UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS DA EDUCAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

Areli Andreia dos Santos

***F-GROUP*: Um *framework* para aplicação de Indicadores de Produtividade Científica a Grupos de Pesquisadores, com ênfase na colaboração**

Florianópolis
2024

Areli Andreia dos Santos

F-GROUP**:** Um *framework* para aplicação de Indicadores de Produtividade Científica a Grupos de Pesquisadores, com ênfase na colaboração

Tese submetida ao Programa de Pós-Graduação em Ciência da informação da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Doutora em Ciência da Informação, área de concentração Gestão da Informação, linha de pesquisa Informação, Gestão e Tecnologia.

Orientador(a): Prof. Moisés Lima Dutra, Dr.

Florianópolis

2024

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Santos, Areli Andreia dos
F-GROUP : Um framework para aplicação de Indicadores de Produtividade Científica a Grupos de Pesquisadores, com ênfase na colaboração /Areli Andreia dos Santos ; orientador, Moisés Lima Dutra, 2024.
210 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro de Ciências da Educação, Programa de Pós-Graduação em Ciência da Informação, Florianópolis, 2024.

Inclui referências.

1. Ciência da Informação. 2. Ciência da Informação. 3. Avaliação da Pesquisa Científica. 4. grupo de pesquisadores. 5. índice h. I. Dutra, Moisés Lima. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Informação. III. Título.

Areli Andreia dos Santos

***F-GROUP*: Um *framework* para aplicação de Indicadores de Produtividade Científica a Grupos de Pesquisadores, com ênfase na colaboração**

O presente trabalho em nível de Doutorado foi avaliado e aprovado, por banca examinadora composta pelos seguintes membros:

Prof. Rogério Mugnaini, Dr.
Universidade de São Paulo (USP)

Prof. Edgar Bisset Alvarez, Dr.
Universidade Federal de Santa Catarina

Prof. Douglas Dyllon Jeronimo de Macedo, Dr.
Universidade Federal de Santa Catarina

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de Doutora em Ciência da Informação.

ICP-Edu
Documento assinado digitalmente
Edgar Bisset Alvarez
Data: 07/05/2024 08:55:15-0300
CPF: ***.516.228-**
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de Pós-Graduação

ICP-Edu
Documento assinado digitalmente
Moises Lima Dutra
Data: 07/05/2024 08:04:37-0300
CPF: ***.825.889-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Moisés Lima Dutra Dr.
Orientador(a)

Florianópolis, 2024.

Dedico este trabalho às meninas Naiany, Joana e Laura.

AGRADECIMENTOS

Grande parte deste trabalho foi desenvolvido durante a pandemia de COVID-19, em isolamento. Então, primeiramente gostaria de agradecer a Deus, por estar viva. Gostaria também de agradecer aos cientistas que se dedicaram na busca pela vacina, por permitirem que na conclusão desse trabalho eu possa abraçar meus entes mais queridos e agradecê-los pessoalmente.

Gostaria de agradecer à minha família, por sempre me dar o incentivo necessário para que eu continue meus estudos, em especial minha mãe, a professora Andreia e meu pai, Janecir. Agradecer meu familiar mais próximo, meu companheiro André, que me incentivou a iniciar, e, insistentemente, a concluir esse trabalho. Gostaria de agradecer aos meus sogros, por me apoiarem e me incentivarem nesses momentos tão turbulentos que todos nós passamos nos últimos quatro anos.

Meu orientador, Prof. Moisés, obrigada por toda paciência, pelas reuniões de orientação que sempre me acalmavam e me direcionavam para a melhor realização deste trabalho. Obrigada por toda dedicação nas entregas de artigos e por todos os ensinamentos. É um prazer poder trabalhar sob sua orientação.

Aos membros da banca, que com muito respeito me direcionaram, já na qualificação, para as melhorias necessárias neste trabalho, muito obrigada. Com muito carinho agradeço ao PGCIN, especialmente ao Samuel Pereira Marcolin, nosso secretário de curso sempre prestativo, e ao Coordenador de Curso, Prof. Edgar Bisset Alvarez, que conduz de forma tão proativa esse excelente programa de pós-graduação.

Aos professores do PGCIN meu agradecimento especial. As aulas que começaram presenciais, de forma abrupta, precisaram ser realizadas de maneira remota. A forma de ensino, mesmo adaptada, permitiu adquirir conhecimentos e habilidades de forma muito rica. Ver os professores e colegas, mesmo pela tela do computador, e aprender sobre assuntos tão diversos e interessantes foi bálsamo em meio ao caos. Muito obrigada aos professores do PGCIN com os quais tive o prazer de ter aula: Adilson Luiz Pinto, Ana Clara Cândido, Camila Monteiro de Barros, Cezar Karpinski, Douglas Dyllon Jeronimo De Macedo, Eliana Maria dos Santos Bahia Jacintho, Gregorio Varvakis, Luciane Paula Vital, Gustavo Medeiros de Araújo, Márcio Matias, Moisés Lima Dutra e Rosangela Schwarz Rodrigues.

Gostaria também de agradecer aos meus colegas de Doutorado. Sempre tão queridos e prestativos, mais um motivo para que eu possa falar com tanto carinho do PGCIN. Ao meu colega de trabalho, por algum tempo meu chefe também, e colega de Doutorado André Fabiano Dyck, que me apresentou o Prof. Moisés como orientador, muito obrigada. À Paty, Amábile, Eugênio, Fernanda, Edson, Camillo, Paula e Zaina, pelas trocas e parceria nos nossos trabalhos acadêmicos.

Muito obrigada Renata, Camilla e Rubens, que realizaram seus cursos de graduação na área de Biblioteconomia e Ciência da Informação, e que me apresentaram o PGCIN. Em especial à Renata, mestra em Ciência da Informação que me deu os caminhos necessários para aprovação no programa. Essas três pessoas são profissionais da informação que tive o prazer de poder trabalhar em conjunto e que me fazem hoje ter muito orgulho e carinho ao dizer que sou também uma cientista da informação.

Gostaria ainda de agradecer aos meus colegas na Coordenadoria de Integração de Sistemas e Administração de Dados – CISAD/SeTIC. Davi, Giovani, Leandro, Roque e Renato que seguraram as pontas nos meus momentos de ausência e me auxiliaram com o ambiente necessário para que eu pudesse concluir esta tese. Gostaria de agradecer à SeTIC, na figura do superintendente Bruno Carlo Celeguim de Amattos, por todas as oportunidades e parceria para meu desenvolvimento pessoal e profissional no decorrer da minha formação.

De coração: Claudinha, Gabi, Rayssa, Elanne, Melise, e Camila, vocês tornaram a SeTIC um ambiente de acolhimento diário e apoio nos momentos mais necessários no decorrer desses quatro anos de doutorado. Agradeço também às demais colegas de trabalho da SeTIC por demonstrarem, com exemplo, que podemos fazer sempre mais e melhor, conciliando a carreira profissional, acadêmica e pessoal.

Obrigada às meninas da minha vida, Naiany, Joana e Laura. Estar com vocês, brincar e cuidar de vocês é o que muitas vezes me deu fôlego para continuar nessa jornada. Querer me tornar um exemplo próximo e singelo de mulher na ciência foi, por muitas vezes, minha motivação, obrigada minhas afilhadas e minha irmã.

Um agradecimento muito especial à essa instituição que é minha segunda casa desde 2009. Na UFSC conheci os melhores profissionais, os professores mais dedicados e as muitas outras pessoas que fazem a minha vida, com certeza, melhor. Obrigada UFSC pelo ensino público, gratuito e de qualidade.

“*Everything is sketchy. The world does nothing but sketch.*”

Florence Nightingale

RESUMO

A produção científica pode ser medida de diferentes formas. Essas medidas são úteis para tomada de decisão relativa ao desenvolvimento de carreiras e à distribuição de recursos para pesquisas científicas. Nas últimas duas décadas, diversos trabalhos propuseram índices para medir o quão produtivo e relevante é o trabalho de um pesquisador. Alguns desses índices visam avaliar um único pesquisador e/ou grupos de pesquisadores aplicando a mesma fórmula para ambas as situações. Porém, em vários casos, esta aplicação não se dá de forma direta. A avaliação de grupos de pesquisadores tem sido realizada de forma não padronizada, o que dificulta a comparação entre diferentes pesquisas. O objetivo deste trabalho é propor um *framework* para aplicação de indicadores de produtividade científica a grupos de pesquisadores, com ênfase na colaboração. Primeiramente, através de uma revisão de literatura, foram identificadas as formas de avaliação da produtividade dos grupos de pesquisadores, bem como os indicadores de produtividade científica mais utilizados para avaliar esses grupos. Verificamos, ainda, que não existe uma metodologia consolidada para a aplicação de indicadores de produtividade a grupos de pesquisadores. Consequentemente, a busca por uma forma estabelecida de avaliar grupos de pesquisadores provou-se incipiente, existindo diferentes indicadores e formas de aplicação destes. Após a revisão da literatura, uma análise acerca da atribuição da ideia de valor ao pesquisador e sua relação com os indicadores de produtividade foi realizada. O intuito desta análise foi o de explorar o problema da avaliação de pesquisadores sob a ótica da sociologia das profissões, trazendo correlações do conceito de valor de Csillag, com a profissão do pesquisador, a partir de textos de Freidson e Dubar, dentre outros autores. A aplicação do índice h em grupos de pesquisadores é apresentada como subsídio para ilustrar os problemas de seleção e agregação de artigos ao avaliar esses grupos. Desta forma, propomos neste trabalho uma forma de selecionar e agregar artigos ao avaliar grupos de pesquisadores, denominada *IN-GROUP*. Esta forma de seleção e agregação prioriza os trabalhos realizados em conjunto pelo grupo que está sendo avaliado. Neste trabalho, propomos também o *F-GROUP*, um *framework* que visa padronizar a avaliação de grupos de pesquisadores. No Framework *F-GROUP* consideramos aspectos relacionados à caracterização dos grupos, diferentes formas de agregação e contagem dos artigos, chegando à avaliação através de indicadores para grupos de pesquisadores. Adicionalmente, apresentamos diferentes possibilidades de utilização da abordagem *IN-GROUP* e do *framework F-GROUP*, com casos de uso e experimentos com dados reais. Duas formas de agregação foram exploradas com dados reais extraídos da *Web Of Science*, na área de Biblioteconomia e Ciência da Informação nos últimos 10 anos. Desta forma, realizamos análises comparativas através de *rankings* com diferentes indicadores para grupos de países e universidades. Como conclusão, podemos destacar que os grupos formados pelos autores nos trabalhos de Ciência da Informação e Biblioteconomia nos últimos dez anos é bastante heterogêneo. Observamos ainda que os artigos realizados por mais de um autor apresentam uma média de citações mais elevada. EUA e China lideram os *rankings* de indicadores totais de artigos, citações e autores, bem como de aplicação do índice h em grupos formados por países. Universidades americanas e chinesas também lideram vários *rankings* de diferentes indicadores.

Palavras-chave: Índice, Indicador, Citação, Produtividade, Publicação, Colaboração, Grupo, Bibliometria.

ABSTRACT

Several metrics are used to assess the productivity of scientific research. These metrics are useful for various management decisions, such as career development and resource allocation for scientific research. In the last two decades, several studies have suggested indices to measure how productive and relevant a researcher's work is. Some of these indices aim to evaluate a single researcher and groups of researchers, applying the same formula for both situations. However, in many cases, this application is not straightforward. The evaluation of groups of researchers has been done in a non-standard way, which makes it difficult to compare different studies. The objective of this work is to propose a framework to help the evaluation of groups of scientific researchers, focusing on the selection and aggregation of articles. First, through a literature review, we recognize ways to evaluate the productivity of groups of researchers, then, we identify the main scientific productivity indicators used to evaluate these groups. The search aimed to find a consolidated methodology to determine the productivity indicators in groups of researchers. As a result, the search for an established way for evaluating groups of researchers was incipient, with different indexes and different ways of applying them. After reviewing the literature, the idea of value attribution for researchers and its relationship with productivity indicators was examined. The purpose of this analysis was to explore the problem of evaluating researchers from the perspective of the sociology of professions, bringing correlations between Csillag's concept of value and the researcher's profession, based on texts by Freidson and Dubar, among other authors. The application of the h index in groups of researchers was presented as a subsidy to illustrate the issues of selecting and aggregating articles when evaluating these groups. In this work, we present a form of selecting and aggregating articles when evaluating groups of researchers, called *IN-GROUP*. This new form to select and aggregate papers prioritizes a group's production cooperatively. We hereby propose the F-GROUP, a framework to standardize the evaluation of groups of researchers, addressing aspects related to the characterization of groups, through forms of aggregation and counting of articles, leading to reaching evaluation through indicators for groups of researchers. Additionally, we present use cases and experiments with real data to illustrate different possibilities for using the *IN-GROUP* approach and the *F-GROUP* framework. The framework and two forms of aggregation are analyzed with real data extracted from the Web Of Science, in Library and Information Science over the last 10 years. Comparative analyses were conducted through rankings with different indicators for groups of countries and universities. For conclusions, we highlight that groups formed by authors in Information Science and Library Science works in the last ten years are quite heterogeneous, in addition, we observed that articles written by more than one author have a greater number of citations. The USA and China lead the rankings when using total indicators of articles, citations and authors, as well as when using the h index in groups formed by countries. American and Chinese universities also lead the rankings when exploring groups formed by universities.

Keywords: Index, Indicator, Citation, Productivity, Publication, Collaboration, Group, Bibliometric.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACM	*Association for Computing Machinery*
BMLR	*Bayesian Multilevel Logistic Regression*
BRAPCI	Base de Dados em Ciência da Informação
BRI	*The Belt and Road Initiative*
BRICS	Brasil, Rússia, China, Índia e África do Sul
CERN	*European Organization for Nuclear Research*
CI	Ciência da Informação
EMI	Estudos Métricos da Informação
ENANCIB	Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação
FSS	*Fractional Scientific Strength*
FSS^N	*Normalized Fractional Scientific Strength*
G7	*Group of Seven*
GC	Gestão do Conhecimento
GI	Gestão da Informação
GT4	Grupo de Trabalho 4
IC	Índice de Colaboração
ICER	*Conference on International Computing Education Research*
IEEE	*Institute of Electrical and Electronics Engineers*
ITiCSE	*Innovation and Technology in Computer Science Education*
JIF	Journal Impact Factor
LISA	*Library and Information Science Abstracts*
NSFC	Fundação Nacional de Ciências Naturais da China
RQ	*Research Question*
RSL	Revisão Sistemática de Literatura
SIGCSE	Special Interest Group – Computer Science Education
UB	Universidade de Belgrado
UPV	Universidade Politécnica de Valencia
UFV	Universidade Federal de Viçosa
UNESP	Universidade Estadual Paulista
USP	Universidade de São Paulo
WoS	*Web of Science*

SUMÁRIO

1 INTRODUÇÃO

O advento da *Internet* e a consequente ampliação do acesso aos textos científicos, sejam estes livros, artigos ou outra forma textual, facilitou a capacidade de medir a produção científica (BOSSY, 1995). Com o passar dos anos e a estruturação das grandes bases de artigos científicos *online*, em conjunto com a crescente evolução dos recursos computacionais, foi possível automatizar a coleta e a divulgação de medidas de produção científica. Atualmente, ao se publicar de forma *online* um novo artigo científico, por exemplo, muitos indicadores de produção científica são atualizados de forma automática. Esta capacidade ampliada e automatizada nos processos de contagem no número de publicações e citações, permitiu uma ampliação dos Estudos Métricos da Informação (EMI). Destacam-se, dentre os EMI impactados, a Bibliometria, mais generalista avaliando as produções bibliográficas, a Cienciometria, mais focada nas métricas relacionadas à produção científica e a Altmetria, que avalia o impacto da produção científica na *Web*, com números de cliques e *downloads*, por exemplo.

Dado o cenário com acesso a publicações científicas, como artigos e periódicos, facilitado pelo uso da *Internet*, diversos índices foram propostos para comparar e avaliar a relevância de uma publicação e seus impactos na sociedade (EGGHE, 2006; KOSMULSKI, 2005; ZHANG, 2009). Esses índices são úteis especificamente para avaliar aspectos quantitativos, medindo o quanto um pesquisador, um grupo de pesquisadores e até mesmo universidades, instituições de pesquisa ou países produzem em termos de ciência. Um dos principais usos desses dados estatísticos é para uma distribuição mais justa de recursos entre as entidades científicas (JOSHI, 2014; ROEMER; BORCHARDT, 2005).

