem

Em trabalhos recentes, alguns autores propuseram variações desse índice para medir não apenas pesquisadores de forma isolada, mas também grupos de pesquisadores (ex.: pesquisadores de um mesmo instituto). Schubert (2007) criou um conceito chamado por ele de índice h sucessivo de segunda ordem, no qual tendo o índice h de cada pesquisador de um grupo é possível calcular um novo *h* (chamado h_2) que representa o índice h desse grupo. Prathap (2006) foi quem criou um esboço desse conceito para apoiar a classificação ou avaliação de institutos de pesquisa, como sendo: "h_2 = h, se uma instituição possuir h indivíduos, onde cada um possui um índice h que é pelo menos *h*". Egghe e Rao (2008) apresentaram três diferentes maneiras para calcular o índice h de grupos de pesquisadores: usando '*successives h-indices*', usando o índice *hp*, que se baseia em um *ranking* autor-publicação e usando o índice *hc* baseado em um *ranking* autor-citação.

2.1.3 Índice g

Outro índice, similar ao índice h, que permite mensurar a produtividade científica, foi criado por Egghe (2006). Dado um conjunto de artigos classificados por ordem decrescente pelo número de citações recebidas por eles, o índice g é o maior e único número, no qual os top g artigos receberam em média pelo menos g citações. Com isso, o índice g busca dar um peso maior aos artigos mais citados, para contornar uma desvantagem presente no índice h. Assim, no momento em que um artigo atinge a posição de estar entre os top g artigos, as citações subsequentes recebidas por ele não afetam mais sua posição. Incluindo essa correção, por consequência, o índice g deverá ser maior ou igual ao índice h correspondente em uma dada análise frente ao mesmo conjunto de dados.

De forma similar à idéia apresentada por Schubert (2007) do índice h sucessivo de segunda ordem, foi proposta uma métrica para avaliação de grupos, porém utilizando o índice g como base fundamental. A definição para esse conceito, consiste em: dado um conjunto de pesquisadores ordenados de forma decrescente de seus índices g, o índice g_1 é o maior e único número o qual os top-g_1 pesquisadores possuem, na média, pelo menos um índice g de g_1 (TOL, 2008).

2.1.4 Análise comparativa dos índices estudados

Esta seção apresentou métricas comumente empregadas na avaliação da produtividade de pesquisadores. Outras métricas mais simples são comparadas na Tabela 2.1. Foi dada

uma maior ênfase à métrica do índice h, bem como suas variantes descritas nas subseções, por serem cruciais na compreensão do restante desta dissertação. Optou-se pelo uso índice h ao invés do índice g, por ser um índice mais aceito na literatura e já possuir ferramentas como o ISI Web Of Science que o implementam em seus relatórios envolvendo citações de artigos.

Tabela 2.1: Comparação entre métricas de produtividade de pesquisadores

Métricas	Vantagens	Desvantagens
Número de artigos	Mede quantidade	Não mede o impacto das publicações
Número de citações	Mede impacto das publicações	Pode-se obter altos valores tendo poucos artigos muito citados com muitos co-autores
Número médio de citações	Permite medir pesquisadores de diferentes períodos (idades)	Recompensa autores com poucos artigos publicados e penaliza autores com muitos artigos
Número de artigos relevantes (número de artigos com y citações)	Elimina as desvantagens das 3 primeiras métricas	– *y* é um valor arbitrário que pode aleatoriamente favorecer ou prejudicar dos pesquisadores; – *y* precisa ser ajustado para diferentes tipos de pesquisadores mais antigos.
Número de citações dos artigos mais relevantes	Elimina as desvantagens das 3 primeiras métricas	– Retorna um conjunto de valores, tornando a análise e obtenção dos dados mais custosa; – Apresenta as mesmas desvantagens da métrica de artigos relevantes.

Fonte: adaptado de (ALTMANN; ABBASI; HWANG, 2009)

2.2 Distribuição de citações

Uma das formas de conexão entre publicações científicas ocorre através das citações, que possuem o papel de remeter o leitor às referências bibliográficas que levaram o autor de um trabalho a desenvolver e melhorar conceitos já publicados. Pesquisadores dividem os créditos de determinadas pesquisas (sobre assuntos correlacionados) citando uns aos outros por meio de artigos (HSU; HUANG, 2011). Com isso, o número de citações de um artigo publicado representa um importante papel como um indicador de impacto, refletindo a importância de uma determinada pesquisa desenvolvida. De forma sucinta, para Redner (1998), as distribuições de citações são: "o número de artigos que foram citados um total de x vezes, $N(x)$".

Alguns estudos abordando esse tema foram desenvolvidos, como em Laherrere e Sornette (1998), trabalhos nos quais um ajuste exponencial alongado (em inglês, *stretched exponential fit*) foi aplicado na modelagem da distribuição de citações baseado em processos multiplicativos. Já no trabalho de Redner (1998), foram analisadas as distribuições de citações presentes nas bases ISI Web of Science (WoS) e Physical Review D (PRD) em determinados intervalos de tempo, considerando 783.339 artigos publicados em 1981 com 6.716.198 citações de 1981 até 1997 no WoS e para o PRD 24.269 artigos publicados em junho 1997 com 351.872 citações obtidas entre 1975 até 1994, observando que para esse problema a mesma exponencial alongada é obtida apenas para um pequeno número

de citações, $x < x_c$, mas para um grande número de citações $x > x_c$ os dados corroboram um comportamento de lei de potência $N(x) \sim x^{-\alpha}$, com $\alpha \approx 3.0$.

Contudo, existem discussões sobre o trabalho de Redner (1998), mostrando que certas conclusões obtidas não são totalmente aceitáveis. Em seu estudo, ele concluiu que as distribuições de citações nas bases estudadas não são descritas por apenas uma função em todo o intervalo de citações, e sim, por duas funções com aspectos diferentes, uma lei de potência e uma exponencial alongada. Em contraponto a esse levantamento de Redner (1998), fazendo uso da abordagem generalizada proveniente da Mecânica Estatística para sistemas não extensivos, Tsallis e Albuquerque (2000) mostraram ser possível descrever as distribuições de citações em uma única função *N(x)* e para isso analisaram os mesmos dados usados por Redner. Outra tentativa de realizar ajuste em uma distribuição de citações, tanto para uma lei de potência como para uma exponencial alongada, foi aplicada por Lehmann, Lautrup e Jackson (2003) na base de dados SPIRES, que continha cerca de 281.717 artigos, e mostrou-se impossível fazer uma distinção entre os dois modelos.

É importante destacar nesse momento qual a finalidade no uso do ajuste, que é comum em várias áreas científicas e utilizado em muitos trabalhos envolvendo distribuições de dados. Trata-se de uma maneira de ilustrar a relação entre uma ou mais variáveis, através de uma equação que represente essa conexão entre as variáveis. Alguns dos propósitos e razões para o uso de ajuste incluem, basicamente:

- obter características fora do conjunto de dados, como: encontrar um ponto máximo ou ponto de inflexão (onde ocorre troca do sinal e muda uma curvatura);
- prover figuras/gráficos de melhor visualização, usando a linha (criada através do ajuste) como uma espécie de guia na leitura e compreensão do gráfico;
- descrever dados a partir de um princípio físico, o ajuste assim fornecerá os parâmetros na equação física correspondente;
- encontrar uma fórmula de pesquisa para uma dependência entre diferentes propriedades físicas.

A seguir, elencaremos alguns ajustes esperados para a distribuição de citações na literatura de bibliometria e cientometria[2].

2.2.1 Exponenciais alongadas (*stretched exponentials*) e comportamento lei de potência na distribuição de citações

Um tipo de função vista e citada na literatura, envolvendo distribuições de citações é a exponencial alongada (ou estendida) (LAHERRERE; SORNETTE, 1998). Esse tipo de função é uma extensão das funções exponenciais, mas com o adendo de um parâmetro adicional. Tem sido usada para descrever predominantemente fenômenos da natureza e economia que não seguem leis de potência (GUO et al., 2008). Possui a vantagem de ter menos parâmetros ajustáveis, isso pode não ser interessante em determinadas análises, mas de um modo geral as tornam mais econômicas (em termos de váriaveis) e objetivas. A fórmula genérica de uma exponencial alongada, também conhecida por função de Kohlrausch, é definida como:

[2]Cientometria é definida como o estudo da mensuração e quantificação do progresso científico, estando a pesquisa baseada em indicadores bibliométricos (SILVA; BIANCHI, 2001).

$$f(x) = c \cdot \exp[-(\frac{x}{x_0})^{\beta}] \tag{2.1}$$

onde $0 < \beta \leq 1$, x_0 sendo um parâmetro que representa dimensões temporais e c é uma constante de acoplamento.

A função de Kohlrausch é convenientemente utilizada como uma função de ajuste, mesmo na ausência de um modelo, dado que ela permite estimar desvios de forma simples para um comportamento 'canônico' de uma exponencial simples através do parâmetro β (BERBERAN-SANTOS; BODUNOV; VALEUR, 2008).

Laherrere e Sornette (1998) foram os primeiros a tratar distribuição de citações de pesquisadores. Eles ranquearam 1120 físicos de acordo com o seu número total de citações. O número de pesquisadores $N(x)$, como uma função de seus números de citações x, segue uma função exponencial alongada: $N(x) = N_0 \exp[-(x/x_0)^{\beta}]$, com $\beta \approx 0.3$. Nota-se que o número de citações n_c é uma variável inteira (*integer*), mas aqui foi considerado seu limite contínuo x.

De forma alternativa, Redner (1998) também tratou esse questionamento através de um enfoque um pouco diferente. Foi estudada a distribuição de probabilidade de citações de 783.339 artigos científicos, não de autores, publicados em 1981, com 6.716.198 de citações obtidas no período entre 1981 e 1997 na base de dados ISI (Institute of Scientific Information). O comportamento de exponencial alongada foi observado para um baixo número de citações $x < x_c$, com $x_c = 200$. Para um grande número citações $x > x_c$, o comportamento de lei de potência é dominante $N(x) \sim x^{-\alpha}$, com $\alpha \approx 3.0$.

2.2.2 Generalização de Tsallis

Os autores Tsallis e Albuquerque (2000) propuseram que o problema levantado em Redner (1998)[3] poderia ser melhor descrito através de uma distribuição generalizada, como segue:

$$\begin{aligned} N_q(x) &= N_0 \cdot \left(\exp_q(-\lambda x)\right)^q \\ &= N_0[1 + (q-1)\lambda x]^{q/(1-q)} \end{aligned} \tag{2.2}$$

onde $\exp_q(x) = [1 + (1-q)\lambda x]^{1/(1-q)}$ é a definição generalizada da função exponencial relacionada ao fator q, ou simplesmente q-exponencial, que surgiu no contexto da Mecânica Estatística não extensiva. Trata-se de uma teoria microscópica para a termodinâmica válida para situações nas quais a entropia de dois subsistemas não é simplesmente a soma das entropias de cada subsistema. Para o momento, deve ser observado que a definição generalizada do logaritmo (q-logarithm): $\ln_q(x) = (x^{1-q} - 1)(1-q)$ é naturalmente o inverso de $e_q(x)$, desde que:

$$\exp_q(\ln_q(x)) = \left[1 + (1-q)\frac{x^{1-q} - 1}{(1-q)}\right]^{1/(1-q)} = x \tag{2.3}$$

Todas as ferramentas matemáticas foram desenvolvidas para apoiar funções generalistas e muitas aplicações podem ser relacionadas a essa generalização apresentada[4]. Na Equação 2.2 o parâmetro λ é extraído, por exemplo, restringindo que o número médio de citações

[3] O problema estudado baseou-se na distribuição de citações dos artigos, diferentemente de citações de autores em artigos como em Laherrere e Sornette (1998).

[4] Ver (TSALLIS, 1999).

por artigo é uma constante, $\langle x \rangle = \sum_{x=0}^{\infty} x P_q(x) =$ constante, onde $P_q(x) = [1 + (q-1)\lambda x]^{q/(1-q)} / \sum_{y=0}^{\infty} [1 + (q-1)\lambda y]^{q/(1-q)}$.

O estudo feito por Tsallis e Albuquerque (2000) é a base da abordagem sobre distribuições de índice h e citações presentes nesta dissertação, na qual também é feito uso da estatística generalizada em conjunto com a dependência quadrática presente no trabalho de Hirsch (2005).

3 EXPLICANDO NOSSOS DADOS E SUA AQUISIÇÃO

A aquisição de dados para o desenvolvimento de experimentos no trabalho ocorreu através do uso de duas bases de dados: ISI Web of Science (WoS) e Google Scholar (GS). O uso do WoS foi focado em duas áreas: Física e Biologia, áreas que publicam tradicionalmente em revistas científicas indexadas, na sua maioria, nessa base de dados. Quanto ao uso do GS, este se deu pela necessidade de obtenção de dados de conferências científicas da área da Ciência da Computação, pouco presentes no WoS, porém com bastante visibilidade através do GS.

Todas as consultas efetuadas passaram por uma etapa de validação (casamento de instâncias) através da Plataforma Lattes, para garantir que os documentos encontrados eram dos autores pesquisados e não de homônimos destes. A Plataforma Lattes realmente é uma base de currículos brasileira bastante confiável e com suporte governamental. É inclusive reconhecida no exterior como relata Lane (2010): “A experiência brasileira com a base de dados Lattes é um poderoso exemplo de boas práticas. Ela provê dados de alta qualidade em 1,6 milhões de pesquisadores e cerca de 4.000 instituições", a autora ainda conclui que esta é uma das bases de dados de pesquisadores mais limpa (baixo ruído nos dados) do mundo.

3.1 ISI Web of Science

Web Of Science[1] faz parte da organização Thomson Scientific e tem sido uma das fontes de referências acadêmicas mais utilizadas por possuir mais de 37 milhões de registros de publicações, tendo como um de seus serviços mais conceituados a divulgação anual do fator de impacto (FI), através do *Journal Citation Report* (JCR), que avalia a importância e influência de revistas cinetíficas. O FI tem sido bastante criticado através dos anos, porém ainda é um dos índices de avaliação científica mais utilizado (MAIER, 2006). O WoS possui ferramentas para pesquisa por tópicos, autores, universidades, nome da publicação, nome do veículo de publicação, entre outras opções mais avançadas de pesquisa. É possível aplicar diversos filtros, a fim de evitar inconsistências na consulta aos dados e ainda possui uma opção ’Create Citation Report’ (Gerar relatório de citações), por meio da qual é feita a sumarização das citações a partir da lista de resultados obtida através da submissão de uma consulta e do refinamento por meio do uso de filtros. Nesse relatório, aparecem todas as publicações envolvidas e são automaticamente computadas informações como: índice h, citações médias por artigo e soma das vezes que um artigo foi citado. O acesso a todos esses serviços é restrito a instituições que possuem cadastramento nesse sistema e por consequência pagam por seu acesso aos dados do WoS.