Expressar quantitativa e qualitativamente um valor intangível como a produtividade científica não é uma tarefa fácil e sempre haverá distorções da realidade. A ideia mais básica é usar o número de artigos publicados, mas ele mede apenas a quantidade, não a qualidade destes trabalhos. Outra métrica comum é o número de citações, que visa expressar o quanto um artigo científico é relevante para

outros pesquisadores. O número de citações é a métrica mais relevante e tradicional para avaliar o impacto e a relevância de um artigo científico (ARAÚJO, 2006).

A fim de combinar quantidade e qualidade, vários índices foram propostos usando o número de artigos e citações (por exemplo: *índice h*, *índice e, índice i10*, entre outros), com o objetivo de medir não apenas a quantidade, mas o quão útil cada artigo científico é. A combinação do número de artigos e de citações é bastante comum e largamente utilizada, sendo utilizada também para medir o Fator de Impacto (ARAÚJO, 2006). Com disso, a autocitação com intuito de aumentar o valor medido por esses índices se tornou um problema. Desde então, outros índices foram propostos para evitar uma distorção provocada pela autocitação excessiva (EGGHE, 2006; FLATT; BLASIMME; VAYENA, 2017, 2017; KOSMULSKI, 2005; ZHANG, 2009). Mais recentemente, aspectos que consideram o ambiente *online* emergiram, trazendo novas métricas como cliques e *downloads*, sendo esse conjunto chamado de Altmetria (TORRES-SALINAS; ROBINSON-GARCIA; JIMÉNEZ-CONTRERAS, 2016). Ainda assim, o número de artigos, citações e colaborações são as métricas mais relevantes para medir a produção e relevância na ciência.

Em relação à forma como os índices são calculados existem alguns aspectos que recebem críticas ao longo dos anos (JOSHI, 2014). Um dos primeiros problemas está relacionado ao aspecto "qualidade *versus* quantidade". Autores que se concentram em artigos de revisão de literatura podem ter um número maior de citações do que outros que estão produzindo conhecimento inédito. O índice mais relevante para contornar esse problema é o *Crown*, que considera apenas os trabalhos identificados como equivalentes por determinadas características, como: área de aplicação, tipo de publicação, período de publicação, entre outros aspectos, para permitir uma comparação equilibrada. Esse indicador, no entanto, utiliza somente o número de citações como critério de avaliação.

Outra questão relacionada aos indicadores baseados no número de citações é que um autor pode ter um valor muito alto desse indicador mesmo após anos de aposentadoria, enquanto novos pesquisadores com novas contribuições podem ser prejudicados, mesmo apresentando qualidade e quantidade em uma carreira ainda curta. Consequentemente, alguns índices foram propostos para contornar esse problema, por meio do uso de um intervalo de tempo para levar em conta os artigos, como a métrica AIF (*Author Impact Factor* ou Fator de Impacto do Autor) (PAN;

FORTUNATO, 2015). Como a citação se tornou um elemento superestimado para avaliação da pesquisa dentre os pesquisadores, surgiu o problema do exagero na autocitação. O uso da autocitação para aumentar esse indicador também se tornou uma questão relevante (FLATT; BLASIMME; VAYENA, 2017). Alguns índices foram propostos aplicando uma transformação matemática ao *índice h*, como considerar o número de citações ao quadrado (KOSMULSKI, 2005), na tentativa de relativizar o problema gerado pelo excesso de autocitações.

Enquanto alguns dos índices propostos são usados para calcular dados para um pesquisador individual (HIRSCH, 2005; ZHANG, 2009), outros são usados no cálculo para um grupo de pesquisadores (ROUSSEAU; YANG; YUE, 2010; TORRES-SALINAS; ROBINSON-GARCIA; JIMÉNEZ-CONTRERAS, 2016; VALLES *et al*., 2020). Este trabalho considera um grupo de pesquisadores como qualquer grupo formado por dois ou mais pesquisadores, que podem ser alunos, professores ou outros acadêmicos, onde seja necessária uma medição da produtividade deste grupo. Este ambiente de grupo pode variar: desde um pequeno conjunto composto por alguns colegas pesquisadores que estão se candidatando a um projeto de pesquisa local, a um grande grupo de pesquisa trabalhando em um projeto internacional com muitos pesquisadores e recursos de diferentes países. Pode ainda representar instituições e países como um todo.

Algumas pesquisas reuniram os índices utilizados para mensurar a produtividade de pesquisadores (AGARWAL *et al.*, 2016; GASPARYAN *et al.*, 2018; JOSHI, 2014). Nestes trabalhos, foram realizadas análises de uma área de conhecimento específica, comparando diferentes índices (AGARWAL *et al.*, 2016; GASPARYAN *et al.*, 2018). Outros trabalhos são mais generalistas, como (JOSHI, 2014), por exemplo, que foca em métricas para analisar artigos e pesquisadores sem especificar uma área ou campo. No entanto, estes trabalhos não apresentam métricas para mensurar a produtividade de grupo de pesquisadores.

O que existe de mais próximo de uma medida que combine número de citações e artigos para medir a produtividade da pesquisa em grupos são as variações de índice h. Alguns trabalhos propuseram a aplicação do *índice h* para um grupo de pesquisadores (KHAN *et al.*, 2013; MITRA, 2006; RAD *et al*., 2010; SCHUBERT, 2007). Estes trabalhos usaram uma transformação, como a média, por exemplo, para aplicar o *índice h* em uma análise de um campo específico ou para comparar

diferentes instituições. Esta transformação do índice h para avaliar grupos pode utilizar médias e totais dentre outras transformações para atribuir a um grupo de pesquisadores um valor único do índice.

Um estudo recente aponta que as temáticas “Medição de impacto de pesquisa e colaboração em pesquisa”, “Rede social” e “Métricas de pesquisa e estudos baseados em citações” são as áreas emergentes da pesquisa em Ciência da Informação (SAHOO *et al.*, 2020). Ou seja, a temática de análise de produção científica para grupos demonstra ser um tema bastante atual e com diversos campos a serem explorados. As métricas e indicadores para avaliar grupos de pesquisadores são escassas e não é trivial encontrá-las na literatura. O problema de analisar a produtividade de grupos de pesquisadores ainda é bastante abstrato, fazendo-se assim necessária uma consolidação dos conceitos já discutidos nesta temática. A aplicação do índice h em grupos ocorre de forma não padronizada. A seleção e agregação dos artigos que representam o grupo também ocorre de variadas formas. Esta falta de padronização impede a comparação entre diferentes estudos.

Na próxima seção são detalhados alguns dos questionamentos atuais e apresentada a questão de pesquisa deste trabalho.

1.1 PROBLEMA E QUESTÃO DE PESQUISA

Mesmo as formas estabelecidas para avaliar grupos de pesquisadores, como número de artigos e citações, levantam dúvidas acerca da forma da avaliação dos grupos de pesquisadores. Dentre essas dúvidas, muitas delas emergem já na seleção dos artigos: quais artigos devem ser selecionados para medir o quão relevante é a produtividade da pesquisa científica de um grupo? A seleção dos artigos pode levar em conta diversos critérios como as diferentes instituições envolvidas no trabalho, o número e a importância dos autores envolvidos na publicação, se todos os artigos desenvolvidos por cada membro do grupo devem ser considerados individualmente.

Figura 1 - Problema de Pesquisa

Fonte: Elaborado pela autora

A Figura 1 apresenta de forma resumida o problema de pesquisa. Dado um grupo de pesquisadores (a), são buscadas as publicações do grupo a ser avaliado (b). Quais artigos devem ser selecionados (c) para que os indicadores representem a produção científica do grupo (d)?

No problema descrito em (c), perguntas elementares surgem, como: *Devem ser considerados todos os artigos de um autor que trabalhou em várias instituições, ou apenas os que foram publicados durante o período de trabalho em sua instituição atual?* Os pesquisadores podem atuar em diferentes locais ao longo da sua carreira. Podem, ainda, ter mais de um vínculo institucional ao mesmo tempo. Como balancear esses vínculos ao analisar a produção de um pesquisador quando o intuito é avaliar o trabalho de um grupo? Mesmo os trabalhos que analisam uma instituição muitas vezes não deixam claro se consideraram os trabalhos dos pesquisadores apenas durante o período em que atuaram na instituição avaliada (ALEIXANDRE et al., 2013b; ÇAKIR et al., 2019; CODINA-CANET, 2012; KUMAR, 2020; PAKKAN et al., 2021; PILCEVIC; JEREMIC; VUJOSEVIC, 2018; ROSAS, 2015). Sendo o vínculo na análise de grupos uma questão ainda em aberto.

O número de autores em cada artigo deve ser considerado? Dependendo da área de pesquisa, da instituição, de questões de avaliação locais e globais, o número de autores nos artigos pode variar bastante. Existem trabalhos que consideram o número de autores na tentativa de mensurar a relevância de um único autor (DE MOYA-ANEGON et al., 2018; SIMON, 2016b). Ao considerar a avaliação da produção em grupos de autores, o número de autores em cada publicação pode influenciar muito no resultado da avaliação, e, portanto, é um aspecto a ser considerado.

Devemos considerar todas as citações em que um dos autores faz parte do grupo? Ou devemos nos concentrar apenas nos artigos em que todos os membros do grupo participam? Essas questões se aplicam aos casos em que o intuito é avaliar um grupo definido de autores, como um laboratório ou grupo de pesquisa, por exemplo. A existência de um(a) pesquisador(a) com índices de produção muito diferentes do restante do grupo pode modificar drasticamente a avaliação do grupo como um todo. Por outro lado, considerar apenas os trabalhos onde todos os membros do grupo participaram, pode desfavorecer a entrada de novos membros.

Em suma, todas essas perguntas que evidenciam o problema da aplicação de indicadores de produção científica em grupos de pesquisadores são apresentadas

através da questão de pesquisa: "Como avaliar quantitativamente e qualitativamente a produção científica de grupos de pesquisadores?". A seguir, são apresentados os objetivos deste trabalho.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Considerando o problema da aplicação de métricas que analisam aspectos quantitativos e qualitativos para um grupo de pesquisadores, o objetivo deste trabalho é propor um *framework* para aplicação de Indicadores de Produtividade Científica a Grupos de Pesquisadores, com ênfase na colaboração.

1.2.2 Objetivos Específicos

Com o intuito de atender o objetivo geral, são elencados os seguintes objetivos específicos:

a) Discutir a relação da ideia de valor do pesquisador e sua relação com os indicadores baseados em quantidade de artigos e citações;
b) Identificar as formas de avaliação existentes para grupos de pesquisadores através de indicadores.
c) Propor novas formas de avaliar grupos de pesquisadores, considerando as lacunas presentes nas formas atuais.
d) Avaliar a aplicabilidade dos dispositivos propostos por meio de casos de uso.
e) Comparar os dispositivos propostos no trabalho com as formas de avaliação consolidadas na literatura.

1.3 JUSTIFICATIVA

A medição da produção científica justifica-se como possível forma de distribuição mais justa de recursos entre as entidades científicas (JOSHI, 2014). As métricas atuais mais famosas focam na produção de um pesquisador, mas a pesquisa

científica é realizada muitas vezes em conjunto, com a colaboração de vários pesquisadores, em grupos.

A atribuição de medidas a grupos já vem sendo explorada sob outras vertentes, como no cálculo do grau de relacionamento entre pessoas através de seus encontros presenciais, monitorados por GPS (SANTOS et al., 2015), tema da dissertação de mestrado da autora. As métricas relacionadas à produção científica podem ser relacionadas aos encontros, aplicando o conceito de que a produção de um artigo pode ser vista como um encontro entre pesquisadores.

Existem diversos trabalhos que mensuram a produção científica em grupos de pesquisadores, a maioria deles com estudos de caso aplicados a países e universidades (ACERO; KLEIN, 2021; ALEIXANDRE et al., 2013a, 2013b; CAKIR et al., 2019; CICERO; MALGARINI, 2020; CODINA-CANET, 2012; GOLICHENKO; MALKOVA, 2017; KUMAR, 2020; LANCHO-BARRANTES; CANTU-ORTIZ, 2020; MESCHINI; ALVES; OLIVEIRA, 2018; MORENO-DELGADO; GORRAIZ; REPISO, 2021; PAKKAN et al., 2021; PILCEVIC; JEREMIC; VUJOSEVIC, 2018; ROSAS, 2015; SAHOO et al., 2019; THOMPSON, 2020; TORRES-PASCUAL; SÁNCHEZ-PÉREZ; ÀVILA-CASTELLS, 2021; YUAN et al., 2018). Existem ainda, em menor número, estudos de caso em conferências (SIMON, 2016b), províncias (ABRAMO; D'ANGELO, 2015) ou eventos (SILVA, 2019), por exemplo. Como indicadores, esses trabalhos utilizam, em sua grande maioria, totais de artigos e totais de citações. É bastante complexo identificar os trabalhos que abordam a mensuração da pesquisa em grupos de pesquisadores. Além disso, existem poucos trabalhos que abordem esta temática de forma direta (WANG et al., 2021).

A combinação de total de artigos como um indicador associado à quantidade, e de citações, como um indicador relacionado à qualidade das produções acadêmicas, é bastante utilizada. Um número elevado de publicações, combinado com um número elevado de citações eleva o *status* de um pesquisador. Consequentemente, as instituições e grupos de pesquisadores também atuam de forma a ampliar seu *status*, incentivando a produção científica e as colaborações. A avaliação da produção que perpassa aspectos sociais têm amplo embasamento na Ciência da Informação (KUMAR et al., 2021; LANCHO-BARRANTES; CANTU-ORTIZ, 2020; VESSURI, 1987; YANG; TANG, 2012).

Uma combinação para total de artigos e de citações foi proposta por Hirsch (2005), o índice h. Hirsch criou este índice para avaliar pesquisadores, ou seja, a produção de um único autor. Em seguida, algumas variações foram propostas para a aplicação do mesmo índice para grupos: i) o *h_1-índice* que considera toda uma instituição ou departamento, ou seja, leva em conta todos os artigos desta instituição/departamento em seu cálculo (MITRA, 2006); ii) a média ou mediana dos índices h dos indivíduos que fazem parte do grupo (KHAN et al., 2013; MUGNAINI; PACKER; MENEGHINI, 2008; RAD et al., 2010) ; iii) o chamado *índice h sucessivo*, que calcula o *índice h* das instituições identificando como principais pesquisadores todos aqueles que possuem o *índice h* maior ou igual a uma pontuação predeterminada (SCHUBERT, 2007).

Desta forma, percebe-se dois pontos principais: a) que a aplicação do índice h para grupos não está padronizada; e b) a avaliação de grupos de pesquisadores em estudos de caso também carece de padronização. Esta falta de padrões dificulta a comparação entre diferentes trabalhos e a realização de estudos mais aprofundados, como meta-análises, por exemplo.

A utilização de uma estrutura que facilite a avaliação de grupos, bem como a comparação entre estudos é de suma relevância para desenvolvimento desta temática. Considerando o conceito de *framework* definido por Macedo e Souza (2023), uma estrutura que possibilite o desenvolvimento de pesquisas, utilizando métricas para avaliar a produtividade de grupos de pesquisadores, pode ser bastante útil para tornar os estudos de casos mais comparáveis.

> Um framework, de maneira geral, é uma estrutura (um esqueleto) que possibilita o desenvolvimento de algo sobre sua base inicial, podendo representar um problema e fornecer a base para a resolução deste problema em um domínio específico. A finalidade dessa estrutura é tornar os resultados de uma pesquisa mais rigorosos e significativos, possibilitar o desenvolvimento de teorias e assegurar a generalização. (MACEDO; SOUZA, 2023, p. 3)

Além da evidente necessidade de padronização na avaliação de grupos, os temas relacionados à colaboração são considerados emergentes na área. Sahoo (2020) apresenta que as temáticas de medição de impacto de pesquisa e colaboração em pesquisa, são apresentadas como tendências na pesquisa em Ciência da Informação. Pesquisas recentes, evidenciam também a temática da Transparência

em Ciência com a proposta de novos modelos para avaliação da ciência (SENA; CARVALHO SEGUNDO; MELO, 2023).

Neste cenário, estudar a avaliação da pesquisa em grupos de pesquisadores pode ser útil para: a) permitir a comparação entre trabalhos similares que avaliam grupos de autores; b) descobrir características dos grupos de autores com produção mais relevante; e c) permitir uma avaliação mais objetiva, utilizando indicadores numéricos.

1.4 RELAÇÃO COM A CI

A origem do termo Cienciometria (também grafado como Cientometria) é controversa. O mais provável é que tenha surgido na Europa, sendo o inglês Derek John de Solla Price associado à criação da Cienciometria (PARRA; COUTINHO; PESSANO, 2019). Para Price, esse campo está focado na análise da dinâmica da atividade científica. A Cienciometria tem a finalidade de "investigar a atividade científica como fenômeno humano, social e mediante parâmetros e indicadores baseados em modelos matemáticos" (PARRA; COUTINHO; PESSANO, 2019 p. 129). Outra definição para o termo é dada por Silva e Bianchi (2001).