[1]Web Of Science: Disponível em: <http://science.thomsonreuters.com/pt/produtos/wos/>.

3.2 Google Scholar

Google Scholar[2] foi desenvolvido pela Google Inc. e disponibiliza acesso a um vasto catálogo de links que apontam para artigos, apresentações, capítulos de livros, teses e resumos de inúmeros sítios públicos presentes na *Web*. Sendo essencialmente um motor de busca na *Web*, o GS tem como objetivo atingir a maior audiência e usuários possíveis, permitindo uma busca rápida e avançada. No modo avançado, a busca pode ser limitada por: palavras do título de um documento, autores, fontes, data de publicação e áreas de atuação. Cada informação de artigo recuperada é mostrada no formato: título, autores e fonte. Em cada artigo recuperado também é possível visualizar o número de citações recebidas e também ver quais documentos são esses que realizaram as citações. O Google Scholar não possui nenhuma ferramenta automática que possibilita filtragens mais específicas, agrupamento de artigos, cálculo de índices de avaliação, portanto, faz-se necessário o uso de ferramentas externas que dêem suporte à consultas no GS. Com a ferramenta Harzing's Publish And Perish[3] é possível buscar dados de citações no GS e então analisá-los para gerar as métricas baseadas em citações. Como resultado, a ferramenta provê informações como: número total de artigos, número total de citações, número médio de citações por artigo, número médio de citações por autor, número médio de artigos por autor, índice h de Hirsch, índice g de Egghe, *contemporary h-index*, *individual h-index* , *age-weighted citation rate*. Contudo, a ferramenta automaticamente não realiza limpeza de dados, como desambiguação de autores e remoção de auto-citações, tampouco consulta outras bases para garantir que os dados obtidos no GS estão realmente corretos.

Um novo serviço disponibilizado em 2011 pelo Google na Plataforma Scholar é o Google Scholar Citations[4]. Nesse serviço, o usuário pode criar um perfil através de sua conta Google e verificar quem está citando seus trabalhos, gerar gráficos comparando o número de citações recebidas em um dado período de tempo e computar os dados através de métricas como: índice h, número de artigos com no mínimo 10 citações recebidas e o número total de citações. Também é possível realizar consultas sobre outros pesquisadores e tornar público seu próprio perfil, dessa forma as consultas ao Google Scholar retornarão dados mais precisos sobre as publicações relacionadas a este perfil. Todavia, esse serviço não pode ser utilizado nesta dissertação por ter sido lançado em novembro de 2011, período no qual a coleta de dados já havia sido finalizada.

3.3 Métodos para obtenção dos dados

A coleta de dados no WoS foi feita utilizando a interface disponível no site, que é restrita a assinantes do serviço, como as universidades federais brasileiras. Existem tantas variáveis a serem consideradas no que diz respeito à determinação se as informações de um artigo realmente são condizentes com o autor relacionado, que a coleta foi realizada de forma manual. O maior problema no WoS são as homografias em função das quais, por exemplo, os nomes 'Renato da Silva' e 'Ricardo da Silva' são representados como 'da Silva, R', além de haver pesquisadores em um mesmo campo de pesquisa com nome de igual grafia. Como o objetivo desta dissertação não é desenvolver métodos para tratar esses problemas, nem mesmo automatizá-los, a coleta foi bastante custosa. A validação dos dados foi feita consultando manualmente a Plataforma Lattes (onde os autores man-

[2]Google Scholar: Disponível em: <http://scholar.google.com/>

[3]Harzing's Publish And Perish: Disponível em: <http://www.harzing.com/pop.htm>

[4]Google Scholar Citations. Disponível em: <http://scholar.google.com/citations>

têm atualizados os dados de suas publicações) e cruzando os dados com os encontrados no WoS, para obter a exata ou casamento mais próximo das instâncias entre as duas bases de dados (sendo observado o título do trabalho e o número total de artigos presentes na Plataforma Lattes em comparção aos dados obtidos nas consultas ao WoS para cada pesquisador). Assim, esse processo foi efetuado para todos os pesquisadores presentes nos programas de pós-graduação em física e biologia consultados. De forma mais precisa, foram considerados 1.203 pesquisadores brasileiros, vinculados a universidades brasileiras, divididos em 19 programas relevantes em física (600 pesquisadores) e em 26 programas relevantes em biologia (603 pesquisadores). Esse método de validação foi uma excelente forma de filtrar e obter dados referentes à pesquisa efetuada por pesquisadores brasileiros.

Em muitas áreas da ciência da computação, os autores são classificados levando em conta não apenas trabalhos publicados em periódicos, mas também publicações em importantes conferências. Existem inúmeras conferências de grande relevância na área e, portanto, para fins de simplificação na coleta de dados, foi selecionada a área de Engenharia de Software como delimitador da pesquisa. Foram computados os índices h e o número total de citações de membros de comitês de programas de 7 diferentes conferências, totalizando 600 pesquisadores. Para essa coleta, foi usado o GS (através do software Harzing's Publish And Perish), por este apresentar uma maior abrangência de dados de anais de conferências da área de ciência da computação. Para cada pesquisa efetuada ao GS, foi realizada uma limpeza dos dados para melhorar a precisão, já que o software permite a seleção de quais documentos retornados devem ser utilizados no cálculo do índices presentes na ferramenta.

É importante ressaltar que esse tipo de distinção na coleta de dados (usando bases diferentes) vêm ao encontro do objetivo de encontrar uma fórmula universal para índice h, fazendo-se necessária a verificação de bases mais adequadas às que não são baseadas estritamente em publicações em periódicos. Áreas de ciências naturais como física, biologia, química e medicina concentram suas publicações de artigos em veículos como períodicos, principalmente pela grande visibilidade alcançada nesse meio. Outras áreas como história e filosofia concentram suas publicações em livros e áreas mais recentes como ciência da computação costumam dar mais ênfase na publicação e apresentação de trabalhos em conferências científicas que permitem uma maior interação entre comunidades de pesquisadores (ALVARENGA; NETO, 2007).

O número de pesquisadores analisados através do GS não foi maior por limitações no número de consultas permitidas ao sistema do Google, o qual efetua um bloqueio após repetidas tentativas.

4 DESCRIÇÃO DOS MODELOS

Partindo dos objetivos definidos anteriormente nesta dissertação, o presente capítulo apresenta a descrição das técnicas utilizadas para atingí-los. A primeira seção aborda as adaptações matemáticas necessárias para o uso dos conceitos de Tsallis e Albuquerque (2000) e Laherrere e Sornette (1998) no contexto das distribuições de índice h, bem como da integração com a dependência quadrática de Hirsch (2005). Foram construídos estimadores para os parâmetros presentes nas equações de modo que o cálculo fosse facilitado, obtendo como resultado uma fórmula para distribuição de índice h a partir da estatística generalizada e também da abordagem de *stretched exponential*. Na segunda seção, temos a descrição dos trabalhos analisados e que embasaram a proposta para *ranking* de pesquisadores desenvolvida nesta dissertação. São apresentadas as definições formais dos conceitos envolvidos, além do algoritmo utilizado para obtenção dos chamados h_2 e h_2 relativo. Por fim, temos a formalização da métrica $s-index$ baseada no índice h sucessivo de segunda ordem.

4.1 Formulação para distribuição de índice h para grupos de pessoas

A caracterização das distribuições envolvendo dados bibliométricos trata da tentativa de determinar de forma científica as características intrínsecas existentes em diferentes campos de pesquisa. Dessa forma, por meio de equações e métodos é possível gerar gráficos que mostrem esse comportamento, além de provar que áreas diferentes de pesquisa seguem uma universalidade (RADICCHI; FORTUNATO; CASTELLANO, 2008). Os estudos de Tsallis e Albuquerque (2000) e Laherrere e Sornette (1998) foram fundamentais para trazer à comunidade científica conceitos de estatística generalizada e exponencial alongada aplicada à bibliometria, respectivamente, mostrando possibilidades de análise bastante abrangentes no que diz respeito à distribuição de citação. Tendo em vista os resultados obtidos em Tsallis e Albuquerque (2000) para distribuição de citações utilizando uma única fórmula e os resultados apresentados em Laherrere e Sornette (1998), foram definidas formulações baseadas nesses estudos para dar suporte à análise de distribuições de índice h.

O primeiro método a ser demostrado faz uso da hipótese descrita em Tsallis e Albuquerque (2000) acerca das distribuições de citações. Nesse sentido, buscamos adaptar a fórmula generalista apresentada na Equação 2.2 para o contexto do trabalho que foca na distribuição de índice h e uso do índice h na caracterização dos pesquisadores em áreas da ciência. As etapas envolvidas na obtenção de uma fórmula para distribuição de índice h por meio da estatística generalizada são as seguintes:

1. Foi elaborada uma primeira modificação, por meio da qual se descrevem as fór-

mulas em um limite contínuo da distribuição normalizada de citações usando a abordagem generalizada. A fórmula usada nessa re-escrita é:

$$\begin{aligned} P_q(x) &= \frac{[1+(q-1)\lambda x]^{q/(1-q)}}{\int_0^\infty [1+(q-1)\lambda y]^{q/(1-q)} dy} \\ &= \lambda[1+(q-1)\lambda x]^{q/(1-q)} \text{ para } 0 < x < \infty \end{aligned} \tag{4.1}$$

Para estimar o λ, fazemos uso do Método dos Momentos (ver Apêndice I) e calculamos, de forma analítica, o primeiro momento dessa distribuição como:

$$\begin{aligned} \langle x \rangle &= \int_0^\infty x P_q(x) dx \\ &= \lambda \int_0^\infty x[1+(q-1)\lambda x]^{q/(1-q)} dx \\ &= \frac{1}{(2-q)\lambda} + \frac{1}{\lambda} \lim_{x\to\infty} \frac{x\lambda+1}{(q-2)(qx\lambda-x\lambda+1)^{\frac{1}{q-1}}} \\ &= \frac{1}{(2-q)\lambda} \text{ se } 1 < q < 2 \end{aligned} \tag{4.2}$$

2. Coletando uma amostra de citações $x_1, x_2, ...x_n$ de n autores, temos uma estimativa para $\langle x \rangle$, que é dada naturalmente pela simples média aritmética $\widehat{x} = (x_1 + x_2 + ... + x_n)/n$. Então, estimamos $\widehat{\lambda} = \frac{1}{(2-q)\widehat{x}}$ e uma expressão híbrida para densidade de citações, considerando que $\widehat{x}$ é um estimador para $\langle x \rangle$, temos:

$$\widehat{P}q(x) = \frac{1}{(2-q)\widehat{x}}[1 + \frac{(q-1)}{(2-q)\widehat{x}}x]^{q/(1-q)} \tag{4.3}$$

3. Dentro do propósito do nosso trabalho, fazemos a relação entre o estudo de Tsallis e Albuquerque (2000) e Hirsch (2005), através da hipótese fundamental de Hirsch, que considera uma dependência quadrática de *h* para o número de citações de um autor x, por exemplo, $x = ch^2$, onde c é uma constante que pode ser determinada através de um ajuste linear adequado, já que nossos dados amostrais suprem o número de citações (x_i) e o índice h correspondente do autor (h_i) com $i = 1, ..., n$, obtemos o estimador para o parâmetro $\widehat{c}$:

$$\widehat{c} = \frac{\sum_{i=1}^{n} h_i^2 x_i - \frac{1}{n}\sum_{i=1}^{n} h_i^2 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} h_i^4 - \frac{1}{n}\left(\sum_{i=1}^{n} h_i^2\right)^2} \tag{4.4}$$

por consequência, a distribuição de índice h de um grupos de autores ou pesquisadores, é dado pela equação:

$$\begin{aligned} H_q(h) &= \widehat{\lambda}[1+(q-1)\widehat{\lambda}(\widehat{c}h^2+\widehat{b})]^{q/(1-q)} \frac{dx}{dh} \\ &= \frac{2\widehat{\lambda}\widehat{c}h}{[1+(q-1)\widehat{\lambda}(\widehat{c}h^2)]^{q/(q-1)}} \\ &= \frac{2\widehat{c}h}{(2-q)\widehat{x}[1-\frac{(1-q)\widehat{c}}{(2-q)\widehat{x}}h^2]^{q/(q-1)}} \end{aligned} \tag{4.5}$$

onde é possível verificar naturalmente a normalização $\int_0^\infty \frac{2\widehat{\lambda}\widehat{c}h}{[1+(q-1)\widehat{\lambda}\widehat{c}h^2]^{q/(q-1)}}dh = 1$.

4. Tendo os parâmetros $\langle x \rangle$ e c sido estimados previamente por $\widehat{x}$ and $\widehat{c}$, o único parâmetro restante para ser ajustado é q.

O segundo método desenvolvido tem sua base teórica no estudo de Laherrere e Sornette (1998), onde foram discutidas distribuições de citações que respeitam uma exponencial alongada. Dessa forma, é importante traçar esse paralelo entre as contribuições de Tsallis e Albuquerque (2000) e Laherrere e Sornette (1998), pois nos dois estudos são utilizados dados de citações e cada um possui uma peculiaridade relacionada à forma de análise dos dados. Com isso, desenvolvemos também um método baseado em exponenciais alongadas para suporte à distribuição de índice h.

Para tanto, a forma normalizada da equação, que faz uso dos conceitos de exponencial alongada vistos em Laherrere e Sornette (1998), é apresentada como:

$$P_\beta(x) = \frac{\beta}{x_0\Gamma(1/\beta)} e^{-(\frac{x}{x_0})^\beta} \tag{4.6}$$

deduzida a partir de

$$\begin{aligned} \int_0^\infty e^{-(\frac{x}{x_0})^\beta} dx &= x_0 \int_0^\infty e^{-z^\beta} dz \\ &= \frac{x_0}{\beta} \int_0^\infty e^{-t} t^{1/\beta - 1} dt = \frac{x_0}{\beta}\Gamma(1/\beta) \end{aligned} \tag{4.7}$$

Dessa forma, podemos estimar x_0 de maneira similar a feita anteriormente, estimamos $\langle x \rangle$ por $\widehat{x}$:

$$\begin{aligned} \langle x \rangle &= \int_0^\infty P_\beta(x) dx \\ &= \frac{x_0}{\Gamma(1/\beta)}\Gamma(\tfrac{2}{\beta}) = \widehat{x} \end{aligned} \tag{4.8}$$

resultando em $\widehat{x}_0 = \Gamma(1/\beta)\widehat{x}/\Gamma(\frac{2}{\beta})$. Através desse resultado, tomando como base a relação $x = ch^2$ e seguindo os mesmos passos anteriormente utilizados na abordagem generalizada, temos uma fórmula para distribuição de índice h, através da abordagem de *stretched exponential*:

$$H_\beta(h) = \frac{2\widehat{c}\beta\Gamma(\frac{2}{\beta})}{\widehat{x}\Gamma(\frac{1}{\beta})^2} h \exp\left[-(\frac{\widehat{c}\Gamma(\frac{2}{\beta})h^2}{\Gamma(\frac{1}{\beta})\widehat{x}})^\beta\right] \tag{4.9}$$

4.2 Proposta para *ranking* de pesquisadores baseado no índice h sucessivo

Outro objetivo definido nesta dissertação trata de uma proposta para a classificação de pesquisadores com base no seu índice h sucessivo de segunda ordem de Schubert (2007), que também foi modelado e estendido de forma teórica em Egghe (2008). Um método para classificar a qualidade de um grupo de pesquisadores a partir de seus índices h foi iniciado e apresentado em Silva et al. (2010). Os autores consideraram a magnitude do índice h e o nível de igualdade desse índice em toda uma população de comitês de programa de eventos da área da Ciência da Computação. Para tanto, a primeira definição importante que deve ser apresentada, trata do índice h sucessivo de segunda ordem (ou simplesmeste h_2) utilizado em Silva et al. (2010), bem como na presente dissertação:

Definição para o h_2

> Um grupo tem índice h_2 se ele possui h_2 pesquisadores com índice h igual a h_2 mas não possui mais de h_2 pesquisadores com índice h maior que $h_2 + 1$.

Em Silva et al. (2010), os autores constataram um fato comum encontrado em conferências nas quais o número de membros do comitê de programa varia em cada conferência, levando a definição de um índice chamado de h_2 relativo, baseado no grupo com tamanho menor. O índice gerado pelos autores assemelha-se com a definição do índice h sucessivo de segunda ordem proposto por Schubert (2007), porém a idéia de representar essa relação com o tamanho dos grupos foi a evolução apresentada. Para o restante desta dissertação, trataremos o índice h sucessivo de segunda ordem apenas pela denominação h_2 para facilitar a leitura.

Os conceitos apresentados em Silva et al. (2010) e Schubert (2007), levaram ao uso de técnicas de amostragem (*sampling*), nas quais o h_2 de um grupo deve ser repesado pelo h_2 relativo (o h_2 que o maior grupo deve ter se o seu tamanho for igual ao grupo pequeno). A seguir, mostramos a definição para o h_2 relativo amostrado:

Definição de h_2 relativo amostrado

> Definimos um grupo A de tamanho n_A, isto é com n_A integrantes, que possui índice h_2 relativo amostrado $h_2(A, B)$ à um dado grupo B com $n_B < n_A$ integrantes, sob um número de amostra n_{sample}, de acordo com
>
> $$h_2(A, B) = \frac{1}{n_{sample}} \sum_{i=1}^{n_{sample}} h_2^{(i)}(n_B|n_A) \tag{4.10}$$
>
> onde $h_2^{(i)}(n_B|n_A)$ denota o índice h_2 de uma particular amostra de n_B pesquisadores coletados sobre n_A pesquisadores do grupo A. Este reflete o potencial do grupo A considerando se ele tivesse a capacidade máxima de pesquisadores do grupo B. Obviamente $h_2(A, B)$ coincide com o próprio h_2 do grupo A se os grupos são iguais (propriedade naturalmente esperada) já que todas as amostras terão tamanho $n_B = n_A$, que é o número total de pesquisadores.

Um fato motivador para o uso do h_2 relativo amostrado encontra-se na impossibilidade de uma instituição com produção moderada atingir o índice h de uma grande instituição, mesmo se a qualidade de suas publicações forem similares ou até melhores devido a sua produção total ser até menor que h (SYPSA; HATZAKIS, 2009). A comparação direta entre o índice h de instituições de tamanhos diferentes é discutível e torna difícil a compreensão da produção científica, pois um grande índice h, como abordado, pode vir de uma grande produção mediana, não comparável a uma média produção de grande impacto, distorcendo o valor obtido pelo índice h.

O Algoritmo 1 foi desenvolvido no trabalho de Silva et al. (2010) para obtenção dos valores de h_2 para um grupo ou o h_2 que faz a relação entre um grupo maior e um menor. Esse algoritmo tem como entrada 4 parâmetros: n, que representa o total de pesquisadores do maior grupo; $nmin$, que representa o total de pesquisadores do grupo menor; $nsample$, que é o número de vezes que a amostra será embaralhada; vetor $hIndexGrupoMaior[]$, que contém o índice h dos pesquisadores do maior grupo. A saída pode ser o h_2 absoluto quando é medido um grupo em relação a ele mesmo ou o h_2 relativo, quando é medido

um grupo de menor tamanho em relação a um maior tamanho. O algoritmo começa com a alocação do vetor $arrayMin[nmin]$ (linha 2). Em seguida, há um laço que irá se repetir n_{sample} vezes e irá executar as tarefas de embaralhar o vetor $hIndexGrupoMaior$ e adicionar ao vetor $arrayMin[nmin]$ os valores de $hIndexGrupoMaior$ até o seu limite alocado. Após, ainda no laço, irá ordenar de forma ascendente o vetor hIndexGrupoMenor e inicializar uma variável local $ihgroup$ para o valor 1. A seguir, existe uma estrutura $While$ (linha 11-13), similar ao cálculo do índice h de Hirsch, porém levando em conta os índice h ao invés das citações. Na linha 14, a variável $ihgroupsample$ armazena os cálculos relativos a uma fatia (de n_{sample}), ou seja, quanto maior o n_{sample}, maior a chance de não repetição da ordem dos dados nos vetores. Por fim, é feita a média (linha 16) do acumulado na variável $ihgroupsample$ dividido pelo número de amostras geradas (n_{sample}), resultando no h_2 absoluto ou h_2 relativo. É importante salientar que ao parâmetro $nsample$ na execução do algoritmo foi utilizado o valor 200 de forma que a amostra de dados ficasse bastante balanceada.

Algorithm 1: Cálculo do h_2 e h_2 relativo

Input: $n, nmin, nsample, hIndexGrupoMaior[]$
Output: h_2

```
1  begin
2      arrayMin[nmin] ⟵ 0;
3      ihgroupsample ⟵ 0;
4      for (i ← 0; i < nsample; i++) do
5          shuffle(hIndexGrupoMaior);
6          for (j ← 0; j < nmin; j++) do
7              hIndexGrupoMenor[j] ⟵ hIndexGrupoMaior[j];
8          end for
9          sort_ascending(hIndexGrupoMenor);
10         ihgroup ⟵ 1;
11         while nmin − (ihgroup + 1) > y[ihgroup − 1] do
12             ihgroup++;
13         end while
14         ihgroupsample ⟵ ihgroupsample + (nmin − (ihgroup + 1));
15     end for
16     h2 ⟵ ihgroupsample/nsample;
17 end
```

Este algoritmo proposto em Silva et al. (2010) é utilizado em nossa proposta de duas formas: (i) na obtenção do h_2 absoluto e (ii) na obtenção do h_2 relativo. No trabalho original, os autores afirmam que uma classificação para conferências (grupos) pode ser estabelecida baseada no h_2 e no coeficiente de Gini. É nesse ponto que aparece a contribuição do nosso trabalho onde: (i) grupos grandes são balanceados (repesados) de forma a manter um bom número de bons pesquisadores; (ii) grupos menores são balanceados por conta de seu 'índice h potencial', pois o tamanho do grupo é levado em conta na análise. Na proposta original de Silva et al. (2010), esse balanceamento era feito através do coeficiente de Gini, aqui utilizamos outras formulações as quais serão detalhadas a seguir.

Denominamos nesta dissertação de índice s (ou $s - index$) a métrica que traz como resultado a possibilidade de comparação entre grupos (no contexto desta dissertação: programas de pós-graduação) com diferentes tamanhos, a fim de montar uma classificação que possa ser usada em conjunto com outras avaliações da qualidade dos pesquisadores.

Definição do $s - index$

O $s - index$ de um grupo de pesquisadores $k = 1, ..., m$, considerando um universo com m diferentes grupos de pesquisadores é dado por

$$s - index = \frac{1}{n_{min}^{(k)}} \sum_{j, n_j < n_k} h_2(k, j) \tag{4.11}$$

onde $n_{min}^{(k)} = \#\text{grupos } j = 1, ..., m | n_j < n_k$. Caso o universo de m grupos estiver classificado de acordo com seu tamanho, o valor para $n_{min}^{(k)}$ é simplesmente o *rank* do grupo k.

O objetivo desta métrica ($s - index$) é sua utilizacão para qualificação de grupos de pesquisadores com diferentes números de membros através do índice h, levando em consideração aspectos como: (i) a homogeneidade presente nos grupos analisados; (ii) a magnitude, através de repesagens a fim de equiparar o tamanho dos grupos em função dos melhores índice h de seus pesquisadores; (iii) manutenção da escala do índice h original, não havendo necessidade de determinar os intervalos de valores obtidos e definir uma nova escala para analisá-los.

Neste capítulo foram apresentados os modelos desenvolvidos nessa dissertação, onde o primeiro trata de duas abordagem para definir a distribuição de índice h, através da teoria da generalização de Tsallis e Albuquerque (2000) e através de *stretched exponential* como em Laherrere e Sornette (1998), partindo das respectivas distribuições de citações. O segundo modelo trata da definição de uma nova métrica baseada no índice h para classificação e comparação de grupos de pesquisadores, que podem possuir tamanhos diferentes, e não obterem boas caracterizações por meio de medidas tradicionais como: índice h absoluto ou média dos índices h.

5 RESULTADOS

Neste capítulo, apresentaremos os resultados obtidos e revisados sobre as formulações desenvolvidas na dissertação. Na seção 5.1, temos a aplicação das formulações geradas e de que forma as abordagens são efetivas na caracterização das distribuições de índice h, comparando o uso de estatística generalizada e exponencial alongada (*stretched exponential*) frente aos conjuntos de dados estudados. Na seção 5.2 é apresentada a metodologia utilizada na validação da métrica $s - index$, bem como os resultados obtidos com sua aplicação na classificação de grupos de pesquisadores, comparando a escores gerados por agências de fomento à pesquisa.

5.1 Formulações para distribuição de índice h

Esta seção apresenta os resultados obtidos através do estudo envolvendo estatística generalizada e *stretched exponential*, com dados de pesquisadores de áreas distintas da ciência (física, biologia, ciência da computação) para distribuição de índice h e citações. Os resultados apresentados para citações foram uma replicação dos métodos propostos em Tsallis e Albuquerque (2000) e Laherrere e Sornette (1998), respectivamente, para os dados experimentais obtidos para esta dissertação (dados sobre pesquisadores nas áreas de física, biologia e ciência da computação).

Primeiramente, foram gerados gráficos apropriados do número de citações em função do quadrado do índice h de cada autor para cada grupo estudado, para assim estimar $\widehat{c}$, e realizar cálculos para os dois grupos estudados. Foram criados dois grupos de dados para análise: Grupo 1, representando Conferências de Ciência da Computação (obtidos através do Harzing/Google Scholar); e Grupo 2, representando a união dos dados obtidos, através do ISI WoS/Lattes-CNPq, de pesquisadores em programas de pós-graduação em física e biologia. Na Tabela 5.1 são apresentados os *slopes*[1] estimados, além do número médio de citações para cada um dos grupos considerando suas respectivas bases de dados diferentes.

Tabela 5.1: Slopes estimados para os grupos estudados

Grupos	$\widehat{c}$	$\widehat{x}$
Conferências de Ciência da Computação (Harzing/Google Scholar)	5.71(11)	936(76)
Pesquisadores de Pós-graduações em Física e Biologia (ISI WoS/Lattes-CNPq)	3.77(5)	473(23)

Os gráficos da Figura 5.1 exibem o comportamento do número de citações em função

[1] *Slope*: em português, é o coeficiente angular da reta sendo um dos parâmetros no cálculo da regressão linear

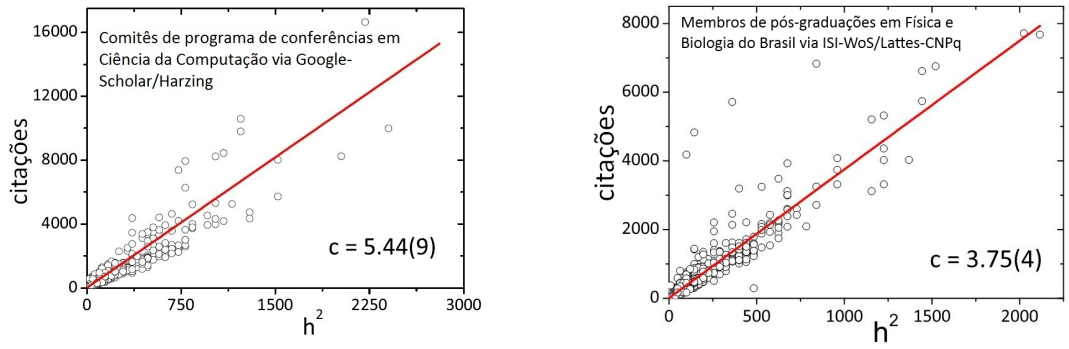

Figura 5.1: Número de citações em função do h^2 (Grupo 1 à esquerda e Grupo 2 à direita)

do h^2 de diferentes autores em cada grupo estudado. A variável c, da relação proposta por Hirsch descrita na forma algébrica na Equação 4.4 apresentada na Seção 4.1, foi obtida de forma numérica e é mostrada junto aos gráficos. Esse primeiro teste foi efetuado para saber se os nossos dados seguiam a relação descrita por Hirsch entre o número de citações e o índice h dos pesquisadores.