> Mensura os métodos e canais para a produção, a comunicação e a colaboração científica nas mais diversas áreas do conhecimento, considerando as características e práticas em pesquisa, bem como as relações e atividades dos cientistas com fins a mapear atividades dos campos científicos e delinear políticas em C&T (SILVA; BIANCHI, 2001, p. 09).

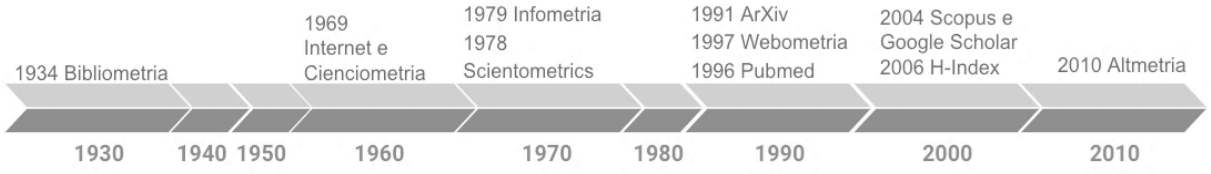
De acordo com as definições apresentadas pode-se destacar a natureza positivista deste campo. Em especial na utilização dos termos “mensura”, “indicadores” e “modelos matemáticos” em ambas as definições. Dessa forma, a Cienciometria está centrada na objetividade auxiliada por instrumentos padronizados e neutros (PARRA; COUTINHO; PESSANO, 2019).

A Cienciometria está inserida na grande área da Ciência da Informação, no tópico de estudos métricos da informação, área que possibilitou a incorporação de bases teóricas e conceituais devido à aproximação com as demandas de organização, representação e padronização dos dados e informações. Essa característica da Cienciometria se encontra nos termos apontados por Pinto e Matias (2012):

> O desenvolvimento, a geração e a análise destes indicadores demandam organização, representação e registros padronizados e adequados do conhecimento para a geração de informações precisas e úteis aos gestores das universidades e das áreas estratégicas de educação e de C&T (PINTO; MATIAS, 2012, p. 13).

Contudo, há estudos que mostram o início da Cienciometria em data anterior à da própria constituição da CI. Essa constatação se encontra na Figura 2, uma linha do tempo que considera o surgimento de conceitos relacionados à Cienciometria e as respectivas datas associadas ao possível surgimento de cada um dos conceitos. Esse recorte temporal tem por base "Genealogia dos Subcampos dos EMI [Estudos Métricos da Informação]" (CURTY; DELBIANCO, 2020).

Figura 2- Linha do tempo de conceitos relacionados à Cienciometria



Fonte: Adaptado pela autora a partir de (CURTY; DELBIANCO, 2020, p.05)

Na Figura 2 são elencados alguns conceitos relacionados à Cienciometria, acompanhados das datas atribuídas ao provável surgimento de cada um dos conceitos. O primeiro é a Bibliometria, com surgimento ocorrido na década de 1930. O ano de 1969 é atribuído à criação da *Internet* e, para Curty e Delbianco (2020), é o ano do surgimento da Cienciometria.

> Cienciometria foi mencionada pela primeira vez em um texto publicado por Nalimov e Mulchenko, em 1969, que sugeria o uso do termo para o estudo quantitativo de todos os aspectos da ciência e da tecnologia, incluindo seus métodos de comunicação e circulação, bem como a evolução e o desdobramento de novos ramos científicos (CURTY; DELBIANCO, 2020, p. 06).

Contudo, os mesmos autores destacam que, de forma efetiva, “A Cienciometria se fortaleceu enquanto campo de estudo a partir do surgimento do periódico *Scientometrics*, em 1978” (CURTY; DELBIANCO, 2020, p.6).

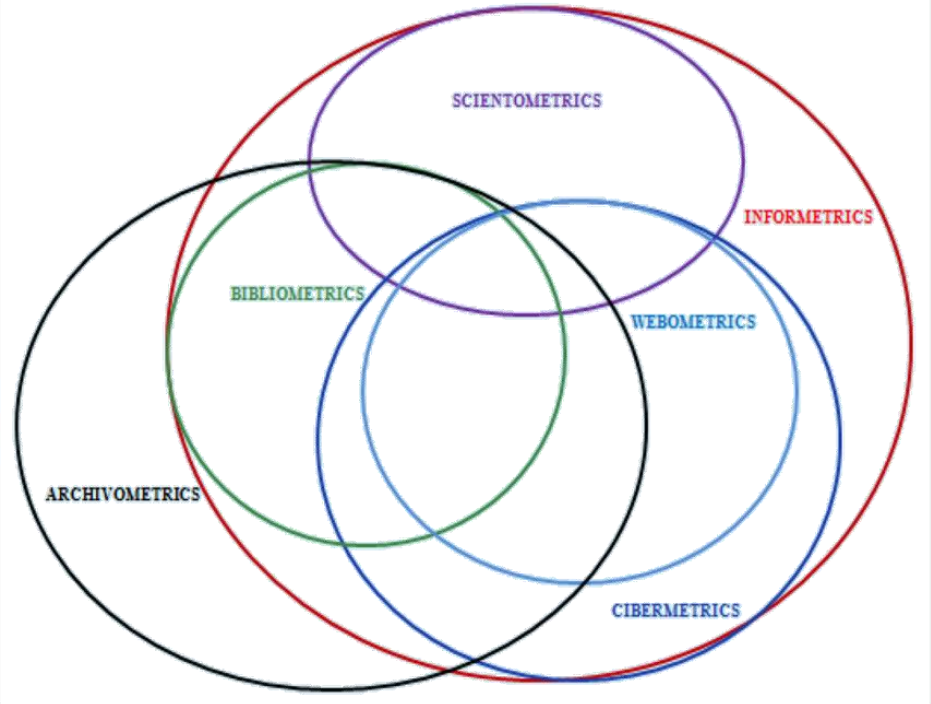
Na linha do tempo, mostra-se que no final da década de 1970 houve a criação da Infometria. Nas décadas 1990 e 2000, a criação de bases de dados como *ArXiv*[1],

[1] https://arxiv.org/

Pubmed[2], *Scopus*[3], além do buscador *Google Scholar*[4]. Consta também a criação do *índice h*, um dos índices mais utilizados no mundo para avaliação de pesquisadores. Em 2010 tem-se a emergência da Altmetria, sendo esse um dos mais recentes fenômenos, que propõe medidas alternativas para avaliar alcance, disseminação e impacto de artigos. A Altmetria está inserida na Cienciometria, por se tratar de uma metodologia para complementar às medidas tradicionais baseadas em citação, com o intuito de medir o alcance da ciência na *Web*.

A Cienciometria deriva da Sociologia da Ciência e sua métrica baseia-se na Cienciologia (PINTO, 2011). Está inserida no contexto dos Estudos Métricos da Informação (EMI), que se relacionam, em última instância, à avaliação da informação. Os EMI englobam ainda informações de natureza social, política e tecnológica sem que haja uma limitação da fonte utilizada. Possui característica interdisciplinar por se relacionar com outras áreas do conhecimento. Mesmo englobando outros documentos, a análise da comunicação científica ocupa uma grande parcela dos estudos métricos da informação (CURTY; DELBIANCO, 2020).

Figura 3 - Diagrama dos estudos métricos da Informação e da Documentação



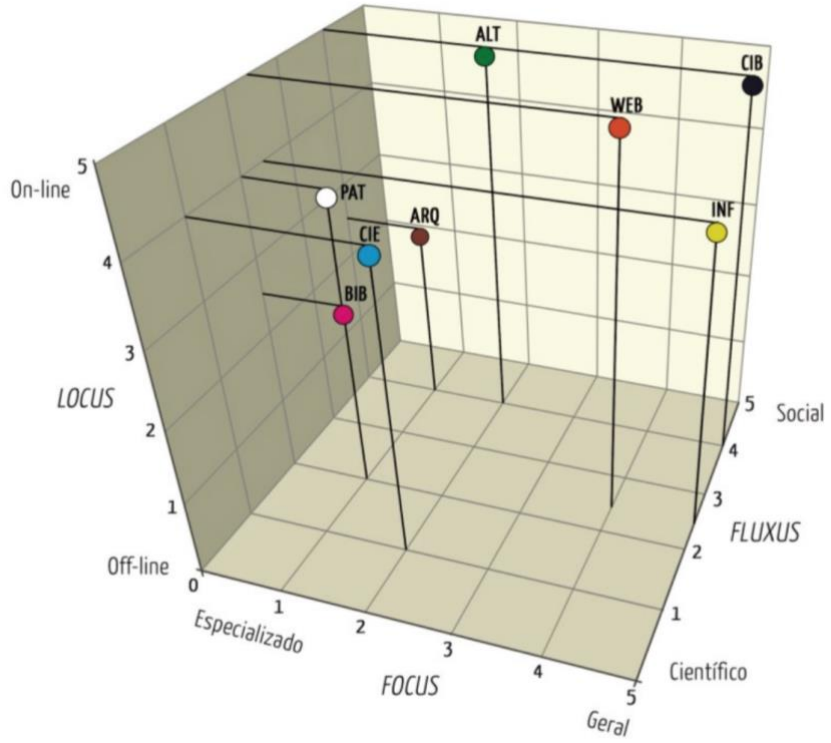
Fonte: (PINTO, 2011, p. 63)

[2] https://pubmed.ncbi.nlm.nih.gov/
[3] https://www.scopus.com/
[4] https://scholar.google.com.br/

A temática dos Estudos Métricos da Informação possui definições consolidadas de acordo com os objetos de estudo. Neste sentido, considerando o diagrama adaptado de Pinto (2011) na Figura 3, o escopo do presente trabalho se encontra na intersecção entre as formas de *SCIENTOMETRICS* e *BIBLIOMETRICS*, consequentemente, a Arquivometria e Informetria também fazem parte do escopo aqui definido. A Informetria, definida como um modo mais abrangente dos estudos métricos da informação e que não se limita ao meio pelo qual a informação é disseminada engloba também a informação científica.

Como destacado, ainda na Figura 2, a Cienciometria tem seus alicerces na Bibliometria. Por isso, é importante destacar a Lei da Produtividade Científica de Autores, proposta por Lotka em 1926. Esta, dentre outras leis, contribuiu de forma significativa para discussões acerca dos estudos métricos da informação. Na Figura 3 é apresentado um diagrama com a interseção dessas áreas. A Figura 4 apresenta uma forma de delimitar os termos associados aos estudos métricos da informação, facilitando a delimitação da Cienciometria (CIE) perante as demais abordagens métricas.

Figura 4- Abordagens métricas e suas dimensões



Fonte: (CURTY; DELBIANCO, 2020, p.16).

Na Figura 4 são apresentadas três dimensões: *Locus* (*online* ou *off-line*); *Focus* (Especializado ou Geral); *Fluxus* (Social ou Científico). Na classificação proposta por Curty e Delbianco (2020), no tocante ao Focus, a Bibliometria (BIB) é mais especializada do que a Cienciometria. Com relação ao *Locus*, a Cienciometria foi classificada mais próxima do *online*, enquanto a Bibliometria ficou bem ao centro indicando que está entre o *online* e o *off-line*. Em relação ao *Fluxus*, a Cienciometria ficou mais relacionada ao Científico, e a Bibliometria ficou mais ao meio, embora com uma demarcação um pouco mais próxima do Científico do que do Social. Desta forma pode-se analisar que, na visão das autoras, a Cienciometria se caracteriza por um viés *online*, científico e especializado.

Diante do cenário atual dos EMI, é preciso refletir sobre a temática da análise de dados e sua relação com a Ciência da Informação e a Cienciometria, levando-se em conta as mudanças ocorridas na chamada Era da Informação, pós-industrial (JAMIL; NEVES, 2000). São considerados, nesta seção, os aspectos positivistas da Cienciometria e sua relação com a análise de dados.

Estamos presenciando uma mudança estrutural na forma de se analisar dados. Anteriormente, a corrente majoritária adotada era a *Data Warehouse*, onde um conjunto de dados do passado era utilizado para tentar predizer tendências no futuro. Com o surgimento do paradigma *Big Data,* neste século, o objetivo passou a ser fornecer um panorama atual, de acordo com múltiplas fontes de dados, ou seja, se obter uma informação mais recente e identificar fatos e tendências enquanto eles ocorrem (AFTAB; SIDDIQUI, 2018). Uma corrente que ganha terreno é a chamada Ciência de Dados, estratégia multidisciplinar que inclui conceitos relacionados a Matemática, Estatística e Computação (em especial, a Inteligência Artificial), e que é aplicada à área com a qual se deseja trazer mais informações.

No aspecto científico, essa mudança recente no campo da análise de dados pode ser melhor compreendida levando-se em consideração a teoria de Popper (1975), no concernente ao rompimento com o indutivismo, que partia de premissas individuais para obter conclusões globais. Pode-se relacionar a perspectiva *Data Warehouse* com o Indutivismo, uma vez que são observados dados do passado para tentar explicar algum fenômeno ou ainda, tentar predizer o futuro. Da mesma forma podemos relacionar a Ciência de Dados ao Falsificacionismo, onde as afirmações são validadas através da apresentação de hipóteses que podem ou não ser verdadeiras.

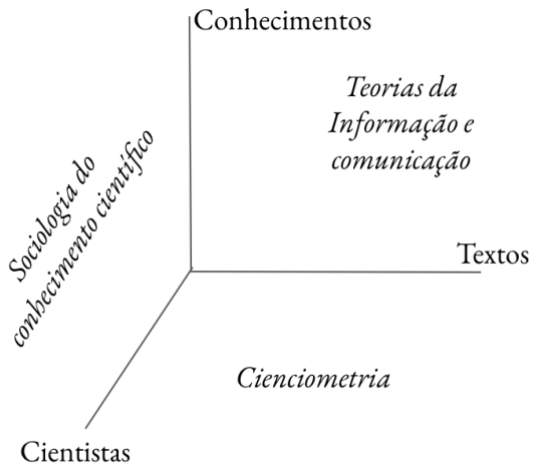
É através do experimento científico que será possível determinar a validade das hipóteses apresentadas.

Popper (1975) delimita o que é científico através da possibilidade de falseabilidade dos fenômenos. Portanto, chama-se Ciência da Informação uma disciplina com base *multi* ou interdisciplinar, cujo objeto de estudo é a Informação. Ou seja, rompe-se, dessa forma, a barreira do tipo de suporte físico onde a informação é armazenada, ou a forma como pode ser recuperada. Assim, a Informação se torna objeto de estudo e fonte de teorias científicas que podem ser falseadas. No entanto, na Cienciometria permanece uma delimitação importante que é a publicação, ou seja, o conhecimento transmitido por meio de textos, independentemente de seu suporte físico. A ciência permanece tendo como recurso essencial a leitura e a escrita de textos científicos. Isso porque, de acordo com Vessuri (1987), a "ciência que não é publicada não existe" (VESSURI, 1987, p. 124).

É neste escopo que se situa a Cienciometria, conforme representação da Figura 5. Neste esquema, apresenta-se a tríade 'Cientistas', 'Textos' e 'Conhecimentos', bem como suas respectivas relações. Sendo definida como Cienciometria a relação entre os cientistas e os textos.

Considerando os conceitos presentes na Figura 5, a Cienciometria é o estudo dos textos científicos. São comuns nestes estudos: número de publicações ao longo dos anos acerca de alguma temática específica; análise do número de citações; formação de indicadores para medir a produção científica de pesquisadores.

Figura 5- O estudo da ciência como um problema multidimensional.



Fonte: (LEYDESDORFF, 2001, p. 4)(tradução nossa).

Concomitantemente, com o surgimento de bases de busca para artigos científicos, também se popularizaram os indicadores de produção científica. Na Figura 6, temos um exemplo da visualização de indicadores de citação de um pesquisador na plataforma *Google Scholar*.

Figura 6- Visualização dos indicadores de um pesquisador.

Fonte: Imagem extraída da plataforma *Google Scholar* (2021).

Sobre a temática tecnológica alinhada à Cienciometria, Cupani (2016) pondera acerca das reflexões de Mumford e o “mito da máquina”. Para os autores, o homem, como ser de dotada capacidade intelectual deve, portanto, dedicar-se mais a atividades intelectuais, em detrimento de atividades mecânicas e fabris. A avaliação de resultados de saída faz parte dos algoritmos de Inteligência Artificial (IA). Essa avaliação ocorre através de indicadores numéricos. Pode-se então traçar uma analogia no sentido de que quanto mais os indicadores de qualidade de produção científica forem aprimorados, mais perto pode-se estar da automatização de algumas tarefas na produção científica. Recentes tecnologias já permitem gerar de forma automatizada o referencial teórico de um trabalho aplicando técnicas de *Text Mining* (FRUNZA *et al.*, 2011). Quanto mais resultados numéricos acerca da qualidade das produções científicas, mais fácil se torna a automatização das atividades relacionadas à produção científica.