Estas estimativas refletem as diferenças entre áreas de pesquisa diferentes, mas principalmente entre as áreas em questão, pesquisadores de Física e Biologia, e comitês de programa de conferência da Ciência da Computação. Foram calculados parâmetros para distribuição de citação para ambos os grupos estudados, considerando diferentes abordagem anteriores da literatura, como Tsallis e Albuquerque (2000) e Laherrere e Sornette (1998), usando novos métodos, reconstruindo alguns pontos para tornar possível posteriormente obter distribuições de índice h através dessas diferentes abordagens e mostrar que estas se ajustam universalmente à distribuição quando testada nos diferentes grupos usados no trabalho.

Nesse ponto, fazemos uso do Método dos Momentos (ver Apêndice I), que nos permite estimar de forma precisa o β da Equação 4.6. Para tanto, devemos calcular os momentos analíticos dessa distribuição, dados por:

$$\begin{aligned} \left\langle x^k \right\rangle &= \frac{\beta}{x_0 \Gamma(1/\beta)} \int_0^\infty x^k e^{-(\frac{x}{x_0})^\beta} dx \\ &= \frac{x_0^k}{\Gamma(1/\beta)} \int_0^\infty e^{-t} t^{\frac{k+1}{\beta}-1} dx \quad = \quad \frac{x_0^k}{\Gamma(1/\beta)} \Gamma(\tfrac{k+1}{\beta}) \end{aligned} \tag{5.1}$$

sendo os momentos experimentais das citações calculados por $\overline{x^k} = (x_1^k + x_2^k + ... + x_n^k)$. Esse método busca comparar $\left\langle x^k \right\rangle$ e $\overline{x^k}$ a vários valores de k (não apenas inteiros) e observar o melhor β que corresponde a melhor concordância entre $\left\langle x^k \right\rangle$ e $\overline{x^k}$.

Neste ponto, observamos a dependência de $\left\langle x^k \right\rangle$ com o momento x_0, que teríamos de usar para obter o estimador necessário. Todavia, contornamos essa dificuldade usando a razão $\Phi_k^{(\beta)}$, não dependendo de x_0.

$$\Phi_k^{(\beta)} = \frac{\left\langle x^k \right\rangle}{\left\langle x \right\rangle^k} = \frac{\Gamma \frac{1}{\beta}^{k-1} \Gamma(\frac{k+1}{\beta})}{\Gamma(\frac{2}{\beta})^k} \tag{5.2}$$

Para fins de comparação com os dados experimentais, definimos também a razão

$$\Phi_k^{(\text{exp})} = \overline{x^k} / \overline{x}^k = (\sum_{i=1}^n x_i^k) / (\sum_{i=1}^n x_i)^k \tag{5.3}$$

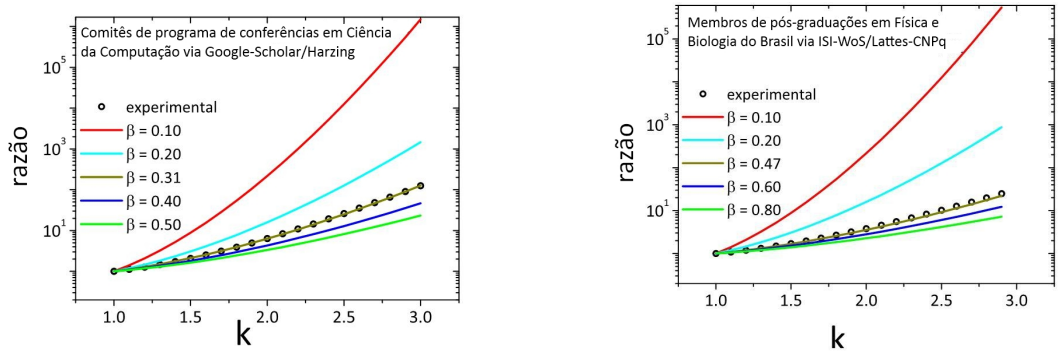

Figura 5.2: Momentos teóricos e experimentais, usando Stretched Exponential, para Grupo 1 e Grupo 2, respectivamente

Assim, demonstramos na Figura 5.2 os momentos teóricos e experimentais para os dois casos estudados, baseados em citações distribuídas na forma exponencial alongada (*stretched exponential*). Nesse mesmo gráfico, são mostrados momentos teóricos para vários valores de β, junto com os momentos experimentais. Nesse ponto devemos observar que o objetivo dessa comparação entre os momentos está na busca pelo valor de β que faz o melhor casamento entre os momentos teóricos e experimentais.

Uma simplificação numérica de $\int(\phi_k^{(\beta)} - \phi_k^{(\exp)})^2 dk$, usando $\delta k = 0.01$ resulta em um $\beta = 0.47$ para pós-graduações em física e biologia obtidas através do WoS e da Plataforma Lattes e um $\beta = 0.31$ para comitês de programa de conferências de ciência da computação extraídos através do aplicativo Harzing's Publish Or Perish (com dados do Google Scholar). É importante salientar que o valor de β para o Grupo 2 (dados de conferências) corrobora as afirmações presentes no trabalho de Laherrere e Sornette (1998) quanto ao β ser aproximadamente 0.31. Esse valor não era esperado pelas diferenças entre os conjuntos de dados usados nesta dissertação e os dados presentes no trabalho de Laherrere e Sornette, no qual neste último foram usadas citações dos 1.120 autores mais citados obtidas por meio do ISI-JCR, durante o período de 1981-1997, enquanto nossos dados contemplam todas as citações da vida científica dos pesquisadores selecionados para o estudo. O resultado obtido para o Grupo 1, mesmo sendo oriundo de uma mesma base (ISI), difere do encontrado em Laherrere e Sornette (1998). Todavia, o resultado para este grupo é exponencialmente próximo para as citações de artigos do periódico Physical Review D (que apresenta β de aproximadamente 0.39) e similar às citações de artigos do ISI ($\beta \approx 0.44$) obtidos por Redner num limite baixo de citações (sendo $x < 500$).

Foram construídos *Zipf plots* para testar os ajustes gerados para cada um dos nossos propósitos. A idéia do *Zipf plot* é ordenar as citações de todos os autores do conjunto de dados estudados na forma $x_1 \geq x_2 \geq x_3 ... \geq x_n$ (ver Apêndice I). Com essa idéia, chegamos à expressão

$$\zeta_j = \int_{x_j}^{\infty} P_\beta(x)dx = 1 - \frac{j}{n} \tag{5.4}$$

onde j é o *rank* de citações x_j. O cálculo de ζ_j, é portanto:

$$\begin{aligned}\zeta_j &= \frac{\beta}{\widehat{x}}\frac{\Gamma(2/\beta)}{\Gamma(1/\beta)}\int_{x_j}^{\infty}\exp\left[\frac{-\Gamma(2/\beta)^{\beta}}{\Gamma(1/\beta)^{\beta}\widehat{x}^{\beta}}x^{\beta}\right] \\ &= \frac{1}{\Gamma(1/\beta)}\int_{\frac{\Gamma(2/\beta)^{\beta}x_j^{\beta}}{\Gamma(1/\beta)^{\beta}\widehat{x}^{\beta}}}^{\infty} z^{1/\beta-1}e^{-z}dz \\ &= \frac{\Gamma(1/\beta,\frac{\Gamma(2/\beta)^{\beta}x_j^{\beta}}{\Gamma(1/\beta)^{\beta}\widehat{x}^{\beta}})}{\Gamma(1/\beta)}\end{aligned} \tag{5.5}$$

onde $\Gamma(a,b)=\int_b^{\infty} z^{a-1}e^{-z}dz$ representa uma função gama incompleta.

A Figura 5.3 mostra os *Zipf plots* (ζ em função de j/n) para ambos os grupos estudados usando os valores de β obtidos pelo método dos momentos ($\beta = 0.31$ para o Grupo 1 e $\beta = 0.47$ para o Grupo 2). Para uma melhor compreensão e comparativo, mostramos um ajuste linear (linha vermelha contínua) e um comportamento esperado exato (linha azul tracejada). Observamos um significativo comportamento linear como esperado para ambos os casos. Contudo, alguns desvios foram identificados, como é possível observar no gráfico para o Grupo 2 onde a distribuição acumulada está distante do comportamento esperado exato e no gráfico para o Grupo 1, onde a distribuição acumulada também apresenta pontos discrepantes em relação ao comportamento esperado. Após a conclusão deste experimento, identificamos a necessidade de um método que fosse mais preciso e com isso executamos testes usando Estatística Generalizada e observamos o comportamento das distribuições.

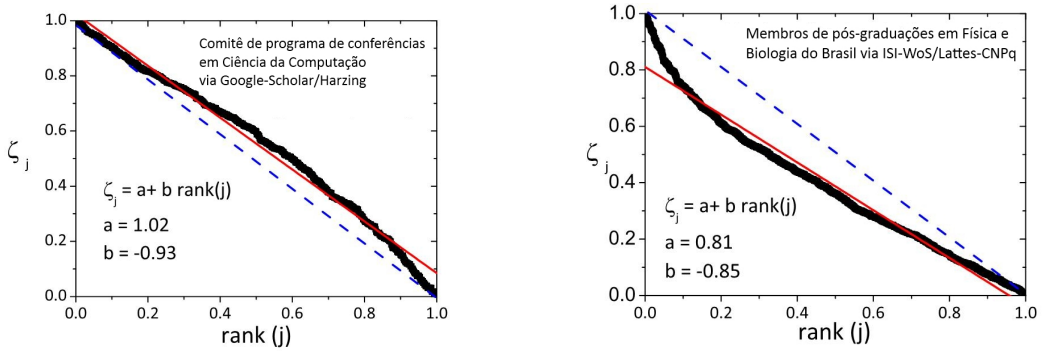

Figura 5.3: *Zipf plots*, usando *Stretched Exponential*, para Grupo 1 e Grupo 2, respectivamente

Da mesma forma com o uso de *Stretched Exponential*, foram executados testes usando Estatística Generalizada para estudar as distribuições de citações de acordo com a Equação 4.3. Foram calculadas numericamente as razões da distribuição de citação, descritas por

$$\Psi_k = \frac{\langle x^k\rangle}{\langle x\rangle_q^k} = \frac{1}{(2-q)\widehat{x}^{k+1}}\int_0^{\infty} x^k\left[1+\frac{(q-1)}{(2-q)\overline{x}}x\right]^{q/(1-q)}dx \tag{5.6}$$

onde apenas o primeiro momento $\langle x\rangle_q$ foi estimado experimentalmente pelo primeiro momento $\widehat{x}$. Do mesmo modo, comparamos Ψ_k com os momentos experimentais ($\Phi_k^{(exp)}$) calculados através da Equação 5.3, os quais podem ser observados na Figura 5.4.

Os melhores resultados obtidos foram $q = 1.37$ para o Grupo 1 (Ciência da Computação) e $q = 1.27$ para o Grupo 2 (Membros da Física e Biologia). Neste caso, para gerar o *Zipf plot*, devemos calcular a distribuição acumulada

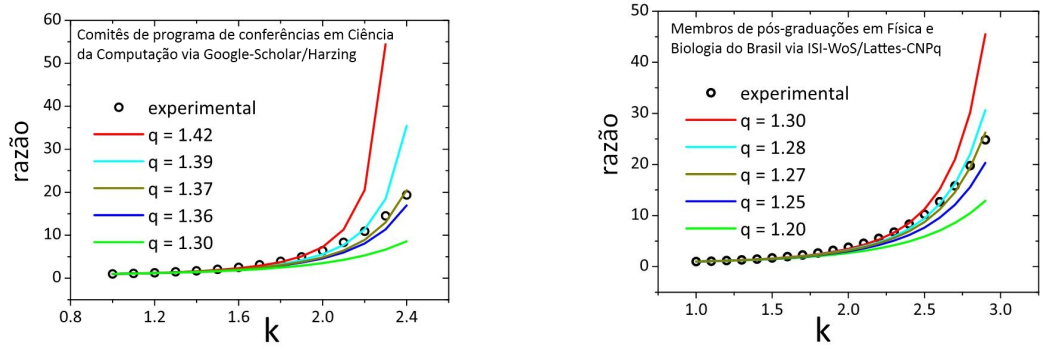

Figura 5.4: Momentos teóricos e experimentais, usando Estatística Generalizada, para Grupo 1 e Grupo 2, respectivamente

$$\begin{aligned}
\zeta_j &= \int_{x_j}^{\infty} P_q(x)dx \\
&= \frac{1}{(2-q)\overline{x}} \int_{x_j}^{\infty} \left[1 + \frac{(q-1)}{(2-q)\overline{x}}x\right]^{q/(1-q)} dx \\
&= \frac{1}{(2-q)\overline{x}} \int_{1+\frac{(q-1)}{(2-q)\overline{x}}}^{\infty} x^{q/(1-q)} dx \\
&= \left(1 + \frac{(q-1)}{(2-q)\overline{x}}x_j\right)^{1/(1-q)}
\end{aligned} \tag{5.7}$$

Usando os valores e os respectivos valores para $\widehat{x}$ obtidos na Tabela 5.1, foram gerados *plots* de ζ_j em função do *rank* (j/n) ilustrados na Figura 5.5.

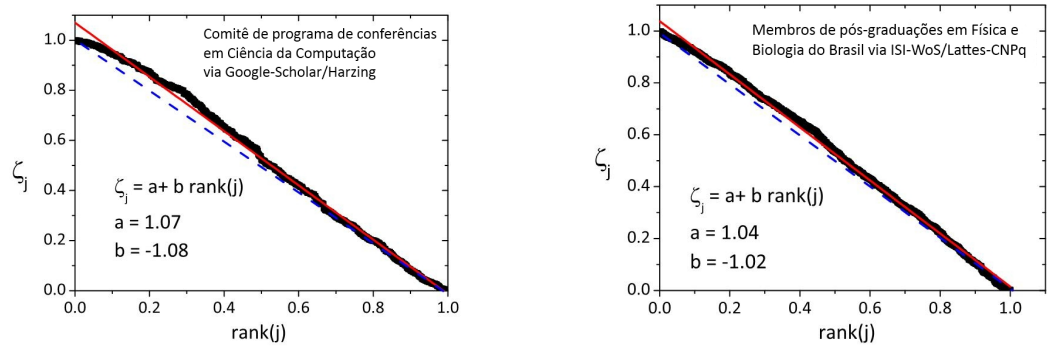

Figura 5.5: *Zipf plots*, usando Estatística Generalizada, para Grupo 1 e Grupo 2, respectivamente

Neste ponto, podemos observar um melhor comportamento linear para o *Zipf plot* para a abordagem usando Estatística Generalizada em comparação ao uso de *Stretched Exponential*. No entanto, uma questão importante a ser observada é o caso do mesmo valor de q ser obtido através de uma distribuição de índice h. Antes de demonstrarmos a aplicação da distribuição de índice h proposta na Equação 4.5, podemos testar a distribuição de índice h dos grupos estudados por meio de *Stretched Exponential*, conforme a Equação 4.6. Usando os estimadores $\widehat{c}$ e $\widehat{x}$ para cada grupo estudado (ver Tabela 5.1) encontramos numericamente o β que minimiza $\chi^2_\beta = \sum_{h=h_{min}}^{h_{max}} \left[f^{(\exp)}(h) - H_\beta(h)\right]^2$ e o q que minimiza $\chi^2_q = \sum_{h=h_{min}}^{h_{max}} \left[f^{(\exp)}(h) - H_q(h)\right]^2$, onde $H_\beta(h)$ é computado pela Equação 4.9, e

$H_q(h)$ é computado pela Equação 4.5. Executamos nosso aplicativo estatístico com um intervalo aceitável para $q(q_{min} = 1.01$ até $q_{max} = 1.99)$ e $\beta(\beta_{min} = 0.01$ até $\beta_{max} = 0.99)$ e a resolução usado foi $\Delta q = \Delta\beta = 0.01$.

Tabela 5.2: Comparação entre parâmetros obtidos com estatística generalizada (q) e com exponencial alongada (β)

Grupos	q(t.)	q(exp.)	β(t.)	β(exp.)
Conferências de C. da Computação (Harzing/Google Scholar)	1.24(1)	1.37(1)	0.64(1)	0.31(1)
Pesquisadores de Pós-graduações em Física e Biologia	1.26(1)	1.27(1)	0.66(1)	0.47(1)

* t. é abreviatura para a palavra **teórico** e exp. para a palavra **experimental**

Foram obtidos bons ajustes para as duas abordagens utilizadas no trabalho (Figura 5.6), contudo é importante observar que o uso de Estatística Generalizada produz um resultado muito melhor entre o obtido por meio da distribuição de citações e distribuição de índice h conforme apresentado na Tabela 5.2, a qual apresenta os momentos teóricos e experimentais para os grupos estudados. Esse fato também foi identificado nas Figuras 5.3 e 5.5, pois através do Zipf plot observamos diferenças importantes entre as duas abordagens. Para o estudo de caso envolvendo pesquisadores da pós-graduação em Física e Biologia, temos um parâmetro estimado de forma exata mostrando sua grande robustez quando comparado a fórmulas alternativas baseadas no trabalho de Laherrere e Sornette (1998). Também é possível notar que para os dados obtidos da Harzing/GS, foram produzidas estimativas mais distantes. Isso acontece pela natureza dos dados obtidos, os quais possuem um nível de ruído muito maior que a base controlada do ISI Web Of Science.

As distribuições ajustadas pelas Equações 4.5 e 4.9 são apresentadas na Figura 5.6 e indicam que o uso de Estatística Generalizada é mais interessante na descrição de distribuições de índice h, provendo fórmulas simples para índice h de grupos bastante distintos em diferentes bases. Esse estudo é importante, pois provê uma distribuição universal de índice h caracterizado por meio do simples parâmetro generalizado q.

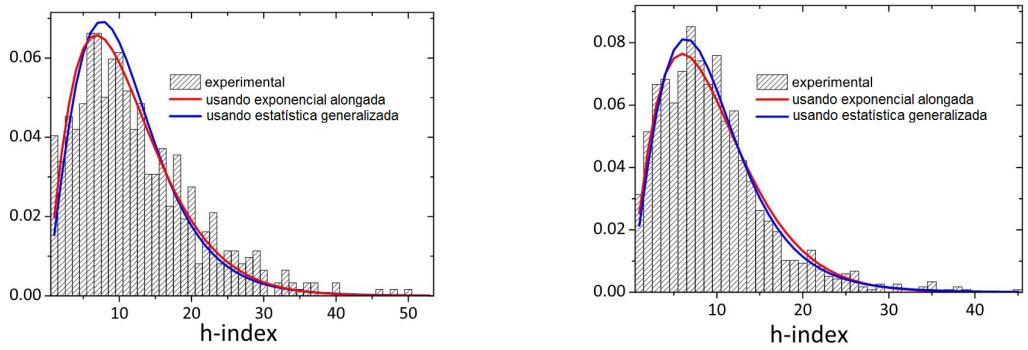

Figura 5.6: Distribuição de índice h para ambos os grupos estudados (a linha vermelha corresponde ao propósito da Equação 4.5 e a linha azul tracejada corresponde ao propósito da Equação 4.9)

5.2 Validação do uso do s-index

Esta seção descreve o processo de validação experimental, de forma a determinar a qualidade e relevância dos resultados obtidos com a aplicação da métrica do $s - index$ para grupos de pesquisadores em pós-graduações brasileiras nas áreas de biologia e física.

O experimento aqui conduzido buscou uma maneira de determinar a correlação entre a classificação obtida através da aplicação do $s - index$ descrita na dissertação e uma classificação já existente, efetuada pela CAPES, a qual faz uso de outras variáveis para avaliação de programas de pós-graduação que não apenas a produtividade por meio de artigos e citações.

5.2.1 Métricas de avaliação

Os experimentos realizados utilizam as seguintes métricas para avaliação dos resultados obtidos: Coeficiente de Spearman, Coeficiente de Kendall, $NDCG$ e testes de significância para comprovação estatística dos resultados. A descrição detalhada das métricas está presente nos Apêndices dessa dissertação.

A escolha destas métricas teve como principal razão o objetivo do modelo proposto, que é de obtenção de um *ranking* que sirva de suporte à avaliação de instituições (ou grupos de pesquisadores). Nesse contexto, as métricas Spearman e Kendall são empregadas quando da necessidade de verificação da correlação entre duas variáveis apresentadas em postos[2].

A métrica $NDCG$ é mais recente e provém da área de Recuperação de Informações, na qual grande parte dos esforços se dão na análise e melhoramento de *rankings* para uso em motores de busca (como Google e Yahoo!), cujo foco está na apresentação de uma lista com os documentos mais relevantes mediante uma consulta feita a um conjunto de documentos. Essa métrica se mostrou útil no contexto do $s - index$, pois mede o quanto uma classificação retornada por um sistema se aproxima de uma classificação teórica ideal. A principal diferença entre as métricas tradicionais Spearman e Kendall para o $NDCG$ está no fato de que nestas todas as posições (postos) de um *ranking* têm influência no resultado final da correlação, além de atribuir um mesmo peso independente da posição em que estão ordenados. Já no $NDCG$, as primeiras posições têm uma maior importância e, portanto, as posições inferiores são penalizadas no cálculo. Para o presente trabalho, essa métrica interessa pelo fato de naturalmente darmos mais atenção às primeiras posições de uma classificação do que as mais inferiores, fato que é verdadeiro ao realizar classificação de grupos de pesquisadores.

5.2.2 Metodologia

Os testes foram divididos em dois grupos, pesquisadores de programas brasileiros de pós-graduação nas áreas de física e de biologia. De posse desses dados, foram aplicadas as etapas de cálculo do h_2 absoluto e cálculo do h_2 relativo de cada instituição presente na listagem de sua respectiva área. Fazendo uso da Equação 4.10, cada instituição teve calculado seu valor de h_2 relativo a outra instituição de menor tamanho. Após, foi aplicada a Equação 4.11 para chegar a um valor final e absoluto que caracteriza uma instituição frente a outra, levando em consideração seu tamanho, índice h de seus pesquisadores e possibilitando a ordenação dos valores em forma de *ranking* sem empates.

Para fins de comparação, o *baseline* utilizado nesse trabalho foi a 'Relação de Cursos Recomendados e Reconhecidos' da CAPES, no qual aparecem os programas de pós-graduação que possuem mestrado e doutorado ou mestrado apenas. Essa avaliação segue a escala de 1 a 7 pontos, na qual 7 é a nota máxima (excelente) e 1 é a nota mínima (péssimo), porém para entrar nessa relação um curso deve ter, após a sua avaliação, obtido

[2]Usamos a palavra postos no contexto da estatística, representando variáveis qualitativas sujeitas a avaliações subjetivas quanto à preferência ou desempenho em um conjunto de observações.

uma nota igual ou maior que 3 (que representa o limiar entre cursos reconhecidos e reprovados). Portanto, na descrição dos resultados obtidos a seguir, a classificação ideal será a classificação dada pela CAPES aos cursos e a classificação calculada será o $s - index$ proposto nesta dissertação. É importante ressaltar que não será discutida aqui a qualidade e precisão do *ranking* construído pela CAPES e sim apresentada uma alternativa baseada no uso do índice h para alcançar resultados semelhantes, ou ainda auxiliar a tomada de decisões em conjunto com outros indicadores. Para construção das tabelas, foi necessário resolver a questão de empates apresentados no *ranking* da CAPES (instituições empatadas) e para isso foi utilizada a técnica de soma dos postos, com o cálculo da média aritmética destes e atribuição de um mesmo valor médio para mais de um posto, assim tornando possível a aplicação da fórmula de Spearman. A aplicação desse procedimento pode ser observada nas Tabelas 5.3 e 5.4, nas quais existem várias instituições que possuem a mesma nota da CAPES impedindo a determinação de um posto específico para cada, assim é necessário atribuir a média dos postos que lhes caberiam caso não houvesse empates.

A tabela 5.3 apresenta os resultados da correlação entre o *ranking* da CAPES e o *ranking* gerado pelo $s - index$ proposto, através do coeficiente ρ de Spearman. Nesta tabela, o valor d_i representa a diferença entre o valor dos postos obtidos na comparação entre os duas classificações e d_i^2 o valor elevado ao quadrado, sendo utilizado na fórmula para obtenção do coeficiente ρ de Spearman apresentada na Equação 5.8, onde n é o número de pares.

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{5.8}$$

Os dados relativos à área de física são representados por programas de pós-graduação brasileiros na proporção de: seis instituições com nota 7, uma instituição com nota 6, quatro instituições com nota 5, cinco instituições com nota 4 e duas instituições com nota 3. Neste contexto, a nota refere-se à avaliação da CAPES, conferindo uma amostra com 18 instituições.

O coeficiente ρ de Spearman gerado para a classificação de programas de pós-graduação em física foi 0.7975, onde $\sum d_i^2 = 190$. Para verificar a significância estatística, utilizamos um teste bi-caudal para erro tipo I com $\alpha = 0.05$. Observando a tabela de valores críticos (presente no Anexo) para $n = 18$ temos o valor de 0.472, que sugere que a hipótese nula (h_0) seja rejeitada já que obtemos um valor de ρ de 0.7975. Com isso, assumimos que existe correlação entre as duas classificações usadas.

Para programas de pós-graduação em biologia, a proporção foi de: quatro instituições com nota 6, cinco instituições com nota 5, cinco instituições com nota 4 e quatro instituições com nota 3, totalizando 18 instituições na amostra. A Tabela 5.4 apresenta os resultados de classificação para programas de pós-graduação em biologia, tendo como coeficiente ρ obtido o valor de 0.6841, onde $\sum d_i^2 = 297$. Igualmente, como para os dados das instituições de física, aplicamos o teste de significância estatítica bi-caudal para erro tipo I com $\alpha = 0.05$, de forma que o valor de ρ 0.6841 é maior que 0.472 para $n = 18$ e corrobora para a rejeição da hipótese nula (h_0). Assumimos que existe correlação entre a classificação da CAPES e do $s - index$ proposto para dados de biologia.

Na coleta de dados foram obtidos mais dados de instituições de biologia do que da área de física, para tanto, se tomou o cuidado de utilizar a correlação de Spearman para um mesmo n, que nesse caso foi 18. Porém, vale salientar que usando todos os dados

Tabela 5.3: Resultados do coeficiente de Spearman para programas de pós-graduação em física

Instituição	Nota CAPES	$s-index$	Rank CAPES	Rank $s-index$	d_i (*)	d_i^2
1	7	21.08	3.5	1	2.5	6.25
2	7	17.69	3.5	2	1.5	2.25
3	7	17.08	3.5	3	0.5	0.25
4	7	16.82	3.5	4	-0.5	0.25
5	7	13.49	3.5	9	-5.5	30.25
6	7	13.86	3.5	8	-4.5	20.25
7	6	14.23	7	6	1	1
8	5	13.10	9.5	10	-0.5	0.25
9	5	13.90	9.5	7	2.5	6.25
10	5	11.24	9.5	14	-4.5	20.25
11	5	12.04	9.5	12	-2.5	6.25
12	4	15.38	14	5	9	81
13	4	11.01	14	15	-1	1
14	4	11.80	14	13	1	1
15	4	12.55	14	11	3	9
16	4	10.29	14	16	-2	4
17	3	9.68	17.5	17	0.5	0.25
18	3	9.08	17.5	18	-0.5	0.25

(*) Diferença entre o Rank CAPES e o Rank $s-index$

de instituições de biologia obtidos, os quais totalizam 26, o coeficiente ρ foi de 0.6766, variando minimamente em relação aquele obtido com 18 instituições. Portanto, pode-se afirmar que o resultado é estatísticamente relevante e mostra uma correlação acima do valor crítico de Spearman entre o *ranking* gerado através do modelo de $s-index$ e o *baseline* adotado, que neste trabalho se apresenta como a avaliação feita pela CAPES.

Tabela 5.4: Resultados do coeficiente de Spearman para biologia

Instituição	Nota CAPES	$s-index$	Rank CAPES	Rank $s-index$	d_i (*)	d_i^2
1	6	15.39	2.5	1	1.5	2.25
2	6	12.26	2.5	5	-2.5	6.25
3	6	11.96	2.5	6	-3.5	12.25
4	6	11.42	2.5	9	-6.5	42.25
5	5	13.93	7	2	5.0	25.00
6	5	12.36	7	4	3.0	9.00
7	5	11.72	7	8	-1.0	1.00
8	5	11.15	7	10	-3.0	9.00
9	5	9.56	7	15	-8.0	64.00
10	4	13.37	12.0	3	9.0	81.00
11	4	11.88	12.0	7	5.0	25.00
12	4	10.33	12.0	11	1.0	1.00
13	4	10.22	12.0	12	0.0	0.00
14	4	9.75	12.0	14	-2.0	4.00
15	3	9.83	16.5	13	3.5	12.25
16	3	8.30	16.5	16	0.5	0.25
17	3	7.47	16.5	17	-0.5	0.25
18	3	7.25	16.5	18	-1.5	2.25

(*) Diferença entre o Rank CAPES e o Rank $s-index$

Conforme Howell (2006) relata, dados na forma de *ranking* não podem ser expressos como uma distribuição normal e não existe um método consolidado para cálculo do erro

padrão de ρ para pequenas amostras. O autor ainda reitera que, apesar das tabelas de valores críticos para o coeficiente de Spearman apresentarem valores de n de 4 até 100, constata-se que para $n >= 28$ os valores apresentados são baseados apenas em aproximações. Em termos reais, uma pessoa dificilmente conseguiria ordenar amostras contendo, por exemplo, 30 valores, já que provavelmente muitos seriam difíceis de classificar por terem um peso semelhante. Tendo ciência disso, o número de dados de instituições coletadas mostra-se adequado para o estudo proposto e sua validação estatística.