Em Silveira e Caregnato (2017) sobre as demarcações epistemológicas dos estudos de citação, vários aspectos teóricos e práticos sobre o fenômeno da citação são levantados, sejam os qualitativos ou quantitativos. Neste estudo são questionadas as "noções de fenômeno e objeto que configuram os estudos de citação no âmbito da intersecção entre estudos métricos, comunicação científica e ciência da informação" (SILVEIRA; CAREGNATO, 2017, p. 146). Como parte central da discussão trata-se do objeto científico que deve ser analisado com métodos adequados e em acordo com as hipóteses levantadas. Para os autores as contribuições dos estudos de citação não deixam claros os fenômenos ou suas práticas, deixando esses elementos implícitos. Defendem então que esses estudos devem expressar os elementos científicos de forma "integrada e ampliada".

A relação do presente trabalho com a Ciência da Informação se dá, portanto, de forma bastante direta. Em especial sobre sua relação direta com a Cienciometria, sendo esta objeto de estudo de diversas publicações em Ciência da Informação (CURTY; DELBIANCO, 2020; PINTO; MATIAS, 2012). De forma análoga, a atribuição de valores à produção científica de grupos de pesquisadores também possui forte relação com a área da Ciência da Informação.

1.5 CONTRIBUIÇÕES DA TESE

Como contribuições principais este trabalho apresenta: i) uma nova forma de selecionar e agregar artigos científicos no propósito de avaliação a produtividade de grupos, e, ainda; ii) um *Framework* para estruturar a avaliação da produção científica de grupos de pesquisadores.

1.6 ESTRUTURA DA TESE

O restante deste documento está divido da seguinte forma: no Capítulo 2 são apresentados os conceitos básicos da presente tese. A metodologia do trabalho com a classificação da pesquisa e passos realizados para a revisão sistemática da literatura, dentre outras coletas de dados, é apresentada no Capítulo 3. No quarto capítulo são apresentados os resultados da revisão sistemática de literatura. No quinto capítulo é apresentada a proposta da nova forma de agregação *IN-GROUP* e do

framework F-GROUP para aplicação de indicadores de produtividade científica em grupos de pesquisadores, com foco na seleção de artigos. O capítulo 6 apresenta a análise e a discussão dos dados com utilização do *framework F-GROUP* com dados da *Web of Science*. E o capítulo 7 apresenta a conclusão, retomando os objetivos estabelecidos e as principais descobertas da pesquisa.

2 CONCEITOS BÁSICOS

Dentre os conceitos básicos da presente tese, temos dois temas centrais. Primeiramente há uma discussão acerca da sociologia do trabalho e da ideia de valor do pesquisador. Este primeiro conjunto de reflexões apresenta o tema dos indicadores de pesquisa de uma maneira mais amplificada. Essa discussão abrange a construção de valor e identidade do profissional Pesquisador. São discutidas também as ideias associadas a valor e a uma aplicação deste conceito à avaliação de pesquisadores. O segundo tema central é resultado de uma revisão exploratória acerca da utilização do índice h e sua aplicação a grupos de pesquisadores. A aplicação do índice h a grupos de pesquisadores permite uma série de reflexões acerca da aplicação de indicadores de produção científica a grupos de pesquisadores.

2.1 OS INDICADORES DE PRODUTIVIDADE ACADÊMICA E A RELAÇÃO DE VALOR COM O PESQUISADOR

O mundo do trabalho tem sofrido mutações significativas, tornando-se ao mesmo tempo mais diverso e especializado no século XXI. Essas mudanças vão desde a expansão do terceiro setor, inclusão das mulheres no mercado de trabalho e a modificação do espaço de trabalho, com o trabalho remoto, por exemplo (ANTUNES; ALVES, 2004). Considerando a ampliação do terceiro setor, a informatização e a introdução de métricas para estimular a competição interna entre os trabalhadores, os meios para medir e qualificar trabalho e produção têm se tornado populares, dentre eles metas de venda e número de emissão de pareceres técnicos, por exemplo.

Nos últimos anos, houve um grande aumento tanto na proposta de novos indicadores de produtividade e qualidade para publicações científicas, quanto na utilização deles. Dentre as métricas mais comuns temos a quantidade de artigos, a quantidade de citações e também indicadores como o índice h (HIRSCH, 2005), por exemplo. Esses indicadores de forma direta medem um conjunto de publicações e atribuem um *score* ao pesquisador, de acordo com o número de suas publicações e da quantidade das citações que suas publicações recebem. Podem, de forma muito simples, atribuir um valor para o qual o total de artigos pode ser considerado um esforço realizado e o total de citações é o resultado desse esforço.

A utilização direta de uma fórmula simples que combina total de artigos e total de citações pode ser uma opção para atribuir um valor ao pesquisador. No entanto, é necessária uma discussão para identificar critérios para atribuição de um número que melhor represente este valor. Adicionalmente, é importante reconhecer os conceitos que tornam um pesquisador mais valoroso, tanto entre seus pares, quanto para entidades externas. A valorização do pesquisador será mais bem percebida quando os profissionais entenderem quais são os atributos buscados, para, assim, externalizar melhor o trabalho que já realizam.

2.1.1 Principais Indicadores para Produção Científica

Para que seja possível traçar um paralelo entre o conceito de valor e o valor relacionado à produção científica de pesquisadores faz-se necessário um estudo das principais formas de avaliação da produtividade científica. Csillag (1955) definiu através de uma fórmula o conceito de valor como sendo a razão entre o Benefícios (Obteve) sobre ao Recursos (Esforços). A definição de (CSILLAG, 1995) foi proposta no sentido de valor percebido pelo cliente, no entanto, o mesmo conceito é aplicado na área da Ciência da Informação (RADOS; VALERIM; BLATTMANN, 1999), dentre outras aplicações. Sendo o valor uma grandeza relacionada a um esforço, em especial a um esforço percebido, faz sentido associar o valor de um pesquisador com a sua produção científica, uma vez que suas publicações são em grande parte o resultado direto de seu esforço. Desta forma, esta seção apresenta os principais indicadores utilizados atualmente para avaliar produção e qualidade dessas produções.

Uma das primeiras ideias quando o assunto é avaliar a produção científica é utilizar a quantidade de artigos publicados (AGARWAL *et al*., 2016; JOSHI, 2014). O total de artigos pode ser expresso em número absoluto, como o total de artigos já publicados por um determinado autor, e pode ainda ter variações de acordo com o contexto. Por exemplo, pode-se fazer uma avaliação mais focada nos últimos 5 anos, nos últimos 12 meses, ou ainda, considerando o ano anterior. Entretanto, a quantidade de artigos, embora seja bastante útil para medir a quantidade da produção científica, não é capaz de determinar ou indicar a qualidade dessa produção.

Uma forma bastante utilizada, pela fácil aplicação, para medir a qualidade de uma produção é o número de citações. Neste contexto, um artigo é considerado

relevante se ele serve muitas vezes de base para outros estudos, pois outros autores costumam referenciar trabalhos a partir de critérios qualitativos. São avaliados aspectos como, por exemplo, os detalhes para que um experimento possa ser replicado, um bom embasamento teórico, entre outros. No entanto, ao se utilizar a citação como indicador observa-se o problema da autocitação, ou seja, para aumentar os indicadores de citação os autores passaram a citar mais seus próprios artigos (FLATT; BLASIMME; VAYENA, 2017; KOSMULSKI, 2005). Além disso, o número de citações pode ser enviesado por um único artigo, atestando a qualidade desse trabalho em específico, mas não de todos os artigos já produzidos por seu autor.

Para balancear a quantidade com a qualidade dos artigos foi proposto o índice h (HIRSCH, 2005). Este índice é obtido a partir da combinação tanto do número de publicações quanto do número de citações que os artigos recebem e seu cálculo é bem simples: se autor tem 1 artigo e esse artigo recebeu 1 citação seu índice h é 1, se o autor tem dois artigos e cada artigo recebeu duas citações seu índice h é 2, e assim sucessivamente, sendo que um autor com índice h de valor 10 deve possuir ao menos 10 artigos publicados, com pelo menos 10 citações em cada um desses artigos. Ou seja, há uma combinação da quantidade dos artigos, mas também é necessário que o número de citações ocorra de forma pulverizada em diversos trabalhos diferentes, para caracterizar um valor mais elevado do índice h.

Outros indicadores surgiram ainda para tentar evitar distorções, como o problema da autocitação. Há ainda o problema relacionado à diferença etária refletir no valor do índice, fazendo pesquisadores mais velhos, e até aposentados, sempre apresentarem valores mais altos de índice h comparados a novos pesquisadores, mesmo que possuam produções relevantes no início da carreira.

Foram propostas então variações do índice h, como, por exemplo, o *h(2)-index* (KOSMULSKI, 2005) que tenta diminuir a distorção causada pelas autocitações, modificando o índice h quanto ao número de citações, que nesse caso deve ser sempre elevada ao quadrado. Outro exemplo é o *g-Index*, que corresponde ao número *g* de publicações que possuem pelo menos g^2 citações juntos (EGGHE, 2006). Outra variante considera somente a média do número de citações que um autor recebeu em suas publicações, denominado A-*index* (JIN *et al.*, 2007). Existem ainda medidas que consideram apenas uma janela temporal para delimitar o período de produção em

pesquisa. O *AIF* (*Author Impact Factor*), por exemplo, considera a média das citações dada uma janela temporal (PAN; FORTUNATO, 2015).

Um importante indicador a ser considerado é o *Crown*, que considera o número de citações de artigos semelhantes (JOSHI, 2014). Neste indicador há uma comparação com uma média geral de citações de artigos semelhantes em período, área de pesquisa e tipo de documento. Se a média de citações do grupo que está sendo analisado for maior do que a média global em 10% o indicador corresponderá ao valor de 1.1. Mas, se a média do grupo de publicações que está sendo analisado for 10% menor do que a média global, o indicador acusará um valor de 0.9.

Outra forma de avaliar o impacto/qualidade das publicações de um pesquisador é o uso de ferramentas que permitem identificar o impacto da pesquisa de forma *online*, como quantidade de referências a um determinado artigo em revistas e jornais não científicos e até de sites da *Internet* (VALLES *et al.*, 2020). Mesmo indicadores mais recentes combinam a produção, dada pelo número de indicadores, com a qualidade ou relevância, dada pelo número de citações ou outras formas de utilização que são passíveis de ser monitoradas através de sistemas automatizados.

A utilização desses indicadores tem moldado a forma de percepção de valor das pesquisas e, consequentemente, a percepção de valor em relação aos profissionais que atuam com pesquisa científica. Esses indicadores têm sido cada vez mais utilizados para distribuição de recursos destinados à realização de pesquisas científicas (JOSHI, 2014; ROEMER; BORCHARDT, 2005). Nesse sentido, é importante que seja estudada a influência da atribuição de valor, e seu relacionamento com esses indicadores, para os profissionais Pesquisadores. A seguir, apresenta-se uma breve discussão sobre esta temática.

2.1.2 A relação entre Indicadores de Produção Científica e a percepção de Valor para Pesquisadores

Muitas vezes os critérios para definição de alocação de recursos para pesquisa se dão a partir de parâmetros mensuráveis, desta forma, se faz relevante a discussão do valor do pesquisador. Deve-se discutir de que forma seria possível analisar os critérios para mensurar o *status*, traduzido em valor para esses profissionais. Larson descreve a importância do *status* e sua relação com a profissão

e o trabalho (ALMEIDA, 2010). Sendo esse *status* construído através de monopólios e segmentação do mercado de trabalho, construindo uma divisão social. O conceito de *status* profissional é naturalmente imensurável, entretanto, o conceito de valor tem sido discutido e aplicado em diferentes segmentos (CSILLAG, 1995).

O *status* do pesquisador está muitas vezes atrelado às suas contribuições como pesquisador, seja pela relevância de suas publicações, seja por parcerias e contribuições construídas a partir do método científico para mudanças na vida cotidiana das pessoas. Esse *status* pode ser ainda comparado com uma medida de valor. Quanto maior o *status* de um pesquisador, ou instituição de pesquisa, maior será o valor percebido por essas entidades na sociedade.

Em geral, a ideia de valor está associada a um produto de uma organização e à percepção de um usuário acerca daquele produto, sendo ainda um conceito muito associado a aspectos econômicos. Uma fórmula muito usada para determinar valor é a que associa este valor ao resultado de um determinado esforço (CSILLAG, 1995). O valor existe somente quando é percebido por alguém, sendo então o conceito de *valor percebido* uma resultante de *benefícios percebidos* sob *esforços percebidos*. Para medir o valor de um produto ou serviço é realizada uma análise que independe do que está sendo avaliado (CSILLAG, 1995).

É comum, com a descrição de valor, associar esse conceito à ideia de produtos fabris, como carros ou canetas, ou ainda pensar nesses conceitos apenas em organizações voltadas à comercialização de alguma mercadoria. No entanto, a ideia de valor é também associada a serviços, como desenvolvimento de sistemas, por exemplo. É possível, inclusive, associar os conceitos de valor aos serviços prestados por uma biblioteca, associando ainda a ideia do valor da informação (SILVA; SCHONS; RADOS, 2006).

O pesquisador, está sujeito a análises acerca de sua produção científica. Ou seja, sua produção é avaliada, muitas vezes de forma automática por sistemas, sempre que publicam um novo artigo. Aspectos como quantidades de artigos e citações são absorvidos por diversas plataformas sempre que um novo artigo científico é publicado e disponibilizado na *Internet*. Os indicadores acerca da produção do profissional são então calculados e disponibilizados em diversas plataformas.

Para que seja possível identificar as relações entre os indicadores apresentados e a ideia de valor atribuída ao pesquisador é necessário estabelecer

como a ideia de valor tem sido associada aos profissionais em uma perspectiva sociológica. Dubar (2005) explicita os modelos definidos por Moore, e suas relações de valor, sendo eles apresentados no Quadro 1.

Quadro 1 - Modelos de valor

Modelo	**Descrição**
a) modelo operário	o valor se dá pelo resultado obtido.
b) modelo oficial	a valorização se dá pela função com uma identificação pelo *status*.
c) modelo físico	valorização ocorre através da formação e identificação com a disciplina/especialidade.

Fonte: adaptado de (DUBAR, 2005, p. 209)

a) modelo operário: Neste modelo o *cargo* e a progressão desses cargos é o que atribui o valor ao profissional. Este cargo é definido por um determinado conjunto de tarefas.

> O núcleo da competência é a FORMAÇÃO IN LOCO, isto é, a capacidade para produzir resultados, proveniente da experiência e do domínio da atividade de trabalho. O salário sanciona a contribuição para a tarefa principal, a que produz valor agregado incorporado ao resultado do trabalho. (...) A codificação principal é a que classifica os cargos segundo sua importância na produção dos resultados. A codificação dos indivíduos decorre da codificação precedente e repousa nas experiências anteriores (cartão de visita, currículo...) e nas aptidões medidas por testes específicos. A carreira não é concebida senão como uma progressão em cargos cada vez mais importantes, suscetíveis de produzir resultados crescentes. O sucesso profissional é medido por esses resultados: é uma "carreira nos cargos", fundada na acumulação "interna" de competências operacionais. (DUBAR, 2005, p. 207)

O conceito de progressão de cargos se dá pela ampliação das experiências, que chancela um conhecimento adquirido através da prática. Neste modelo, há um conjunto de requisitos definidos para cada cargo, havendo uma valorização de autodidatas.

b) modelo oficial: A *função* é o que determina o valor do profissional, sendo essa função atribuída através de títulos oficiais. É essa *função* que carrega atos de responsabilidade, seu valor está atrelado ao peso delas, sendo que a progressão funcional é o aumento dessas responsabilidades. É o grupo de profissionais com a mesma função e *status* que determina os requisitos mínimos para que outros possam exercer tal função (DUBAR, 2005, p. 210).

c) modelo físico: A *especialidade* é o que define o valor do profissional neste modelo, ou seja, a competência para exercer o trabalho vem de conhecimentos adquiridos. "O que se busca, antes de tudo, é o reconhecimento pelos pares, e o engajamento profissional está profundamente condicionado à esperança de um aumento desse reconhecimento ancorado, com frequência, na concepção de "vocação" (...)" (DUBAR, 2005, p. 210). Para Hughes, o pesquisador estaria enquadrado neste modelo.

Os modelos apresentados trazem uma perspectiva de valor relacionada ao profissional de forma direta. Não há uma valorização do resultado do trabalho, ou de como um trabalho realizado poderia influenciar na percepção de valor do profissional. Existem diversas críticas aos modelos propostos, uma vez que tentam encaixar em um só modelo todo o universo do trabalho (ANGELIN, 2010; URTEAGA, 2008). Mueller (2004) defende que esses aspectos devem ser baseados em concepções macro históricas, considerando a influência de condições sociais no processo de profissionalização.

Nos modelos apresentados o pesquisador se enquadraria no modelo físico, uma vez que sua valorização ocorre a partir de sua especialidade. No entanto, com um maior acesso à educação formal e um número crescente de pesquisadores com formação elevada, não seria possível estabelecer uma diferenciação entre os pares. Essa diferenciação se torna necessária quando há uma disputa por recursos e necessidade de destaque e valorização.

Freidson (1996) lembra que a educação "superior" é assim chamada e socialmente elevada em relação a outras formações como técnica, ou secundarista. Ainda segundo Freidson, a classe de pesquisadores científicos seria a de mais elevado grau de estudo especializado. Desta forma, elevados valores nos diferentes indicadores de produção científica contribuem para reconhecimento do trabalho realizado. Esses conceitos podem ser associados às ideias de Dubar acerca da estruturação da profissão, a qual deve ser reconhecida tanto pelo público interno quanto pelo externo.

Para Freidson, o ensino das profissões cria uma distinção entre profissionais e autoridades acadêmicas. O conceito sociológico convencional de profissão "liga corpos de conhecimento, discurso, disciplinas e campos aos meios sociais, econômicos e políticos por meio dos quais seus expoentes humanos podem ganhar

poder e exercê-lo" (FREIDSON, 1996). O autor destaca que houve um aumento no interesse pelo estudo das profissões nos últimos anos. Antes, o estudo das profissões na sociologia se moldava em correntes anglo saxãs, no entanto, com o aumento da força de trabalho com acesso à educação universitária esse panorama mudou.

Esse recente aumento de interesse pelas profissões pode ser explicado em termos práticos pelo fato de as profissões e os profissionais terem se tornado tão numerosos e importantes, em especial nos países industriais avançados. Houve um constante aumento nas ocupações de formação universitária, que ganharam posições privilegiadas tanto no serviço público civil como no mercado privado, o mesmo ocorrendo quanto à proporção de profissionais na força de trabalho como um todo. Não obstante, devido à falta de qualquer consenso sobre as ocupações que deveriam ser estudadas e sobre o tipo de informação que deveria ser coletado a seu respeito, a maioria dos estudos é apenas toscamente comparável, mesmo quando eles examinam a mesma ocupação (FREIDSON, 1996).

Essa comparação dificultada destacada por Freidson também é refletida nas métricas atuais para análise de produtividade dos pesquisadores. Considerando indicadores relacionados à citação, esse número pode variar muito de acordo com a área em que atua o pesquisador. O número de congressos específicos e revistas especializadas em determinada temática pode favorecer ou não a carreira de um pesquisador, trazendo oportunidades diferentes para que os profissionais publiquem seus trabalhos em destaque e sejam, assim, citados por outros. Ou seja, de acordo com a área e oportunidades, trabalhos de mesmo rigor técnico e qualidade podem receber número de citações muito diferentes de acordo com o local onde são publicadas.

2.1.3 A Composição de Valor do Profissional Pesquisador

Como descrito por Harari (2018), a ficção científica e a ciência real são muito distantes e acabam frustrando a população em relação aos avanços da Ciência. A frustração da população em relação à classe científica se dá pelo descrito por Harari, ou seja, com ideias fantasiosas acerca da ciência, seja com carros voadores ou mesmo com cura de doenças aguardadas há décadas. Mas esse fenômeno ocorre

também pela distância social da classe de pesquisadores em relação às camadas mais pobres da população, em especial em países com grande desigualdade social.

A ideia de pesquisador perante a população é muitas vezes limitada a estereótipos, como a do profissional com jaleco branco e tubos de ensaio. A construção de estereótipos deve ser repelida dentro das profissões, a passagem através do Espelho de Hughes, descrita por Dubar (2005) apresenta essa reflexão. Nessa passagem, que busca afastar estereótipos no campo profissional, uma revisão clara e objetiva deve ocorrer de forma constante, fazendo com que os indicadores reflitam qualidade e sejam imparciais. Essa revisão de estereótipos pode ser comparada tanto à definição de valor do pesquisador, quanto da utilização de indicadores definidos pela própria classe de pesquisadores.

Essa imparcialidade passa ainda pelas revisões de artigos. Que deve ser o mais imparcial possível para que se evite a criação de estereótipos na profissão. Ou seja, para que as revistas de maior impacto e, consequentemente, visibilidade, tenham cada vez mais qualidade e credibilidade, deve-se defender avaliações mais neutras e imparciais. Um estudo demonstrou que as avaliações de artigos, quando realizadas com identificação dos autores, tendem a ser construídas com diferentes adjetivos utilizados para caracterizar os trabalhos de diferentes gêneros (GYULA NAGY, 2018). Dessa forma, demonstra-se que os estereótipos podem influenciar de maneira negativa uma avaliação profissional.

> A ideologia do generalismo cultivado não é a única inimiga que o profissionalismo deve cooptar ou neutralizar. Também o são as ideologias da economia liberal e do comunismo, ambas as quais, de formas diferentes, se opõem à garantia de status privilegiado para a qualificação técnica, por mais temperada e aprofundada que seja por uma educação liberal. Embora de modos diversos, ambas defendem que, na medida em que os seres humanos se comprometem com alguma meta, seja o ganho material ou a criação de uma nova sociedade -,eles são capazes de atingir esse fim sem depender de especialistas. (FREIDSON, 1996, p. 06).

Como apresentado por Freidson (1996), as questões ideológicas também devem ser neutralizadas. É com essa neutralização de estereótipos e de ideologias que o *status* relacionado à qualificação técnica se mantém. Freidson ainda atenta à definição de metas ideológicas como um artifício para perseguir esse objetivo, sem depender de especialistas. Em um mundo polarizado, a questão levantada por Freidson fica ainda mais clara, uma vez que para satisfazer os critérios ideológicos os indivíduos ignoram especialistas em ciência e fortalecem suas opiniões com base

apenas em suas ideologias. Logo, as metas precisam de uma revisão, de forma que não somente os quantitativos de ciência voltada apenas à própria ciência sejam valorizados, mas sim que aspectos sociais também façam parte dessas metas e indicadores.

A valorização do pesquisador passa então por formas mais justas e equânimes na avaliação das publicações científicas. Este quesito permite utilizar de forma mais direta a ideia de valor associada à produção e relevância. Esse valor, dada a conjuntura atual, pode ser calculado de forma mais direta através do esforço do pesquisador, representado por suas publicações e ainda pelo reconhecimento da qualidade e contribuição das suas publicações, sendo esse reconhecimento dado através da utilização de seus resultados, seja esse uso expresso em forma de citação em outros artigos científicos, seja pela utilização da comunidade não científica como indústria e governo.

Assim sendo, a composição do valor do pesquisador de Ciência se dá ao somar suas contribuições, dividindo pelo esforço para alcançar esses resultados. As contribuições são o que foi alcançado através do seu esforço, ou seja, as citações, as utilizações na indústria e governo, a colaboração com membros da sociedade civil etc. O esforço se dá pelas publicações, que determinam o resultado direto do trabalho do pesquisador. Ao dividir esses elementos temos então o valor do pesquisador.

Os indicadores que medem a produtividade de pesquisadores têm se tornado cada vez mais populares e são utilizados de forma automatizada em diversas plataformas (AGARWAL *et al*., 2016). De fato, a atribuição de valores numéricos a pesquisadores, pode, de certa forma, reduzir, desqualificar ou ainda, precarizar a forma de trabalho, em especial de uma comunidade tão relevante quanto a científica. No entanto, os indicadores atuais já são utilizados de forma generalizada, e carecem da contribuição e aprimoramento com a inclusão de outras fontes, de forma que não privilegiem somente uma parcela dos pesquisadores. "O que ocorre no mercado de trabalho, arena em que, segundo Abbott, predominam negociações e hábitos, e na qual valem resultados mais que discursos?" (MUELLER, 2004).

Os principais indicadores para avaliar a produção científica podem, portanto, ser relacionados a indicadores com os critérios para composição do valor. O valor do pesquisador, assim como o valor da informação, será positivo quando os benefícios superarem os recursos utilizados (RADOS; VALERIM; BLATTMANN, 1999). Desta

forma, para que o benefício seja mais bem percebido, é necessário que mais aspectos sejam verificados na composição desse valor.

Dentre esses aspectos, o social, no sentido das produções em conjunto, pode caracterizar um conjunto de métricas relevante na valorização dos pesquisadores. Indicadores que apresentem aspectos relacionados a características dos grupos, como proporção de membros da mesma universidade ou país, podem ser utilizados para identificar os grupos que melhor atendam critérios de colaboração em pesquisa, por exemplo. Nos casos em que o objetivo é avaliar um grupo, essa valorização da qualidade das pesquisas também deve ser refletida no grupo como um todo. Dentre as formas de avaliar grupos de pesquisadores que visam analisar quantidade e qualidade, temos as variações do índice h aplicado a grupos, detalhados na próxima seção.

2.2 APLICAÇÃO DO *ÍNDICE H* A GRUPOS DE PESQUISADORES

Modelos, classificações e estruturas baseadas em índices podem contribuir para o acesso e gerenciamento de dados dentro de instituições científicas. Um dos principais usos desses dados estatísticos é para uma distribuição mais justa de fundos e outros recursos entre as entidades (JOSHI, 2014). Neste sentido, a presente seção enfoca no problema da aplicação de métricas que consideram aspectos quantitativos e qualitativos da produção científica para grupos de pesquisadores. É realizada uma avaliação da aplicabilidade de índices de produtividade de pesquisa em grupos, buscando mitigar o problema da seleção de artigos como insumo para avaliar um grupo de pesquisadores.

Mensurar um valor intangível como a produtividade científica não é fácil, e sempre pode haver alguma distorção da realidade. Ao utilizar apenas métricas simples como número de artigos e de citações, podemos deixar de expressar valores importantes, como qualidade ou quantidade. O número de citações é a métrica mais tradicional para avaliar o impacto e a relevância de um artigo científico. O número de citações é um indicador plenamente utilizado em trabalhos recentes para propor novos modelos de classificação (SOHN; JUNG, 2015) e até mesmo por mapear a evolução das estruturas intelectuais (GONZÁLEZ-VALIENTE *et al*., 2019). A seguir, após uma

revisão de dados exploratória, são apontados desafios e oportunidades acerca da mensuração de produtividade para grupos de pesquisadores.

Vários índices têm sido propostos combinando o número de publicações e citações para medir a quantidade e relevância de um artigo. No entanto, o número de artigos e citações ainda são as mais relevantes métricas para medir produção e relevância na ciência. Esta seção concentra-se em índices que combinam o número de artigos e citações para avaliar grupos de pesquisa.

Enquanto alguns dos índices propostos são usados para calcular dados para um pesquisador individual (HIRSCH, 2005, 2019; ZHANG, 2009), outros são usados no cálculo de um grupo de pesquisa (JIN *et al*., 2007; TORRES-SALINAS; ROBINSON-GARCIA; JIMÉNEZ-CONTRERAS, 2016; VALLES *et al.*, 2020), conforme sumarizado no Quadro 2. Neste trabalho, consideramos um grupo de pesquisa como qualquer grupo de dois ou mais pesquisadores, que podem ser estudantes, professores e outros acadêmicos que trabalham/publicam juntos. Esta configuração pode variar de um pequeno grupo, composto por alguns colegas pesquisadores que se candidatam a um projeto pequeno e não remunerado a um grande grupo de pesquisa trabalhando em um projeto com financiamento internacional.

Na literatura, algumas pesquisas abordam os índices usados para avaliar pesquisadores (AGARWAL *et al.*, 2016; GASPARYAN *et al.*, 2018; JOSHI, 2014). Enquanto alguns se concentram em uma área de conhecimento específica (AGARWAL *et al.*, 2016; GASPARYAN *et al*., 2018), outros são mais generalistas (HIRSCH, 2019; JOSHI, 2014). Esses trabalhos se concentram em métricas para analisar artigos e pesquisadores individualmente, mas não há ênfase específica para um grupo de pesquisadores. Há trabalhos que propõem algum tipo de variação no valor do *índice h*, como a média ou mediana, para analisar um campo de conhecimento específico ou comparando instituições (KHAN et al., 2013; MITRA, 2006; MUGNAINI; PACKER; MENEGHINI, 2008; RAD et al., 2010; SCHUBERT, 2007). No entanto, nenhum desses trabalhos identificados analisou os índices para avaliar grupos de pesquisadores.

Alguns dos trabalhos propostos afirmam que é possível aplicar diretamente as métricas existentes para avaliar pesquisadores individuais a grupos de pesquisadores (JACSÓ, 2009; MITRA, 2006), contudo, esses trabalhos propõem variações do *índice*

h. Nesta seção são categorizadas as diferentes formas para aplicar o *índice h* para grupos de pesquisadores: i) o *h_1-índice* (MITRA, 2006), que considera toda uma instituição ou departamento, reunindo todos os artigos desta instituição/departamento em seu cálculo; ii) (KHAN et al., 2013; MUGNAINI; PACKER; MENEGHINI, 2008; RAD et al., 2010) consideram a média ou mediana dos índices h dos indivíduos que fazem parte de um grupo; iii) o chamado *índice h sucessivo* (SCHUBERT, 2007), calcula o *índice h* das instituições, identificando como principais pesquisadores todos aqueles que possuem o *índice h* maior ou igual a uma pontuação predeterminada. É uma aplicação da fórmula do *índice h* para instituições, que considera o próprio *índice h* dos pesquisadores como número de citações e a quantidade de pesquisadores ao invés da quantidade de artigos. De acordo com esta proposta, identificar as principais instituições de um país com o *índice h* maior ou igual a essa pontuação predeterminada permite calcular o *índice h* para todo o país, por isso é chamado de sucessivo.

No entanto, mesmo as formas convencionais de estender o *índice h* para avaliar grupos de pesquisadores ensejam algumas questões: Deve-se considerar todos os artigos publicados por um autor que trabalhou em várias instituições ou apenas naquelas publicadas durante o período no qual esta pessoa trabalha em sua instituição atual? Deve-se considerar todas as citações dos autores, independentemente da coautoria? Deve-se considerar o número de autores em cada artigo? Essas perguntas ainda não respondidas podem ser agrupadas em uma grande questão: *quais trabalhos devem ser selecionados para avaliar a produtividade científica de um grupo de pesquisa?*

2.2.1 Variações do índice h para avaliar grupos

O *índice h*, e outros indicadores bibliométricos para pesquisa química, foram comparados por Van Raan (2006). Para aplicar o *índice h* em grupos de pesquisa, o autor usou todos os artigos produzidos por cada um dos membros do grupo. O estudo foi realizado com 147 grupos de pesquisa em química de universidades holandesas. Foi observada uma correlação entre o *índice h* e a citação, porém, para grupos menores e com "menos tráfego pesado de citações", o indicador *Crown,* que considera

artigos semelhantes como base para diferenciar os artigos mais citados, foi considerado mais adequado.

Mitra (2006) propôs duas variações do *índice h* para avaliar a pesquisa das instituições. O primeiro, *índice* h_1, corresponde a: $h_1 = h$ quando a instituição publicou *h* artigos, com pelo menos *h* citações cada. Enquanto o *índice h* tem como entrada artigos de um determinado autor, a mesma ideia é proposta em Mitra (2006) para avaliar uma instituição. Esta é uma métrica simples que pode ser facilmente aplicada. No entanto, assim como no *índice h*, quando a publicação é um esforço de duas ou mais instituições, o mesmo artigo é contabilizado integralmente para ambas.

O segundo índice proposto por Mitra é o *índice* h_2, essa métrica considera o *índice h* calculado para cada pesquisador e dá uma ideia dos "principais indivíduos" de uma determinada instituição. Nesse cenário, uma instituição possui um *índice* $h_2 = 1$ se um de seus pesquisadores tem um *índice h* igual a 1, outra instituição tem um *índice* $h_2 = 2$ se dois de seus pesquisadores têm um *índice h* de valor 2 e, consequentemente, uma instituição terá um *índice* $h_2 = 50$ quando 50 de seus pesquisadores têm um *índice h* de pelo menos 50, e assim por diante. No entanto, este índice pode ser tendencioso por um grupo de pesquisa da mesma instituição, cujos membros tendem a publicar juntos. Neste caso, os mesmos trabalhos serão considerados duas ou mais vezes no cálculo.