Tabela 5.5: Mapeamento da avaliação de cursos da CAPES e o uso no DCG

Nota CAPES	Valor correspondente no DCG
7	5
6	4
5	3
4	2
3	1

Foi realizado um mapeamento na definição dos valores de relevância para cálculo do DCG e posteriormente o $NDCG$. A Tabela 5.5 apresenta o mapeamento, no qual a nota mais relevante dada pela CAPES é 7 e a menos relevante é 3, fazendo um mapeamento para valores entre 5 e 1 no cálculo do DCG. O valor zero (irrelevante) não aparece no mapeamento, pois a classificação da CAPES já faz essa filtragem, excluindo cursos que não atingem a nota mínima 3 e portanto temos cursos pouco relevantes, porém nenhum irrelevante. É importante salientar que o uso da palavra "irrelevante" está associado apenas ao conceito de obtenção de um *ranking* e não em termos de educação e trabalho efetuado por outras instituições que não foram selecionadas para fazer parte da amostra no estudo de caso proposto. De forma breve, o DCG reduz a relevância de cada item de acordo com uma função logarítmica, por meio da qual são mais valorizados os itens que aparecem nas primeiras posições e valorizando menos as últimas posições de uma classificação.

A Tabela 5.6 mostra valores obtidos aplicando ρ de Spearman, τ de Kendall e o $NDCG$. O uso do coeficiente τ assemelha-se ao coeficiente ρ, porém sua interpretação difere em termos da magnitude dos resultados, já que o coeficiente ρ é calculado em termos da proporção da variabilidade entre os postos e o coeficiente τ representa a diferença entre a probabilidade de um dado observado estar na mesma ordem que o dado amostral e o fato de não estarem na mesma ordem.

Tabela 5.6: Resultados usando ρ de Spearman, τ de Kendall e $NDCG$ para os grupos de biologia e física

Grupo (n=18)	Spearman Rho	Kendall Tau	NDCG
Física	0.7975	0.6835 ($p = 0.0001$)*	0.9827
Biologia	0.6841	0.5512 ($p = 0.002$)*	0.9365

* Valor p obtido através do teste de permutação monte carlo

Para os dados apresentados na Tabela 5.6, podemos concluir que os valores de Kendall para física e biologia corroboram os experimentos usando o coeficiente ρ de Spearman, pois apresentam um valor de p de 0.0001 e 0.002, respectivamente, para um nível de significância $\alpha = 0.05$. Isso nos leva a rejeitar a hipótese nula (h_0) de que os dados não estão correlacionados. Os valores obtidos aplicando a métrica $NDCG$, presentes na Tabela 5.6, apresentam uma magnitude alta para o intervalo $0 \leq ndcg \leq 1$. Quanto mais próximo de 1 o valor do $ndcg$, mais próximo está a classificação obtida de uma classificação teórica

ideal para determinado conjunto de documento (ou valores do $s - index$ para programas de pós-graduação). Sendo assim, $NDCG$ é uma métrica que representa a nitidez no *ranking* obtido.

Portanto, concluímos que a ordenação das instituições usando a métrica do $s - index$ possui boa correspondência com a ordenação obtida observando a classificação da CAPES, no que diz respeito aos de programas de pós-graduação. Lembramos que a métrica proposta mantém instituições com alta avaliação da CAPES nos primeiros postos e instituições com avaliação mínima nos últimos postos; sendo assim, a variação maior de postos ocorre em instituições com notas intermediárias (4 até 6) na CAPES.

6 CONSIDERAÇÕES FINAIS

O presente trabalho apresenta de forma original uma fórmula para distribuição de índice h de um grupo de pesquisadores, baseando-se em métodos existentes na literatura, tais como a caracterização de distribuições de citações através de uma fórmula única apresentada em Tsallis e Albuquerque (2000) e o uso dos conceitos de Laherrere e Sornette (1998) sobre distribuição de citações que seguem um comportamento caracterizado por uma exponencial alongada. A partir desses estudos prévios e com a inclusão da dependência quadrática de Hirsch (2005) para o total de citações de um autor e seu índice h, foi possível definir formulações para distribuição de índice h, aplicação dessas fórmulas aos dados coletados (por meio do ISI Web Of Science e Google Scholar), bem como a validação destes por meio de gráficos que apresentam o ajuste e a universalidade obtida para as distribuições em questão. Foram ainda aplicados os métodos de Tsallis e Albuquerque (2000) e Laherrere e Sornette (1998) para distribuição de citações nos dados coletados, por meio de adaptações nas fórmulas originais que facilitaram os cálculos e a obtenção de resultados importantes sobre pesquisadores presentes nas áreas estudadas (física, biologia e ciência da computação). Um resultado importante obtido mostra que o uso da chamada q-exponencial apresenta melhores ajustes e que o valor de q obtido nos dois grupos avaliados são próximos, sugerindo um comportamento universal em relação ao parâmetro q.

Entre as principais conclusões obtidas neste trabalho destaca-se a aplicação da fórmula para distribuição de índice h com base na estatística generalizada, pois apresenta bons ajustes e pode ser utilizada como ferramenta para classificação de grupos de pesquisadores. Seu uso é interessante para se ter uma visão quantitativa da produção científica dos pesquisadores através do índice h, dando um enfoque menos qualitativo e mais quantitativo. É importante salientar que as áreas de física e biologia foram utilizadas em conjunto na validação do modelo de distribuição de índice h, pois são áreas que possuem uma mesma prática de publicação, enfatizando publicações em períodicos. Dessa forma, conseguimos obter uma amostra consistente de dados, totalizando dados de 1.203 pesquisadores brasileiros nestas duas áreas em específico, além de avaliar uma área que publica predominantemente em conferências, que é a área da ciência da computação. Tendo sido obtido este modelo de distribuição a partir do índice h, conseguimos ter uma visão abrangente da forma como os pesquisadores se comportam nas suas áreas de estudo e observamos quais leis regem esse comportamento científico que envolve publicações de artigos e citações.

O trabalho também definiu uma nova métrica, baseada no índice h de Hirsch (2005) e no índice h sucessivo de segunda ordem de Schubert (2007), bem como nas contribuições apresentadas em Silva et al. (2010) para obtenção de um índice que permite a avaliação e comparação de grupos de pesquisadores com diferentes tamanhos, fazendo uso de

técnicas de *sampling*, por meio da qual cada grupo (ou instituição, como nos dados coletados) é repesado em função de seu índice h em potencial em relação aos outros grupos presentes na amostra. Esta técnica permitiu chegarmos a um novo índice, chamado de $s - index$, com as seguintes características: homogeneidade, magnitude e escala. Para validação desta técnica foram usados os dados coletados para física e biologia, e foi obtido o $s - index$ para cada instituição avaliada. Posteriomente foram feitos cálculos acerca da correlação entre a métrica definida no trabalho e as avaliações sobre programas de pós-graduação efetuadas pela CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) a fim de detectar as proximidades entre os dois métodos de avaliação, que são bastante distintos no número de variáveis consideradas. Como resultados, para a amostra relativa a área de física observamos correlações positiva de 0.7975 através do coeficiente de Spearman, 0.6835 através do coeficiente de Kendall e 0.9827 através da métrica de classificação $NDCG$. Com os dados relativos a área de biologia, constatamos correlações positivas de 0.6841 através do coeficiente de Spearman, 0.5512 através o coeficiente de Kendall e 0.9365 através da métrica de classificação $NDCG$. Com esses resultados, conseguimos observar que existe uma correlação relevante entre a métrica $s - index$ e a classificação apresentada pela CAPES, mostrando que a redução no número de variáveis usadas na métrica $s - index$, desenvolvida nesta dissertação, não trouxe prejuízos aos resultados obtidos e mostra-se promissora à medida que pode ser facilmente calculada. Portanto, através do índice proposto é apresentada uma alternativa apenas baseada no índice h para classificação de grupos de pesquisadores. A principal vantagem está no número reduzido de variáveis utilizadas e na facilidade de obtenção, através de consultas na web, dos dados necessários para chegar a uma classificação coerente. Esta classificação pode dar suporte a tarefas como a destinação de recursos para grupos competentes e focando em um item importante da ciência que é a disseminação de conhecimento por meio da produção científica dos pesquisadores.

Como produção científica, foi publicado o seguinte artigo:

Autores	Roberto da Silva, Fahad Kalil, Alexandre Souto Martinez, J. Palazzo de Oliveira
Título	**Universality in Bibliometrics**
Periódico	Physica A
Ano	2011
Resumo	Nesse artigo foram definidos métodos para obtenção de uma distribuição universal para índice h, através de diferentes abordagens e utilizando diferentes bases de dados para distribuição de citações

Como trabalhos futuros, foi verificada a necessidade de estender as análises experimentais para outros conjuntos de dados (áreas científicas) e também para grupos aleatórios de pesquisadores que englobem pesquisadores de diferentes áreas. Desta forma, será possível observar se o modelo de distribuição de índice h desenvolvido é aplicável a mais áreas da ciência, demonstrando universalidade nas formas de publicação. Quanto à métrica $s - index$, também podemos expandir a análise a outras áreas além da física e biologia, ou ainda realizar testes usando dados mais precisos sobre citações obtidas e índice h. Incentivamos ainda a criação de um sistema computacional completo que permitirá a inserção de dados de pesquisadores e retornará *rankings*, além de realizar o ajuste das distribuições automaticamente a partir dos estudos e equações propostas no trabalho. Com isso, teremos uma aplicação prática do conteúdo desenvolvido nesta dissertação e um número maior de pessoas poderiam fazer uso e validar as propostas apresentadas.

REFERÊNCIAS

ALONSO, S.; CABRERIZO, F.; HERRERA-VIEDMA, E.; HERRERA, F. H-Index: a review focused in its variants, computation and standardization for different scientific fields. **Journal of Informetrics**, Oxford, v.3, n.4, p.273–289, out. 2009.

ALTMANN, J.; ABBASI, A.; HWANG, J. The RP-Index and the CP-Index for Evaluating the Productivity of Researchers and their Communities. **IJCSA**, [S.l.], v.6, n.2, p.104–118, 2009.

ALVARENGA, P. J. L.; ARAUJO NETO, B. Ribeiro de. **Um estudo sobre referências bibliográficas na área de ciência da computação**. 2007. 50 f. Dissertação (Mestrado em Ciência da Computação) — Curso de Pós-Graduação em Ciência da Computação, Universidade Federal de Minas Gerais, Belo Horizonte.

ANASTASIADIS, A. D.; ALBUQUERQUE, M. P. de; ALBUQUERQUE, M. P. de. A characterization of the scientific impact of Brazilian institutions. **Brazilian Journal of Physics**, São Paulo, v.39, n.2a, p.511–518, ago. 2009.

ARAÚJO, C. Bibliometria: evolução histórica e questões atuais. **Em Questão**, Porto Alegre, v.12, n.1, p.11–32, jan./jun. 2006.

BARRETO, A. **Mitos e lendas da informação**: o texto, o hipertexto e o conhecimento. Disponível em: <http://www.dgz.org.br/fev07/Art_02.htm>. Acesso em: out. 2010.

BERBERAN-SANTOS, M.; BODUNOV, E. N.; VALEUR, B. History of the Kohlrausch (stretched exponential) function: pioneering work in luminescence. **Annalen der Physik**, Berlin, v.17, n.7, p.460–461, 2008.

BORNMANN, L.; MARX, W. The h index as a research performance indicator. **European Science Editing**, [S.l.], v.3, n.37, p.77–80, ago. 2011.

COSTA, R. A comunicação eletrônica e a alteração de tempo e espaço na produção do conhecimento científico. **Ciência da Informação**, Brasília, v.36, n.2, p.7–15, maio/ago. 2007.

EGGHE, L. Theory and practise of the g-index. **Scientometrics**, Budapest, v.69, n.1, p.131–152, 2006.

EGGHE, L. Modelling sucessive h-indices. **Scientometrics**, Budapest, v.77, n.3, p.377–387, 2008.

EGGHE, L.; RAO, I. K. R. Study of different h-indices for groups of authors. **JASIST**, [S.l.], v.59, n.8, p.1276–1281, 2008.

FRANCESCHINI, F.; MAISANO, D. A. Analysis of the Hirsch index's operational properties. **European Journal of Operational Research**, [S.l.], v.203, n.2, p.494–504, jun. 2010.

GUO, L.; TAN, E.; CHEN, S.; XIAO, Z.; ZHANG, X. The stretched exponential distribution of internet media access patterns. In: ACM SYMPOSIUM ON PRINCIPLES OF DISTRIBUTED COMPUTING. **Proceedings...** New York: ACM, 2008. p.283–294.

HALL, A. O. **Associação entre variáveis**. Disponível em <http://www2.mat.ua.pt/pessoais/AHall/TEA/Capcorrel.pdf>. Acesso em: janeiro de 2011.

HIRSCH, J. An index to quantify an individual's scientific research output. **Proceedings of the National Academy of Sciences**, Washington, v.102, n.46, p.16569–16572, 2005.

HOWELL, D. C. **Statistical methods for psychology**. 6.ed. [S.l.]: Wadsworth Publishing, 2006.

HSU, J.; HUANG, D. Dynamics of citation distribution. **Computer Physics Communications**, [S.l.], v.182, n.1, p.185–187, jan. 2011.

JARVELIN, K.; KEKALAINEN, J. IR evaluation methods for retrieving highly relevant documents. In: ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 23. **Proceedings...** New York: ACM, 2000. p.41–48.

JARVELIN, K.; KEKALAINEN, J. Cumulated gain-based evaluation of IR techniques. **ACM Trans. Inf. Syst.**, New York, v.20, n.4, p.422–446, out. 2002.