Índices h sucessivos são um modelo proposto por Schubert (2007). O índice h do autor é calculado usando seu número de artigos e citações. Em nível institucional, a ideia é utilizar o *índice* h_2 proposto por Mitra. Para níveis mais altos, por exemplo, toda uma região ou país, a mesma ideia com base no índice h_2 é aplicada. A modelagem de índices h sucessivos propõe que um país poderia ser representado pelos índices h_2 de suas instituições, por exemplo, sendo essa abordagem aplicável a outros níveis. Por exemplo, se no Brasil ao menos 50 instituições possuíssem um índice h mínimo de 50, o índice h do Brasil seria 50.

O CP-*index* foi apresentado por Altmann, Abbasi e Hwang (2009). É baseado no *RP-index*, proposto no mesmo trabalho. O CP-*index* é definido de forma semelhante ao índice h, utilizando o RP-*index* no lugar do número de artigos e citações. Por sua vez, o RP-*index* é baseado em um cálculo que reúne número de citações, idade das publicações e a contribuição do pesquisador em cada publicação. É como o modelo de *índice h* sucessivo, mas ao invés de utilizar o *índice h* de cada

pesquisador, usa sua própria métrica *RP-index*. Sendo assim, se ao menos 10 pesquisadores de uma determinada instituição possuírem um RP-*index* de 10 ou mais, o valor CP-*index* desta instituição será 10.

Em 2009, Jacsó aplicou o *índice h* para países sul-americanos na *Web of Science* e *Scopus*. O autor comparou os 10 primeiros países com maior valor de índices h, formando um *ranking* para cada base de dados. O estudo considerou que o *índice h* para *Scopus* e *Web of Science* foram robustos, pois mesmo com números de artigos diferentes, ambos os *rankings* mostram quase o mesmo resultado para os países avaliados. A única exceção ocorreu com o *ranking Scopus,* no qual a Argentina ficou classificada em segundo lugar no *ranking*, com o Chile em primeiro. Desta forma a ordem foi invertida em relação aos resultados obtidos utilizando os dado da *WoS* (JACSÓ, 2009).

Uma aplicação da média do *índice h* foi proposta em para avaliar a pesquisa em radiologia (RAD et al., 2010). Este estudo selecionou programas de radiologia e obteve os índices h, bem como o número de citações e publicações para os radiologistas selecionados. Neste estudo foi realizada uma análise de regressão para determinar quais variáveis se associaram melhor ao *ranking* acadêmico. Uma correlação foi identificada entre *índice h* e o *ranking* acadêmico obtido na pesquisa: quanto mais alta a posição no *ranking*, maior o valor do índice h.

O *índice* h_2/h_1 foi proposto em 2010 (ROUSSEAU; YANG; YUE, 2010). Neste trabalho, os autores destacam que os índices propostos por Mitra (2006) podem ser calculados de diversas maneiras: seja considerando todos os trabalhos de um determinado autor ou apenas aqueles que têm o endereço das instituições no corpo do artigo. Eles também apontam que duas instituições com o mesmo *índice* h_2 podem ter diferentes perfis de publicações. Além disso, autores com *índice h* médio podem não se sentir valorizados em suas instituições. Eles propuseram dividir h_2/h_1 para dar um indicador estrutural da instituição. Isto é, se uma determinada universidade possui ao menos 10 autores com índice h de 10 ou mais, e esta mesma universidade possui ao menos 20 artigos com 20 citações ou mais, seu índice h_2/h_1 será de 10/20, ou seja, 0,5.

Outra aplicação da média do *índice h* para avaliar um departamento de neurocirurgia foi proposta por Khan (2013). A média do *índice h* foi analisada considerando sexo, nível acadêmico, anos de prática, subespecialidade e instituição.

Concluem que a média do *índice h* pode distinguir produtividade para classificação acadêmica, subespecialidade e anos de prática. Apontam que a utilização da média do índice h é mais adequada do que quando o índice é calculado de forma cumulativa. Justificam que o índice h de forma cumulativa pode ser mais influenciado por um maior número de novos docentes, por exemplo, enquanto a média é mais facilmente influenciada pela produtividade acadêmica de todos os membros do grupo.

Uma sumarização dos indicadores elencados está presente no Quadro 2. Neste quadro temos a métrica proposta, uma indicação sobre se a variação apresentada pode ser aplicada ou não para grupos de pesquisadores, a fonte que é utilizada para calcular a variação do índice h em questão, além de uma breve descrição de cada métrica.

Quadro 2 – Variações do índice h para avaliar grupos de pesquisadores

Métrica	**Usado para**	**Fonte**	**Breve descrição**
índice h (Hirsch, 2005)	Individual	N. de artigos e citações	Total de *h* artigos com ao menos *h* citações cada.
Crown Indicator	Grupo e Individual	N. de citações	Compara o artigo com a média de citações de artigos semelhantes.
h_1*-index* (Mitra, 2006)	Grupo	N. de artigos e citações	Total de *h* artigos com ao menos *h* citações cada.
h_2*-index* (Mitra, 2006)	Grupo	N. de pesquisadores o índice h de cada pesquisador	Total de *n* autores com um índice h de ao menos *n* cada.
Successive h-indices (Shubert, 2007)	Grupo	índice h	Aplicação sucessiva do índice h usando outros valores.
CP-index (Altmann *et al*., 2009)	Grupo	RP-*index*	Número de pesquisadores *n* com um total de *n* pesquisadores com um RP-*index* de *n* ao menos cada.
h_2*/* h_1 *–ratio* (Rousseau *et al*., 2010)	Grupo	h_1-*index*, h_2-*index*	h_2-*index* / h_1-*index*

Fonte: Adaptado e traduzido de Santos e Dutra (2021).

Esta seção apresentou as variações do *índice h* para avaliar grupos de pesquisadores, mostrando diferentes aplicações e resultados de acordo com cada campo. Na próxima seção, são apresentados os desafios e oportunidades para medir a pesquisa de grupos de pesquisadores.

2.2.2 Desafios para avaliar grupos de pesquisa

O *índice h* é a métrica mais popular para combinar aspectos quantitativos e qualitativos para a avaliação da produtividade em pesquisa científica. A ideia principal deste índice é bastante prática e direta. Ele fornece um número que combina uma medida quantitativa relacionada à produtividade com a relevância dos trabalhos de acordo com o número de citações recebidas. No entanto, mesmo sendo um índice bastante útil, algumas questões foram trazidas à luz nos últimos anos. Essas questões estão sumarizadas no Quadro 3:

Quadro 3 - Síntese desafios na utilização do índice h.

	Grupo/Individual	**Desafio**
1	Individual	O problema “qualidade *versus* quantidade”.
2	Individual	O *índice h* se mantém alto mesmo após o encerramento das atividades acadêmicas.
3	Individual	Autores com total de artigos e citações bem distintos podem ter o mesmo *índice h*.
4	Grupo	Quais artigos científicos devem ser usados para calcular o índice.
5	Grupo	Cálculo da produção de um grupo em conjunto.

Fonte: elaborado pela autora

Uma das primeiras questões está relacionada com o problema “qualidade *versus* quantidade”. Autores que produzem majoritariamente revisões sistemática*s*, provavelmente aumentarão seu número de citações em comparação com outros autores que produzem novos conhecimentos. O mais relevante índice para superar esse problema é o Crown, que considera apenas artigos identificados como equivalentes por características específicas, como área de aplicação, tipo de

publicação, período de publicação, entre outros aspectos, para proporcionar uma comparação no número de citações.

Outra questão relacionada ao *índice h* é que um autor pode manter um *índice h* bastante alto mesmo após anos de sua aposentadoria. Ao mesmo tempo, novos pesquisadores com novas contribuições podem ser prejudicados, mesmo apresentando qualidade e quantidade em uma carreira ainda iniciante. Assim, alguns índices foram propostos para superar este problema por meio do emprego de um intervalo de tempo para levar os artigos em consideração, como a métrica AIF (PAN; FORTUNATO, 2015).

Autores com números de artigos e citações bastante diferentes podem ter o mesmo *índice h*, então o uso de autocitação para aumentar seus índices h também se tornou um problema relevante. De fato, o *índice h* em certos campos não representa muito bem a realidade. Alguns índices foram propostos aplicando-se uma transformação matemática no *índice h*, por exemplo, considerando citações em valores quadrados (FLATT; BLASIMME; VAYENA, 2017; JIN et al., 2007).

Ao abordar a avaliação de grupos, muitos outros problemas se tornam aparentes. O primeiro está relacionado a quais artigos científicos devem ser usados para calcular os índices. Muitas propostas se concentram no nível institucional, considerando todos os artigos desenvolvidos no seio de uma determinada instituição. Porém, não está claro se todos os artigos de cada pesquisa da instituição são considerados ou se há algum processamento especial para os trabalhos de um pesquisador específico que possui vários vínculos ou que mudou seu local de trabalho ao longo dos anos. Esta questão é essencial para ser considerada nas variações do *índice h* aplicadas a grupos de pesquisadores: índice h_1, sucessivos ou médias dependem de um conjunto de artigos como entrada.

Outra questão relevante não abordada nos trabalhos anteriores é a força de um grupo como um todo, ou como um grupo fechado trabalha em conjunto. Há de alguma forma uma união de partes separadas que não se comportam como um grupo inteiro. É possível, portanto, estabelecer formas que permitam a avaliação e comparação, quando possível, entre diferentes tipos de grupo de pesquisadores, sejam grupos maiores como instituições e universidades, ou grupos menores, focando em quantitativos do quanto cada grupo consegue produzir sozinho e de quando são

necessários mais pesquisadores. Na próxima seção, apresentamos algumas oportunidades para mitigar esses problemas.

2.2.3 Oportunidades na avaliação de grupos de pesquisa

Com base nas questões destacadas na seção anterior e sumarizadas no Quadro 3, é possível observar que algumas das lacunas apontadas nos índices existentes já foram superadas. No entanto, uma das questões pendentes é a avaliação do prestígio, qualidade e quantidade da produção de conhecimento de um grupo de pesquisa baseado em suas publicações.

Artigos e projetos científicos geralmente são um esforço conjunto de muitas pessoas. Por esta razão é comum avaliar novas propostas de projetos de pesquisa com base nos fundamentos e critérios relacionados ao próprio projeto. Assim, quando várias propostas de projetos estão bem formuladas, um novo conjunto de critérios mais objetivos deve ser usado para fornecer uma avaliação mais justa. Novas formas de avaliação podem considerar o número de pesquisadores, a qualidade de seu trabalho em grupo e os aspectos de cada pesquisador e do pesquisador principal.

No entanto, antes de propor tais formas de medida, é necessário abordar duas questões principais destacadas neste trabalho: como selecionar os trabalhos necessários para avaliar um grupo de pesquisa? Como avaliar um grupo de pesquisadores?

No próximo capítulo, apresentamos a metodologia do presente trabalho.

3 CAMINHOS METODOLÓGICOS

A metodologia deste trabalho está dividida em quatro partes. Na primeira é realizada uma classificação da pesquisa, com base nas definições clássicas de Silva e Menezes (2005). Na segunda parte são descritas as principais etapas da pesquisa. As estratégias de coleta dos dados para revisão sistemática de literatura estão presentes na terceira etapa. Na quarta e última seção está descrita a forma de coleta e o tratamento dos dados utilizados para os experimentos com dados reais.

3.1 CLASSIFICAÇÃO DA PESQUISA

Como forma de classificar a pesquisa temos as definidas por Silva e Menezes (2005) como: i) natureza; ii) forma de abordagem do problema; iii) objetivos; e iv) procedimentos técnicos. Sendo os dois últimos critérios definidos por Gil (1991). Do ponto de vista de sua natureza, a presente pesquisa caracteriza-se como pesquisa aplicada. Pois seus objetivos se encaixam em critérios que preveem a aplicação prática, podendo ser utilizada por exemplo, na gestão do ambiente universitário. A proposta de mecanismos para avaliação de grupos de pesquisadores pode auxiliar na melhor alocação de recursos humanos e financeiros, dentre outros. Pode também ser aplicada na avaliação de novos projetos de pesquisa, identificando trabalhos de grupos de pesquisadores que trabalham bem em conjunto ou, ainda, de áreas que podem ser mais bem exploradas em termos de interdisciplinaridade.

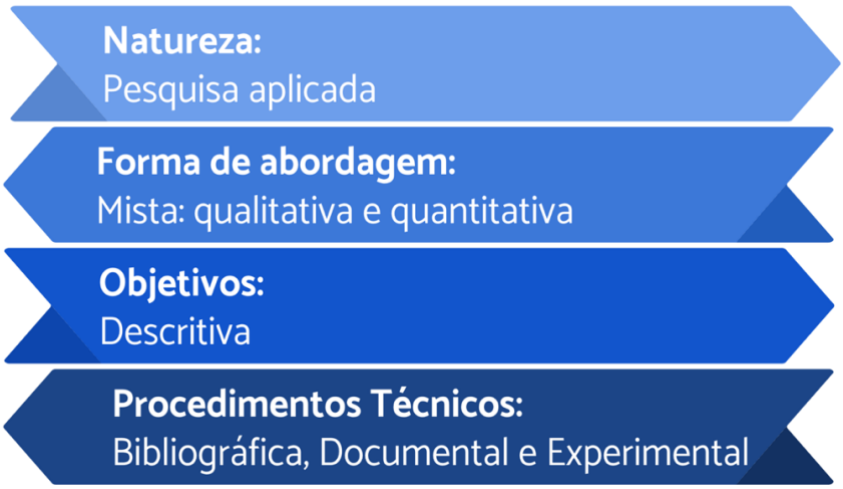
Quanto à forma de abordagem, esta pesquisa é caracterizada de forma mista, tanto qualitativa quanto quantitativa (CRESWELL, 2010). Qualitativa em sua primeira parte, uma vez que busca analisar as métricas existentes e identificar de maneira descritiva a aderência ao problema da avaliação de grupos de pesquisadores. A característica quantitativa dá-se pela análise de artigos recuperados e classificação destes artigos de acordo com suas características. Também há uma análise quantitativa da aplicação dos artefatos propostos no decorrer da tese. São, portanto, realizadas análises da representação numérica nas diferentes formas de agregar os trabalhos de grupos de pesquisadores.

Em relação ao ponto de vista de seus objetivos (GIL, 1991), caracteriza-se como descritiva, uma vez que objetiva descrever e analisar as relações entre as

variáveis, em especial a influência das variáveis acerca das publicações em grupos, nos resultados de quantidade de publicações e citações. Para isso, são utilizadas técnicas padronizadas para a coleta dos dados, com o fim de realizar um levantamento de dados (SILVA; MENEZES, 2005).

Em relação ao ponto de vista dos procedimentos técnicos, esta pesquisa é bibliográfica, documental e experimental. A caracterização bibliográfica se dá pela utilização de materiais já publicados, sendo essa a fonte primária da revisão de literatura. Foi utilizada a base de dados ArXiv[5], com o intuito de analisar as tendências de publicações na temática deste trabalho, caracterizando assim uma pesquisa documental, uma vez que pode identificar materiais que ainda não passaram pela revisão, ou receberam tratamento analítico (SILVA; MENEZES, 2005). A Figura 7 apresenta a classificação da pesquisa quanto a Natureza, Forma de Abordagem, Objetivos e Procedimentos Técnicos, de forma resumida.

Figura 7 - Classificação da Pesquisa



Fonte: Elaborado pela autora

As etapas para a realização da presente pesquisa são apresentadas na próxima seção.

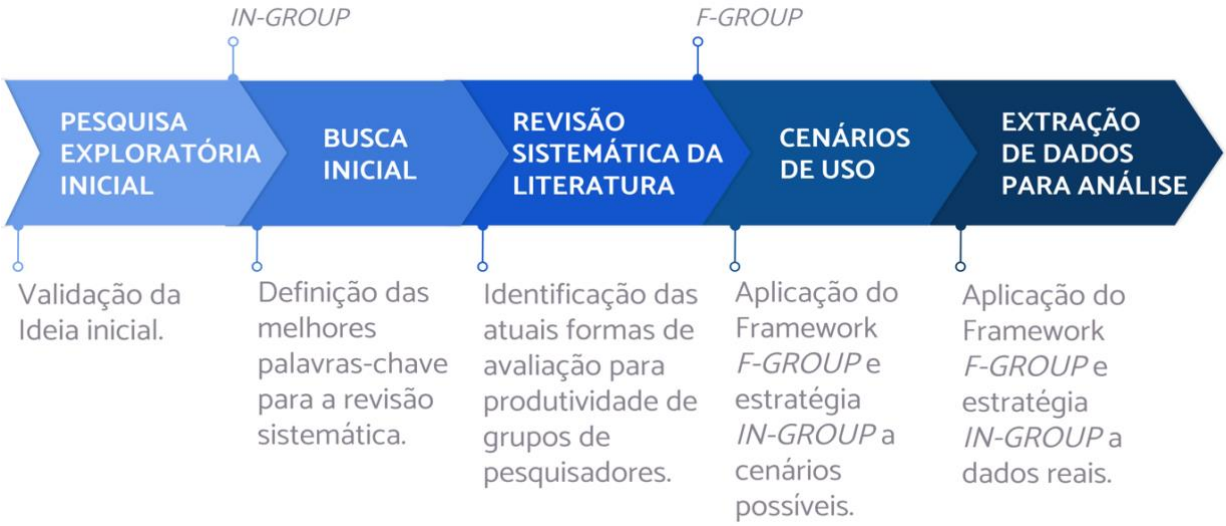
3.2 ETAPAS DA PESQUISA

A presente pesquisa foi realizada seguindo uma série de procedimentos envolvendo coleta e análise de dados. Dentre esses procedimentos destacamos

[5] https://arxiv.org/

aqueles realizadas com buscas em diferentes bases de dados, compilação desses resultados, ou, ainda, com criação de cenários de uso. Na Figura 8 apresentamos as principais etapas realizadas no decorrer da pesquisa.

Figura 8 - Etapas da pesquisa



Fonte: Elaborado pela autora

A primeira busca por artigos científicos relacionados à área de pesquisa foi realizada de forma exploratória em 2020. Os resultados da Pesquisa Exploratória Inicial foram utilizados como subsídio para discussão do objetivo específico “Discutir a relação da ideia de valor do pesquisador e sua relação com os indicadores baseados em quantidade de artigos e citações”. Esta pesquisa permitiu a identificação de diferentes formas de aplicação do índice h para grupos. Consequentemente, também se relaciona com o objetivo “Identificar as formas de avaliação existentes para grupos de pesquisadores através de indicadores”. Seus resultados serviram como referência para a proposta da abordagem *IN-GROUP*, sendo assim se relacionando também com o objetivo: “Propor novas formas de avaliar grupos de pesquisadores, considerando as lacunas presentes nas formas atuais”.

Em seguida, uma busca inicial foi realizada na *Web Of Science* para aprimorar os critérios de busca, antes da Revisão Sistemática. Esta análise inicial dos principais termos utilizados é detalhada na seção 3.3.2. A terceira etapa compreende a Revisão Sistemática da Literatura; etapa esta que se relaciona com o objetivo específico “Identificar as formas de avaliação existentes para grupos de pesquisadores através

de indicadores". Os resultados da Revisão Sistemática de Literatura estão presentes no Capítulo 4 – Estado da Arte. Adicionalmente, foram utilizados como base para a proposta do *framework F-GROUP*, estando assim relacionada com o objetivo: "Propor novas formas de avaliar grupos de pesquisadores, considerando as lacunas presentes nas formas atuais".

Com o intuito de apresentar cenários de aplicação, tanto para a abordagem *IN-GROUP* quanto para o *framework F-GROUP*, foram apresentados Cenários de uso na quarta etapa. Desta forma, a quarta etapa está relacionada com o objetivo: "Avaliar a aplicabilidade dos dispositivos propostos por meio de casos de uso". A última fase realizada foi a extração de dados para aplicação do *framework* proposto em dados reais. Esta etapa se relaciona com o objetivo "Comparar os dispositivos propostos no trabalho com as formas de avaliação consolidadas na literatura". A descrição dos passos utilizados para extração e tratamento destes dados está presente na seção 3.4 deste capítulo.

A apresentação destas etapas é de fundamental importância para relacionar o percurso realizado no decorrer da tese e sua relação com os objetivos apresentados. Na próxima seção são apresentados os critérios utilizados na revisão sistemática da literatura, que compreende as duas primeiras etapas "BUSCA INICIAL" e "REVISÃO SISTEMÁTICA DA LITERATURA".

3.3 REVISÃO SISTEMÁTICA DA LITERATURA

A revisão sistemática da literatura (RSL) foi conduzida com base na abordagem proposta por Bárbara Kitchenham (2009). As seguintes perguntas de pesquisa (RQ) foram tomadas como ponto de partida:

- RQ1: Como tem sido realizada a avaliação da produtividade dos grupos de pesquisadores nos últimos anos?
- RQ2: Quais indicadores de produtividade científica têm sido utilizados para avaliar esses grupos de pesquisadores?
- RQ3: Existe uma metodologia consolidada para a aplicação de indicadores de produtividade em grupos de pesquisadores?

Dada a escassez de trabalhos que tratam da avaliação de grupos de pesquisadores, optou-se por perguntas de pesquisa mais abrangentes, que focam em métricas para avaliação de grupos de pesquisadores. Com o intuito de responder os questionamentos destacados anteriormente e cumprir os objetivos deste trabalho foram realizadas pesquisas em bases de dados, sendo estas listadas a seguir:

- Web of Science;
- Arxiv;
- ACM;
- IEEExplorer;
- Science Direct;
- Scielo;
- SCOPUS;
- BRAPCI.

Houve um levantamento inicial e exploratório acerca dos indicadores existentes que buscam avaliar a produtividade do pesquisador de forma individual, sendo considerada a aplicabilidade desses indicadores para grupos, esta revisão inicial encontra-se no capítulo de conceitos básicos. De forma a elaborar uma pesquisa mais robusta, foi então conduzida uma revisão sistemática da literatura. O objetivo da RSL foi o de identificar trabalhos que atendam aos seguintes critérios:

1. Que proponham indicadores, métricas, *frameworks* ou metodologias para avaliar a produtividade de pesquisadores com base na produção científica deles;
2. Que estes indicadores tenham o propósito de avaliar grupos de pesquisadores ou que apliquem indicadores existentes que analisam a quantidade de artigos e de citações para grupos de pesquisadores.

3.3.1 Critérios de inclusão e exclusão

Após realizar as buscas nas bases definidas, os artigos recuperados foram lidos e selecionados para compor a RSL com base nos seguintes critérios:

- Inclusão
  - Trabalhos que estejam escritos em língua Portuguesa, Inglesa ou Espanhola;

- Trabalhos que proponham alguma forma de avaliar a produtividade de grupos de pesquisadores, conjunto de pesquisadores ou qualquer outra nomenclatura utilizada para designar dois ou mais autores;
- Trabalhos que apliquem métricas específicas para avaliar a produtividade de grupos de pesquisadores.

- Exclusão
  - Trabalhos que tenham sido publicados antes do ano de 2007, com intuito de analisar os últimos 15 anos;
  - Trabalhos que não estejam em língua Inglesa ou Portuguesa ou Espanhola;
  - Trabalhos que não apliquem ou proponham métrica/*framework*/metodologia de avaliação da produção científica a grupos de pesquisadores (fuga de tema).

A limitação de data considera um intervalo de 15 anos, entre 2007 e 2021, inclusos integralmente. Este intervalo foi definido com base em dois critérios principais: a atualidade dos estudos e a posterioridade da proposição do *índice h*, um dos principais indicadores da área. As consultas foram realizadas considerando os argumentos que abrangem título, palavras-chave e *abstract* sempre que possível, ou em todos os metadados quando não era possível selecionar estes campos.

3.3.2 Busca Inicial

Uma busca inicial foi realizada com intuito de aprimorar os termos a serem utilizados na Revisão Sistemática de Literatura. Nesta análise inicial os termos presentes no Título, Palavras-chave e *Abstract* foram verificados. A análise considerou nuvens de palavras, lista de palavras mais utilizadas e, ainda, a lei de Zipf (1949) com o intuito de identificar as palavras-chave mais relevantes. Para compor os critérios da busca inicial foram utilizados os termos "*index*", "*bibliometric*" e "*grupo*", incluindo sinônimos e correlatos, Quadro 4.

O argumento TS foi utilizado por englobar título, palavras-chave e *abstract*, Quadro 5. Como critério temporal, a pesquisa ficou delimitada para artigos após 2007, conforme definido nos critérios de inclusão. A consulta foi realizada na base *Web of*

Science[6], sendo realizada em 22 de outubro de 2021. Como resultado 2066 artigos foram recuperados.

Quadro 4 - Conceitos principais da busca incial

Conceito	**Sinônimos ou equivalentes**
Index	Métrica, Índice
Bibliométrico	Cienciométrico
Grupo	Colégio Invisível, departamento, grupo de pesquisa, universidade, coautores, coautoria

Fonte: elaborado pela autora

Quadro 5 - *String* de busca inicial realizada no *Web Of Science* e total de resultados recuperados.

Pesquisa Realizada	**Total de resultados**

`TS=(Metrics OR metric OR index OR Indice OR Indices OR indexes)`
`AND TS=(scientometric OR cienciometric OR bibliometric OR citation OR citations)`
`AND TS=("author group" OR coauthor OR coauthorship OR "research group" OR "invisible college" OR department OR UNIVERSITY OR universities OR departments OR college)`
	2066

Fonte: elaborado pela autora

A primeira análise foi realizada considerando a Lei de Zipf (1949). Esta Lei consiste em analisar a frequência das palavras-chave, retornando uma lista ordenada da ocorrência de cada termo, somando em todos os textos. Para isto, as palavras-chave foram ordenadas de acordo com o número de ocorrência nos 2066 artigos retornados na busca.

[6] www.webofscience.com

Foram identificados 3419 termos diferentes nas palavras-chave. A palavra mais frequente foi INDEX com 729 registros. A Tabela 1 apresenta as primeiras 15 palavras com o maior número de ocorrências nos artigos retornados da busca inicial. A primeira palavra INDEX foi utilizada como critério na busca inicial. A segunda (SCIENCE) é uma palavra importante, porém pouco específica para o escopo da pesquisa. As Palavras CITATION e COLLABORATION também são importantes e foram classificadas como relevantes para uma possível inclusão no critério de busca.

Tabela 1 - Palavras-chave com o maior número de ocorrências

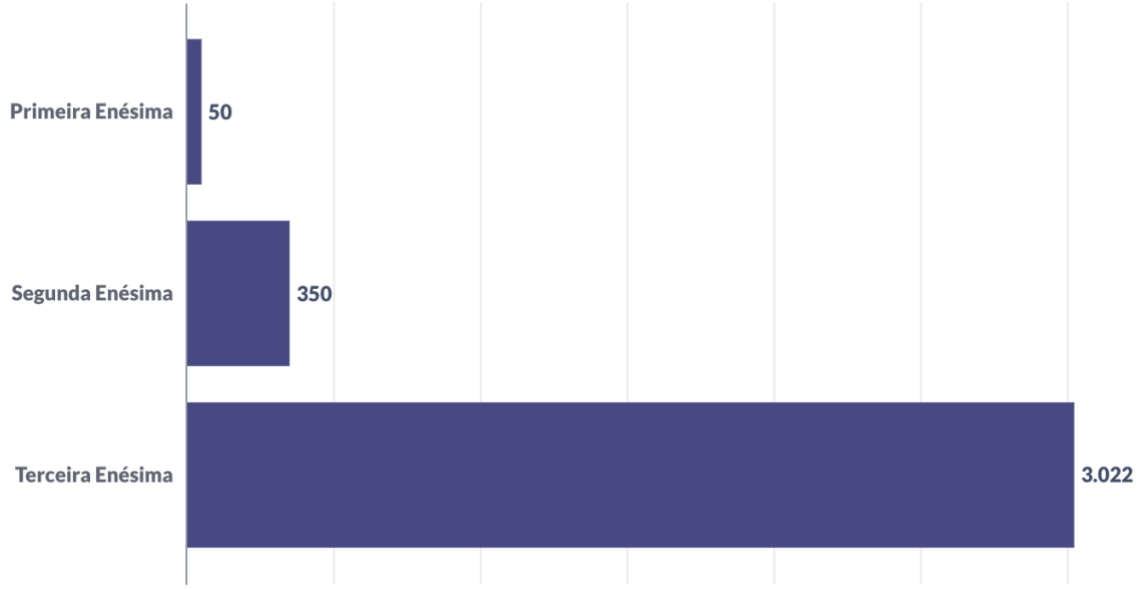
Total de ocorrências	**Palavra-Chave**
729	INDEX
680	SCIENCE
558	CITATION
535	IMPACT
456	SEARCH
446	MEN
312	PUBLICATION
304	H-INDEX
284	INDICATOR
281	PERFORMANCE
276	JOURNALS
261	LABOR
256	INDICATORS
255	AGE
251	COLLABORATION

Fonte: elaborado pela autora

A lista completa de palavras-chave foi dividida em 3 partes, denominadas enésimas (Figura 9), de acordo com a Lei de Zipf (1949). Cada parte correspondendo a ⅓ do total das ocorrências dos termos identificados nas palavras-chave. A primeira enésima, com o total de ⅓ das correspondências em palavras-chave foi de 50 palavras. A segunda enésima ficou 350 palavras e a terceira e última 3022 palavras.

Consequentemente, o primeiro conjunto de palavras-chave possui o mesmo número de ocorrências dos demais grupos. Por este motivo a Lei de Zipf é chamada de Lei do mínimo esforço, por permitir analisar um menor grupo de conceitos, identificando qual é o grupo de palavras-chave mais relevantes.

Figura 9 - Enésimas da Lei de Zipf para identificação de palavras-chave mais relevantes.



Fonte: elaborado pela autora

Os Títulos dos Artigos foram analisados através da extensão do Google Docs[7] que permite a criação de uma nuvem de palavras (Figura 10) e listagem das 10 primeiras palavras com maior número de ocorrências nos textos. Dentre as palavras mais comuns temos RESEARCH e ANALYSIS, que não constam na busca inicial, e, não figuram na listagem das palavras-chave do Quadro 4. BIBLIOMETRIC é a terceira palavra mais utilizada nos títulos e constava na pesquisa inicial, bem como CITATION, UNIVERSITY e INDEX.

Os termos SCIENCE, SCIENTIFIC e IMPACT remetem a conceitos que também foram identificados na análise das palavras-chave, portanto, isso reforça que se trata de conceitos que devem ser analisados para inclusão nos critérios de busca. Além dos termos já identificados através da Tabela 2, ao analisar a Figura 10 os termos BIBLIOMETRIC, INDEX e UNIVERSITY se mostram relevantes para análise e

[7] https://workspace.google.com/marketplace/app/word_cloud_generator/360115564222

possível inclusão. Outros termos também podem ser relevantes para explorar a análise dos grupos como SOCIAL e NETWORK.

Tabela 2 - Palavras com o maior número de ocorrência no Título dos Artigos

Total de ocorrências	Palavra-Chave
747	RESEARCH
591	ANALYSIS
478	BIBLIOMETRIC
263	SCIENCE
242	SCIENTIFIC
186	IMPACT
181	CITATION
177	UNIVERSITY
173	STUDY
155	INDEX

Fonte: elaborado pela autora

Figura 10 - Nuvem de 50 palavras mais utilizadas nos Títulos dos Artigos

Fonte: elaborado pela autora

De forma semelhante, os *abstracts* de todos os Artigos foram analisados através da mesma extensão do Google Docs, criando também uma nuvem de 50 palavras (Figura 11) e listagem das 10 palavras mais utilizadas nos *abstracts* - Tabela

3. Dentre as palavras mais comuns temos RESEARCH sendo utilizada 2245 vezes, mais do que o dobro da segunda palavra. O termo PUBLICATIONS também teve um alto número de utilização, com 1082 ocorrências nos *abstracts*, sendo necessário avaliar a inclusão destes termos no critério de busca. As demais palavras foram encontradas também na lista de palavras-chave e títulos.

Tabela 3 - Palavras com o maior número de ocorrência nos abstracts

Total de ocorrências	Palavra-Chave
2245	RESEARCH
1082	PUBLICATIONS
995	INDEX
927	STUDY
918	ARTICLES
904	SCIENTIFIC
900	CITATION
871	SCIENCE
835	UNIVERSITY
830	NUMBER

Fonte: elaborado pela autora

Figura 11 - Nuvem de 50 palavras mais utilizadas nos abstracts

Fonte: elaborado pela autora

Considerando as palavras trazidas na Figura 11, outros termos interessantes

também passam a figurar. Levando-se em conta que quanto maior o número de ocorrências maior o tamanho da palavra na nuvem, os termos mais comuns nos *abstracts* possuem uma uniformidade maior em relação à nuvem criada com as palavras dos títulos, Figura 10. As palavras INDICATOR, PRODUCTIVITY, SOCIAL e COLLABORATION foram consideradas relevantes para análise e inclusão no critério aprimorado de busca.