LAHERRERE, J.; SORNETTE, D. Stretched exponential distributions in nature and economy: fat tails with characteristic scales. **The European Physical Journal B - Condensed Matter and Complex Systems**, [S.l.], v.2, n.4, p.525–539, maio 1998.

LANE, J. Let's make science metrics more scientific. **Nature**, [S.l.], v.464, n.7288, p.488–489, mar. 2010.

LEHMANN, S.; LAUTRUP, B.; JACKSON, A. D. Citation networks in high energy physics. **Phys. Rev. E**, [S.l.], v.68, n.2, p.026113, ago. 2003.

MAIER, G. Impact factors and peer judgment: the case of regional science journals. **Scientometrics**, [S.l.], v.69, p.651–667, 2006.

MALVA, M. **Coeficiente de correlação Ró de Spearman**. Disponível em: <http://bit.ly/gbDU72>. Acesso em: jan. 2011.

MCCABE, G. P.; MOORE, D. S. **Introduction to the practice of statistics**. [S.l.]: W.H. Freeman & Company, 2005.

MUGNAINI, R. **Caminhos para adequação da avaliação da produção científica brasileira**: impacto nacional versus internacional. 2006. 254 f. Tese (Doutorado em Ciência da Computação) — Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo.

PRATHAP, G. Hirsch-type indices for ranking institutions' scientific research output. **Current Science**, Bangalore, v.91, n.11, p.1438–1438, 2006.

RADICCHI, F.; FORTUNATO, S.; CASTELLANO, C. Universality of citation distributions: toward an objective measure of scientific impact. **Proceedings of the National Academy of Sciences**, Washington, v.105, n.45, p.17268–17272, nov. 2008.

RAMACHANDRAN, K.; TSOKOS, C. **Mathematical statistics with applications**. [S.l.]: Academic Press, 2009.

REDNER, S. How popular is your paper? An empirical study of the citation distribution. **The European Physical Journal B - Condensed Matter and Complex Systems**, [S.l.], v.4, n.2, p.131–134, ago. 1998.

ROUSSEAU, R. Reflections on recent developments of the h-index and h-type indices. **COLLNET Journal of Scientometrics and Information Management**, [S.l.], v.2, n.1, p.1–8, jun. 2008.

SCHREIBER, M. An empirical investigation of the g-index for 26 physicists in comparison with the h-index, the A-index, and the R-index. **J. Am. Soc. Inf. Sci. Technol.**, New York, v.59, n.9, p.1513–1522, 2008.

SCHUBERT, A. Successive h-indices. **Scientometrics**, Budapest, v.70, n.1, p.201–205, 2007.

SHESKIN, D. J. **Handbook of parametric and nonparametric statistical procedures**. 4.ed. [S.l.]: Chapman & Hall/CRC, 2007.

SILVA, J. A. da; BIANCHI, M. L. P. Cientometria: a métrica da ciência. **Paidéia (Ribeirão Preto)**, [S.l.], v.11, p.5–10, jul./dez. 2001.

SILVA, R. da; OLIVEIRA, J. P. M. de; LIMA, J. V. de; MOREIRA, V. Statistics for Ranking Program Committees and Editorial Boards. **CoRR**, [S.l.], v.abs/1002.1060, 2010.

SIMONS, K. The Misused Impact Factor. **Science**, [S.l.], v.322, n.5899, p.165, 2008.

SORMUNEN, E. Liberal relevance criteria of TREC - Counting on negligible documents? In: ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 25. **Proceedings...** New York: ACM, 2002. p.324–330.

SYPSA, V.; HATZAKIS, A. Assessing the impact of biomedical research in academic institutions of disparate sizes. **BMC medical research methodology**, [S.l.], v.9, p.33, maio 2009.

TOL, R. S. A rational, successive g-index applied to economics departments in Ireland. **Journal of Informetrics**, [S.l.], v.2, n.2, p.149 – 155, 2008.

TSALLIS, C. Nonextensive statistics: theoretical, experimental and computational evidences and connections. **Brazilian Journal of Physics**, [S.l.], v.29, p.1 – 35, 03 1999.

TSALLIS, C.; ALBUQUERQUE, M. de. Are citations of scientific papers a case of nonextensivity? **The European Physical Journal B - Condensed Matter and Complex Systems**, [S.l.], v.13, n.4, p.777–780, 2000.

WIKIPEDIA. **Spearman's rank correlation coefficient**. Disponível em: <http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient>. Acesso em: 15 dez. 2010.

WIKIPEDIA. **Discounted cumulative gain**. Disponível em: <http://en.wikipedia.org/wiki/Discounted_cumulative_gain>. Acesso em: 20 jan. 2011.

ZAR, J. H. Significance Testing of the Spearman Rank Correlation Coefficient. **Journal of the American Statistical Association**, [S.l.], v.67, n.339, p.578–580, 1972.

APÊNDICE I ESTIMATIVAS DE PARÂMETROS

1.1 Método dos mínimos quadrados

O método dos mínimos quadrados é bastante usado em situações envolvendo a estimação de parâmetros em modelos lineares. Esta técnica de otimização matemática permite a seleção do melhor estimador individual utilizando um processo que minimiza a soma dos quadrados das diferenças entre o valor estimado e os dados observados, sendo essa diferença descrita na literatura como resíduo. Esse método, portanto, consiste na obtenção de um estimador capaz de minimizar a soma dos quadrados dos resíduos da regressão, de modo a maximizar o grau de ajuste do modelo aos dados observados. Para tanto, calculam-se os parâmetros a e b da reta que minimiza estas diferenças (ou o erro) entre Y (observado) e Y' (calculado).

Para estimar o valor de determinada variável Y, usamos a equação:

$$Y = a + bx + \varepsilon \tag{1.1}$$

onde a é coeficiente linear, b é o coeficiente angular e ε é o erro, que representa a variação de Y não explicada pelo modelo.

As estimativas de mínimos quadrados das constantes a e b são então aqueles valores de a e b, os quais tornam mínima a expressão apresentada na Equação 1.2

$$S(a,b) = \sum_{i=1}^{n}(y_i - a - bx_i)^2 \tag{1.2}$$

que representa a soma dos quadrados dos resíduos.

Os candidatos a ponto de mínimo da Equação 1.2 são aqueles nos quais as derivadas parciais de $S(a,b)$ em relação a cada um dos parâmetros a e b são nulas. Desse modo, obtemos:

$$\frac{\partial S}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \tag{1.3}$$

$$\frac{\partial S}{\partial b} = -2\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0 \tag{1.4}$$

Igualar as derivadas a zero é necessário para constituir o mínimo de uma função. Essa série de derivadas produzirá um conjunto de equações em função dos parâmetros necessários, no caso ilustrado a e b, possibilitando definir os valores dos parâmetros em função do conjunto de pontos (x e y) teóricos.

Podemos então, distribuir e dividir a Equação 1.3 por $2n$ para obter a equação para o coeficiente a:

$$\frac{-2\sum_{i=1}^{n} y_i}{2n} + \frac{2\sum_{i=1}^{n} a}{2n} + \frac{2\sum_{i=1}^{n} bx_i}{2n} = \frac{0}{2n} \tag{1.5}$$

$$\frac{-\sum_{i=1}^{n} y_i}{n} + \frac{\sum_{i=1}^{n} a}{n} + \frac{b\sum_{i=1}^{n} x_i}{n} = 0 \tag{1.6}$$

$$-\bar{y} + a + b\bar{x} = 0 \tag{1.7}$$

$$a = \bar{y} - b\bar{x} \tag{1.8}$$

onde $\bar{y}$ é a média amostral de y e $\bar{x}$ é a média amostral de x. Fazendo a substituição na Equação 1.4, temos:

$$-2\sum_{i=1}^{n} x_i \left(y_i - \bar{y} + b\bar{x} - bx_i\right) = 0 \tag{1.9}$$

$$\sum_{i=1}^{n} \left[x_i \left(y_i - \bar{y}\right) + x_i b \left(\bar{x} - x_i\right)\right] = 0 \tag{1.10}$$

$$\sum_{i=1}^{n} x_i \left(y_i - \bar{y}\right) + b\sum_{i=1}^{n} x_i \left(\bar{x} - x_i\right) = 0 \tag{1.11}$$

$$b = \frac{\sum_{i=1}^{n} x_i \left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} x_i \left(x_i - \bar{x}\right)} \tag{1.12}$$

Desse modo, resolvendo as Equações 1.8 e 1.12, obtemos a equação que representa a reta ajustada para um conjunto de pontos, tal que a soma dos quadrados da distância dos pontos para a linha de ajuste seja minimizada. Essa minimização pode ser feita tanto na direção horizontal como vertical. Caso a regressão seja no eixo X, então a linha é ajustada de modo que os desvios horizontais dos pontos até a linha sejam minimizados. Caso a regressão seja no eixo Y, então significa que a distância dos desvios verticais sejam levadas em conta. A Figura 1.1 ilustra esses comportamentos, na qual as flechas mostram as distâncias $(y_i^0 - y_i)$ que devem ser calculadas. Essa distância refere-se à relação entre o ponto e a linha traçada.

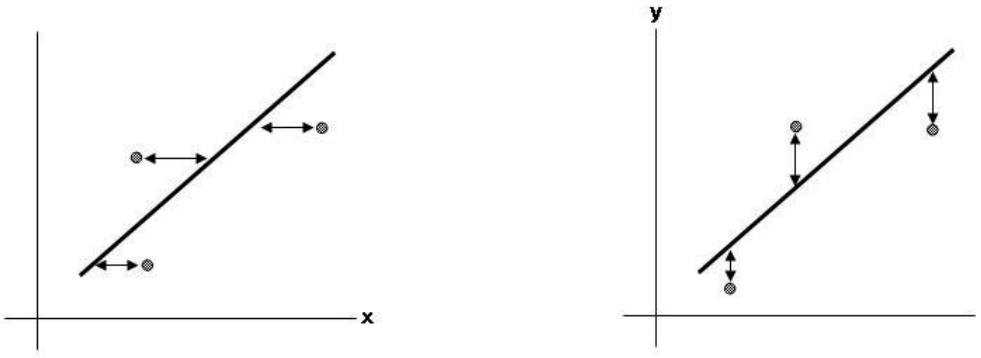

Figura 1.1: Distância minimizada na direção x e y, respectivamente

1.2 Método dos momentos

Um dos métodos mais tradicionais e antigos para estimativa de parâmetros é o **método dos momentos**. Um dos seus pontos fortes é que normalmente esse método é facilmente empregado e quase sempre atinge boas estimativas. Contudo, esse método retorna, muitas vezes, estimativas que devem ser melhoradas para serem utilizadas ou ainda, no pior caso, estimativas sem relação com os dados observados levando a busca por métodos mais eficazes.

O método dos momentos baseia-se no casamento dos momentos da amostra com os momentos populacionais (da distribuição) correspondentes. Este método busca equacionar os momentos (média, variância, etc.) implícitos no modelo estatístico da distribuição da população com os momentos observados na amostra. Desse modo, a amostra é tratada como sendo uma representação em miniatura da população e assume-se que as relações presentes na população também estão expressas na amostra coletada. Pelo fato dos momentos da população $\mu_j(\theta_1, ..., \theta_k)$ serem normalmente funções dos parâmetros da população, é possível equacionar correspondendo os momentos populacionais e amostrais e resolver as estimativas através dos momentos. Segundo Ramachandran e Tsokos (2009), para determinação dos momentos, as etapas enumeradas abaixo devem ser respeitadas. Supondo que temos k parâmetros para serem estimados, definidos como $\theta = (\theta_1, ..., \theta_k)$:

1. Encontrar k momentos da população, $\mu_j(\theta_1, ..., \theta_k)$ = E[X_j] para $j = 1, ..., k$. Dessa forma, μ_j conterá um ou mais parâmetros θ_1,...,θ_k

2. Encontrar k momentos correspondentes da amostra, $m_j(\theta_1, ..., \theta_k)$ para $j = 1, ..., k$. O número de momentos amostrais deve ser igual ao número de parâmetros a serem estimados

3. A partir do sistema de equações, $\mu_j = m_j = \mathrm{E}[X]_j$ para $j = 1, ..., k$, resolver para os parâmetros $\theta = (\theta_1, ..., \theta_k)$; o que será um estimador do momento $\hat{\theta}$.

APÊNDICE II MÉTRICAS DE AVALIAÇÃO DE *RANKINGS*

2.1 Coeficiente de correlação de Spearman

O coeficiente ρ de Spearman mede a intensidade da relação entre variáveis ordinais, ou seja, que respeitam um *ranking*. Usa-se, ao invés do valor observado, apenas a ordem das observações. Deste modo, este coeficiente não é sensível às assimetrias na distribuição, nem à presença de *outliers*[1], não exigindo, portanto, que os dados provenham de duas populações normais. É usado quando não é possível medir a correlação entre variáveis através do coeficiente de Pearson, no qual não se pode violar a normalidade (MALVA, 2007). A fórmula para o cálculo do coeficiente ρ é dada pela Equação 2.13, onde n é o número de pares (xi, yi) e d_i são os postos de x_i dentre os valores de x subtraídos dos postos de y_i dentre os valores de y. Se os postos de x forem exatamente iguais aos pontos de y, então todos os d_i serão zero e ρ será 1.

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{2.13}$$

O grau de associação existente entre dois conjuntos de dados pode ser medido através do coeficiente de correlação de Spearman, que varia entre -1.0 e $+1.0$, sendo que o valor 0 (zero) representa uma correlação nula, ou seja, não apresenta nenhuma correlação entre as variáveis. Com isso, as análises de correlação buscam identificar se as variáveis seguem nas seguintes formas: mesmo sentido (coeficiente de correlação positivo), em sentidos opostos (coeficiente de correlação negativo) ou não há correlação entre as variáveis (coeficiente de correlação zero).

A Tabela 2.1 apresenta como devem ser dispostos os dados para obter d_i^2 de cada comparação entre as classificações geradas. O exemplo mostrado é sobre a determinação da correlação entre o as variáveis '*QI*' e as '*Horas frente à TV por semana*' de um grupo de crianças em um conjunto de dados com 10 casos. As duas variáveis em questão têm seus dados colocados lado a lado e o *ranking* de cada uma é feito com base na magnitude dos números. Na variável *QI*, o menor valor é 86 e para este é dado o posto 1 na classificação e na variável *Horas frente à TV por semana*, o menor valor é 0 e seu posto no *ranking* também é o primeiro, sendo repetido esse processo até o final da amostra. Após a etapa de ordenação e criação dos *rankings*, é calculada a diferença d_i para cada par de postos gerado e depois é calculado o d_i^2 como forma de normalização para que não apresentem valores negativos que anulem as diferenças presentes. Por fim, é obtido o valor de ρ fazendo uso da Equação 2.13.

[1] *Outlier* é uma observação que foge do padrão predominante apresentado em uma distribuição e é facilmente observada em gráficos do tipo *scatterplot* e histogramas (MCCABE; MOORE, 2005).

Tabela 2.1: Tabela de classificação para o coeficiente de Spearman

QI, X_i	Horas frente à TV por semana, Y_i	Rank x_i	Rank y_i	d_i	d_i^2
86	0	1	1	0	0
97	20	2	6	-4	16
99	28	3	8	-5	25
100	27	4	7	-3	9
101	50	5	10	-5	25
103	29	6	9	-3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

Fonte: adaptado de (WIKIPEDIA, 2010)

Existem situações envolvendo dados no formato ordinal que apresentam empates nos conjuntos de dados, no qual mais de um posto possui o mesmo valor de ordenação e com isso correções devem ser aplicadas no cálculo. A razão para isso é que na presença de empates a Equação original 2.13, mesmo que de forma mínima (quando na presença de poucos empates), aumenta o valor final absoluto do coeficiente trazendo um valor incorreto (SHESKIN, 2007). O procedimento para fazer a correção segue as etapas a seguir:

1. Calcular valores para T_x e T_y, que representam os empates nos *rankings* x e y. Aplica-se a Equação para $T_x = \sum_{i=1}^{s}(t_{i(x)}^3 - t_{i(x)})$, na qual é feito o somatório do conjunto de observações com mesmo valor de *ranking*. A Equação para T_y é a mesma que para T_x.

2. Calcula-se $\sum x^2$ e $\sum y^2$, com as Equações $\frac{n^3-n-T_x}{12}$ e $\frac{n^3-n-T_y}{12}$, respectivamente.

3. Por fim, a Equação que faz a correção para classificações com postos empatados é: $\frac{\sum x^2+\sum y^2-\sum d^2}{2\sqrt{\sum x^2 \sum y^2}}$, onde $\sum d^2$ é o somatório das diferenças como apresentado na Equação 2.13.

É importante ressaltar que quando não há presença de empates nos dados o valor obtido como Coeficiente de Spearman é o mesmo obtido aplicando o Coeficiente de correlação de Pearson, vastamente utilizado e conhecido na literatura, similaridade que se dá justamente pelo Coeficiente de Spearman ser derivado do Coeficiente de Pearson[2]. Na presença de empates, de acordo com Howell (2006), o Coeficiente de Pearson será equivalente ao Coeficiente de Spearman Corrigido, sendo mais conveniente (pelo menor número de cálculos) usar o Coeficiente de Pearson, levando em conta que os dados estejam previamente na forma de *rankings*.

2.2 Coeficiente de correlação de Kendall

O coeficiente τ de Kendall possui função similar ao de Spearman, porém apresenta algumas vantagens em casos específicos. As vantagens são observadas no caso de: (i)

[2]O Coeficiente de Pearson é definido pela Equação r = $\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2(Y_i-\overline{Y})^2}}$, onde $\overline{X}$ e $\overline{Y}$ são as médias aritméticas dos conjuntos.

ter conjuntos de dados com amostras de dimensão muito reduzida e valores repetidos; (ii) possibilidade de generalização do coeficiente τ de Kendall para correlações parciais que são correlações medidas entre duas variáveis após remoção do efeito de uma possível terceira variável sobre ambas.

Um exemplo do surgimento de uma terceira variável na análise de correlação entre um dado x e um dado y ocorre quando é feito um estudo da relação entre o domínio de linguagem e a altura de crianças em idade escolar. Certamente existe relação entre essas variáveis, todavia, uma não é diretamente consequência da outra e sim estão relacionadas por uma terceira variável, que nesse caso é a idade das crianças.

Um ponto a ressaltar é o fato de que tanto o coeficiente de Spearman como de Kendall possuem o mesmo objetivo, medir a associação entre variáveis, porém a forma de atingir é distinta em cada um dos métodos. A interpretação e a impossibilidade de comparação entre os valores finais obtidos nos dois métodos fazem necessário optar por um deles, ou usar ambos, porém analisando cada um de forma isolada (HALL, 2004).

O coeficiente τ de Kendall é calculado através da Equação 2.14, na qual n é o número de itens, e P é o somatório, sobre todos os itens, dos itens classificados depois de um determinado item em ambos os *rankings*. P pode ser interpretado como o número de pares concordantes. O denominador na definição de τ pode ser visto como o número total de pares de itens. Portanto, um valor alto de P significa que a maioria dos pares são concordantes, indicando consistência entre dois *rankings*.

$$\tau = \frac{2P}{\frac{1}{2}n(n-1)} - 1 = \frac{4P}{n(n-1)} - 1 \tag{2.14}$$

No caso de haver muitos empates entre os pares, estes não podem ser considerados concordantes ou discordantes e o número total de pares (no denominador) deve ser ajustado de forma a minimizar este impacto. Na literatura, existem as variações de τ, considerando-se empatado um par $(x_i, y_i), (x_j, y_j)$ quando $x_i = x_j$ ou $y_i = y_j$. Na constatação de empates nos dados, o coeficiente deve ser modificado de maneira a manter o intervalo de -1 e +1. A descrição a seguir engloba os três tipos de coeficientes τ disponíveis:

τ_a **(Tau** a**)** Testa o grau de associação das tabulações quando ambas variáveis são medidas em um nível ordinal, mas não faz ajustes para empates;

τ_b **(Tau** b**)** Semelhante ao Tau a, mas faz ajustes para pares com empates e é mais indicado para tabelas quadradas;

τ_c **(Tau** c**)** Semelhante ao Tau b em sua definição, porém é indicado ao trabalhar com tabelas retangulares.

Para o cálculo do coeficiente τ_b, é usada a Equação 2.15, onde N_s é o número de pares iguais, N_d é o número de pares diferentes, T_x é o número de pares empatados na variável independente (x) e T_y é o número de pares empatados na variável dependente (y).

$$\tau_b = \frac{N_s - N_d}{\sqrt{(N_s + N_d + T_x)(N_s + N_d + T_y)}} \tag{2.15}$$

No caso do coeficiente τ_c, é usada a Equação 2.16, onde N é o número total de casos, m é o valor mínimo, o que for menor do número de linhas ou do número de colunas.

$$\tau_c = \frac{N_s - N_d}{\frac{1}{2}N^2[(m-1)/m]} \tag{2.16}$$

Da mesma forma que no ρ de Spearman, os valores obtidos com τ de Kendall variam de -1 (100% associação negativa, ou inversão perfeita) até +1 (100% associação positiva, concordância perfeita). Quando o resultado é o valor zero, esse indica a falta de associação entre as variáveis.

2.3 Normalized Discounted Cumulative Gain (NDCG)

As métricas *Discounted Cumulative Gain* (DCG) e o *Normalized Discounted Cumulative Gain* (NDCG) têm sido aceitas como bons indicadores para avaliação de sistemas de classificação de documentos. O NDCG é utilizado em situações nas quais não se utiliza a noção binária $0, 1$ para a relevância de documentos (JARVELIN; KEKALAINEN, 2000) (JARVELIN; KEKALAINEN, 2002).

A métrica DCG parte da premissa de que documentos altamente relevantes são mais úteis que documentos pouco relevantes, sendo assim, quanto mais distante a classificação de um documento relevante, menos útil ele será para o usuário, por ele sempre acessar apenas os primeiros resultados apresentados nos motores de busca (WIKIPEDIA, 2011). Esse tipo de classificação de relevância é conhecida por *graded relevance*[3] e contrasta com a classificação binária (documento relevante ou não relevante), sendo assim mais completa por ter um poder maior de classificação.

Para entendermos o DCG devemos nos remeter a sua métrica de origem, chamada de Ganho Acumulado (CG). Esta métrica consiste na acumulação do valor de relevância dos itens recomendados, ou seja, o valor real da classificação atribuída pelo usuário. Desse modo, quanto maior o ganho acumulado melhor será a lista de resultados do ponto de vista do utilizador. A Equação 2.17 apresenta a fórmula do CG numa posição p de um *ranking*:

$$CG_p = \sum_{p}^{i=1} rel_i \tag{2.17}$$

onde p representa o número de itens considerados na lista de itens recomendados e rel_i a clasificação de relevância dada pelo usuário ao item da posição i.

O CG não leva em conta a ordenação dos primeiros p itens, valorizando igualmente o primeiro e o último item da lista. O valor de CG obtido não é, portanto, afetado por mudanças na ordenação dos resultado. Para resolver essa limitação foi proposto o DCG, que reduz (penaliza) a relevância de cada item com base em uma função logarítmica, para assim valorizar mais os itens que aparecem no topo da listagem. A Equação 2.18 define o DCG, no qual a base do $\log$ a ser usada é normalmente 2:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log(1+i)} \tag{2.18}$$

onde k é o número total de documentos retornados e rel_i é o rótulo de relevância designado ao enésimo documento retornado.

[3] *Graded relevance* é um tipo de escala onde existem diferentes níveis de relevância como, por exemplo: documento muito relevante, documento razoavelmente relevante, documento marginalmente (pouco) relevante (SORMUNEN, 2002).

A avaliação dos resultados obtidos com o DCG pode ser obtida através da sua comparação com uma lista ordenada de melhor resultado teórico possível (MRTP). Essa lista é definida levando em conta os índices de relevância utilizados na coleção de referência. Um exemplo seria classificar documentos relevantes em uma escala em que são utilizados os valores: 3 (muito relevante), 2 (relevante), 1 (pouco relevante), 0 (irrelevante). Portanto, uma lista ordenada do melhor resultado teórico seria $MRTP = \{3, 3, 3, 2, 2, 2, 2, 1, 1, 1, 0, 0, ...\}$.

Nota-se que o valor dos ganhos é obtido através de um somatório, sendo a avaliação isolada do valor de ganho uma forma pouco útil de aferir se uma lista de itens gerada é relevante para o usuário. Para tornar a avaliação mais concisa, é necessário comparar o valor do ganho com um valor de ganho ideal (ou MRTP), levando a uma razão entre DCG obtido e o DCG ideal. Essa razão resulta na Equação 2.19, que normaliza a métrica em um valor entre 0.0 e 1.0, sendo que 1.0 representa correspondência total entre o *ranking* gerado e o melhor *ranking* possível para este determinado conjunto de dados.

Esse tipo de métrica não possui significância estatística e não há um método formal de averiguação de seus resultados. Serve como uma ferramenta adicional para indicar o quão os resultados classificados usando determinado método estão próximos à melhor ordenação possível em que os resultados mais relevantes aparecem nos primeiros postos.

$$NDCG_p = \frac{DCG_p}{DCG\_ideal_p} \tag{2.19}$$

ANEXO VALORES CRÍTICOS PARA ρ DE SPEARMAN

Valores críticos do coeficiente ρ de Spearman para probabilidades bicaudais e unicaudais, $\alpha(2)$ e $\alpha(1)$, respectivamente

α(2)	0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
α(1)	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.005
n									
4	0.600	1.000	1.000						
5	0.500	0.800	0.900	1.000	1.000				
6	0.371	0.657	0.829	0.886	0.943	1.000	1.000		
7	0.321	0.571	0.714	0.786	0.893	0.929	0.964	1.000	1.000
8	0.310	0.524	0.643	0.738	0.833	0.881	0.905	0.952	0.976
9	0.267	0.483	0.600	0.700	0.783	0.833	0.867	0.917	0.933
10	0.248	0.455	0.564	0.648	0.745	0.794	0.830	0.879	0.903
11	0.236	0.427	0.536	0.618	0.709	0.755	0.800	0.845	0.873
12	0.224	0.406	0.503	0.587	0.671	0.727	0.776	0.825	0.860
13	0.209	0.385	0.484	0.560	0.648	0.703	0.747	0.802	0.835
14	0.200	0.367	0.464	0.538	0.622	0.675	0.723	0.776	0.811
15	0.189	0.354	0.443	0.521	0.604	0.654	0.700	0.754	0.786
16	0.182	0.341	0.429	0.503	0.582	0.635	0.679	0.732	0.765
17	0.176	0.328	0.414	0.485	0.566	0.615	0.662	0.713	0.748
18	0.170	0.317	0.401	0.472	0.550	0.600	0.643	0.695	0.728
19	0.165	0.309	0.391	0.460	0.535	0.584	0.628	0.677	0.712
20	0.161	0.299	0.380	0.447	0.520	0.570	0.612	0.662	0.696
21	0.156	0.292	0.370	0.435	0.508	0.556	0.599	0.648	0.681
22	0.152	0.284	0.361	0.425	0.496	0.544	0.586	0.634	0.667
23	0.148	0.278	0.353	0.415	0.486	0.532	0.573	0.622	0.654
24	0.144	0.271	0.344	0.406	0.476	0.521	0.562	0.610	0.642
25	0.142	0.265	0.337	0.398	0.466	0.511	0.551	0.598	0.630
26	0.138	0.259	0.331	0.390	0.457	0.501	0.541	0.587	0.619
27	0.136	0.255	0.324	0.382	0.448	0.491	0.531	0.577	0.608
28	0.133	0.250	0.317	0.375	0.440	0.483	0.522	0.567	0.598
29	0.130	0.245	0.312	0.368	0.433	0.475	0.513	0.558	0.589
30	0.128	0.240	0.306	0.362	0.425	0.467	0.504	0.549	0.580
31	0.126	0.236	0.301	0.356	0.418	0.459	0.496	0.541	0.571
32	0.124	0.232	0.296	0.350	0.412	0.452	0.489	0.533	0.563
33	0.121	0.229	0.291	0.345	0.405	0.446	0.482	0.525	0.554
34	0.120	0.225	0.287	0.340	0.399	0.439	0.475	0.517	0.547
35	0.118	0.222	0.283	0.335	0.394	0.433	0.468	0.510	0.539
36	0.116	0.219	0.279	0.330	0.388	0.427	0.462	0.504	0.533
37	0.114	0.216	0.275	0.325	0.383	0.421	0.456	0.497	0.526
38	0.113	0.212	0.271	0.321	0.378	0.415	0.450	0.491	0.519
39	0.111	0.210	0.267	0.317	0.373	0.410	0.444	0.485	0.513
40	0.110	0.207	0.264	0.313	0.368	0.405	0.439	0.479	0.507

Fonte: (ZAR, 1972)