A análise das palavras-chave, *abstracts* e títulos permitiu que algumas conclusões fossem alcançadas. A primeira é que o termo Cienciométrico não traz um número de resultados relevante em nenhum dos campos analisados, podendo ser removido do critério de busca. Outra conclusão é que o termo colaboração possui diversas grafias diferentes, que somadas, tornam este termo ainda mais relevante. Portanto, todas as grafias precisam ser incluídas no critério de busca para a Revisão Sistemática de Literatura.

Outra palavra importante que não estava sendo utilizada é o termo INDICATOR, que havia sido traduzido como INDEX apenas. Ao analisar os conceitos chave apresentados no Quadro 4, ficou evidente a necessidade de se adicionar novos conceitos. O critério de busca foi expandido, passando a contar com 5 (cinco) conceitos principais, que são apresentados no Quadro 6, com seus respectivos sinônimos, equivalentes e traduções.

Através da análise realizada foi possível identificar termos que não estavam contribuindo com a pesquisa, pois não se mostraram relevantes em nenhum dos campos analisados. Adicionalmente, identificamos novos conceitos chave para serem incluídos nos critérios de pesquisa. Sendo assim, propomos um novo critério para busca com palavras mais relevantes e com inclusão dos termos identificados na Busca Inicial.

3.3.3 Realização da busca com os termos definidos

A busca foi realizada em 8 bases de dados, em 04 de janeiro de 2022, com a identificação inicial de cerca de 2 mil artigos. Após análise de título, *abstract* e texto completo, 38 artigos foram selecionados de acordo com os critérios definidos.

No Quadro 7 é possível visualizar as pesquisas realizadas, juntamente com a base de dados onde foi feita cada pesquisa, a *string* de busca utilizada e o total de

resultados recuperados. Há diferença entre o poder de expressividade nas bases de dados e, em algumas, foi necessário realizar várias consultas para conseguir buscar todos os termos necessários ou adaptar a *query*. Em todas as consultas foi estabelecido o critério de acordo com o definido no Quadro 6.

Quadro 6 - Conceitos principais da busca revisado

Conceito	**Sinônimos ou Equivalentes**	**Traduções**
Indicador	Métrica, Índice	*Index, Metric, Indicator*
Colaboração	colaboração, coautoria, co-autoria, coautor, coautores, pesquisadores, rede	*collaboration, coauthor, coauthorship, co-author, co-authors, co-authorship, researchers, college, network*
Produtividade	Performance, Qualidade, Produção	*Performance, Quality, Productivity, Production*
Citação	Citado, Citar	*Cite, Cited, Citation*
Bibliométrico	-	*Bibliometric*

Fonte: elaborado pela autora

Quadro 7 - Pesquisas realizadas e total de resultados

Base de Dados	***String* de Busca**	**Total de resultados**
Web of Science	`(TS=(index OR indexes OR metric OR metrics OR indicator OR indicators OR métrica OR métricas OR indicador OR indicadores OR índice or índices) AND TS=(bibliometric OR bibliometrics OR bibliometria OR bibliométrico) AND TS=(cite OR cited OR citation OR citar OR citado OR citação) AND TS=(performance OR quality OR productivity OR production OR qualidade OR produtividade OR produção) AND TS=(collaboration OR coauthor OR coauthorship OR co-author OR co-authors OR co-authorship OR colaboração OR coautoria OR co-autoria OR coautor OR coautores)) AND LANGUAGE:(English OR Portuguese OR Spanish)` ● Com critério de ano entre 2007 e 2022	**729**
SCOPUS	`( TITLE-ABS-KEY ( index OR indexes OR metric OR metrics OR indicator OR indicators OR metrica OR metricas OR indicador OR`	**637**

	indicadores OR indice OR indices) AND TITLE-ABS-KEY (bibliometric OR bibliometrics OR bibliometria OR bibliometrico) AND TITLE-ABS-KEY (cite OR cited OR citation OR citar OR citado OR citacao) AND TITLE-ABS-KEY (performance OR quality OR productivity OR production OR qualidade OR produtividade OR producao) AND TITLE-ABS-KEY (collaboration OR coauthor OR coauthorship OR co-author OR co-authors OR co-authorship OR colaboracao OR coautoria OR co-autoria OR coautor OR coautores) AND PUBYEAR > 2006)	
ACM Digital Library	[[All: index] OR [All: indexes] OR [All: metric] OR [All: metrics] OR [All: indicator] OR [All: indicators] OR [All: indices]] AND [[All: bibliometric] OR [All: bibliometrics] OR [All: bibliometria] OR [All: bibliométrico]] AND [[All: cite] OR [All: cited] OR [All: citation] OR [All: citar] OR [All: citado] OR [All: citação]] AND [[All: performance] OR [All: quality] OR [All: productivity] OR [All: production] OR [All: qualidade] OR [All: produtividade] OR [All: produção]] AND [[All: collaboration] OR [All: coauthor] OR [All: coauthorship] OR [All: co-author] OR [All: co-authors] OR [All: co-authorship] OR [All: colaboração] OR [All: coautoria] OR [All: co-autoria] OR [All: coautor] OR [All: coautores]] AND [Publication Date: (01/01/2007 TO 01/31/2022)]	**594**
IEEE Xplore	("All Metadata":bibliometric OR "All Metadata":bibliometrics OR "All Metadata":bibliometria OR "All Metadata":bibliométrico) AND ("All Metadata":index OR "All Metadata":indexes OR "All Metadata":metric OR "All Metadata":metrics OR "All Metadata":indicator OR "All Metadata":indicators OR "All Metadata":indice OR "All Metadata":indices) AND ("All Metadata":cite OR "All Metadata":cited OR "All Metadata":citation OR "All Metadata":citar OR "All Metadata":citado OR "All Metadata":citação) AND ("All Metadata":collaboration OR "All Metadata":coauthor OR "All Metadata":coauthorship OR "All Metadata":co-author OR "All Metadata":co-authors OR "All Metadata":co-authorship OR "All	**43**

	Metadata":colaboração OR "All Metadata":coautoria OR "All Metadata":co-autoria OR "All Metadata":coautor OR "All Metadata":coautores) ● Com critério de ano entre 2007 e 2022	
Science Direct	Title, abstract or author-specified keywords (index OR indicator) AND (bibliometric) AND (cite OR citation) AND (performance OR production) AND (collaboration OR co-author) ● Com critério de ano entre 2007 e 2022 ● Pesquisa limitada a 8 operadores booleanos: combinados os conceitos de forma que trouxesse o maior número de resultados ● Plurais são incluídos de forma automática ● A tradução para português não retornou resultados	**54**
Scielo	TS=(index OR indexes OR metric OR metrics OR indicator OR indicators OR métrica OR métricas OR indicador OR indicadores OR índice or índices) AND TS=(bibliometric OR bibliometrics OR bibliometria OR bibliométrico) AND TS=(cite OR cited OR citation OR citar OR citado OR citação) AND TS=(performance OR quality OR productivity OR production OR qualidade OR produtividade OR produção) AND TS=(collaboration OR coauthor OR coauthorship OR co-author OR co-authors OR co-authorship OR colaboração OR coautoria OR co-autoria OR coautor OR coautores) ● Com critério de ano entre 2007 e 2022	**84**
arXiv.org	order: -announced_date_first; size: 200; include_cross_list: True; terms: AND all=bibliometric; AND all=index; AND all=citation; AND all=collaboration	**13**
BRAPCI	INDIC* CO*AUT*	*82*
	INDIC* BIBLIOM* COAUT*	*52*
	METR CO*AUTOR*	*125*
	Total BRAPCI sem duplicados	**149**
TOTAL		**2311**

Fonte: elaborado pela autora

O total de trabalhos recuperados, sem verificação de duplicados, foi de 2301. Através da ferramenta StArt[8] foi possível identificar cerca de 98 artigos replicados (Figura 12). A ferramenta StArt permite dividir a RSL em 3 etapas: a primeira é a etapa de Planejamento, onde foram definidas as questões de pesquisa, critérios de inclusão e exclusão e bases de dados que seriam analisadas.

Figura 12 - Imagem extraída da ferramenta StArt, onde os passos da revisão sistemática são demonstrados

Fonte: Extraído da ferramenta StArt (2022).

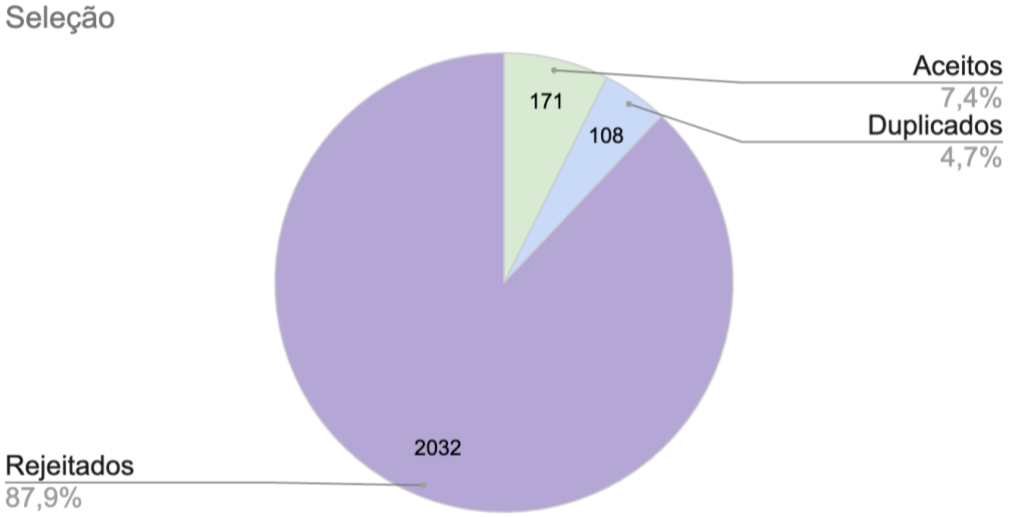
Em seguida, vem a etapa de Execução (Figura 12), que é dividida em Identificação do Estudo, Seleção e Extração. Na Identificação do Estudo são realizadas as buscas de acordo com as *strings* definidas no Quadro 7. Após a recuperação dos dados foram aplicados os critérios de inclusão e exclusão, definidos na subseção 2.2.1. Foram lidos os títulos e *abstracts*, sendo selecionados 171 artigos (Figura 13) para leitura completa. Nesta etapa, foram identificados 108 artigos duplicados e 2032 foram rejeitados por não atenderem os critérios definidos.

Grande parte dos trabalhos recuperados atendem o critério de estarem escritos em língua Portuguesa, Inglesa ou Espanhola. A maioria das rejeições ocorreu devido à falta de utilização de métricas para avaliar um grupo de pesquisadores, terceiro critério de inclusão. Muitos trabalhos traziam análises descritivas dos grupos analisados, não apresentando qualquer métrica ou indicador sobre a produção de artigos científicos relacionados a um grupo. Alguns trabalhos realizavam apenas análises visuais, outros, traziam predição de performance acadêmica, ou análises de redes sociais, mas sem indicadores numéricos. Também foram identificados trabalhos

[8] http://lapes.dc.ufscar.br/tools/start_tool

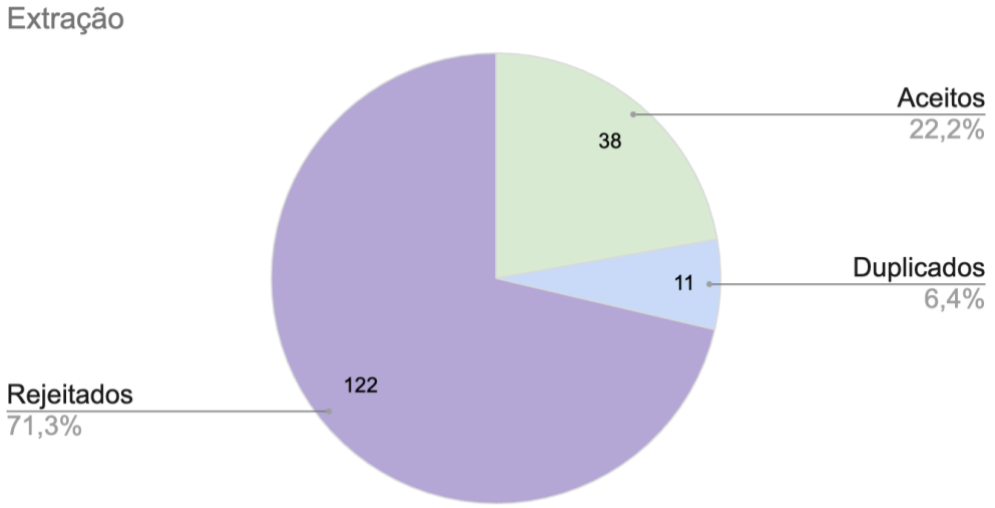
que apresentavam indicadores sobre patentes e inovações, sendo excluídos por também não atenderem aos critérios de inclusão e exclusão.

Figura 13- Gráfico de Pizza dos artigos selecionados na fase de Seleção



Fonte: Adaptado da ferramenta StArt (2023)

Figura 14- Gráfico de Pizza dos artigos Aceitos na fase de Extração

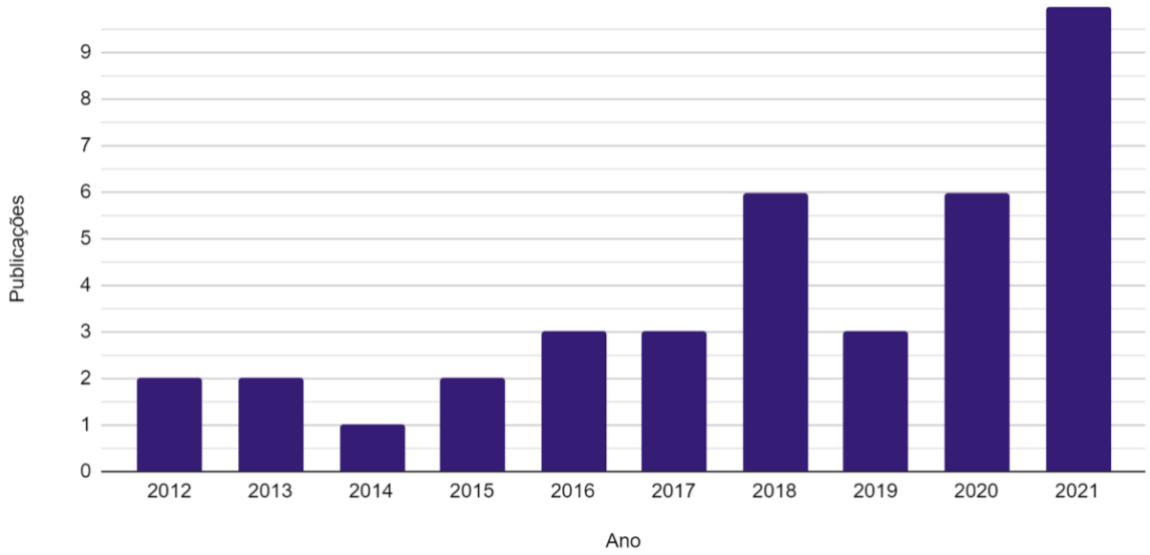


Fonte: Adaptado da ferramenta StArt (2023).

Os 171 artigos foram então analisados de acordo com os critérios definidos na seção 2.2.1. Nesta etapa foram identificados mais 9 artigos duplicados e 121 deles foram rejeitados por não atenderem os critérios. No total, 38 artigos que atendiam a todos os critérios de inclusão e exclusão foram identificados (Figura 14). Os resultados foram divididos em 3 grandes grupos: 1) Estudos de casos que utilizam indicadores por país ou instituição; 2) Trabalhos que utilizam análises de redes de coautoria; 3) Trabalhos que propõem novos índices ou novas metodologias para análise de grupos

de pesquisadores. Dentre os 38 artigos selecionados, destacamos na Figura 15 um histograma que evidencia a crescente de trabalhos sobre o tema pesquisado ao longo dos anos.

Figura 15 - Histograma por ano das 38 publicações selecionadas



Fonte: elaborado pela autora

Após o cumprimento das etapas de Planejamento e Execução, procedeu-se a uma sumarização. Nesta etapa, cada artigo foi lido e resumido de acordo com as necessidades identificadas nas perguntas de pesquisa. A sumarização dos 38 artigos selecionados resultou no Estado da Arte, que é apresentado no Capítulo 4. A seguir, mostra-se como foi feito o tratamento de dados aplicado na análise com dados reais.

3.4 EXTRAÇÃO E TRATAMENTO DE DADOS PARA EXPERIMENTOS COM DADOS REAIS

A última etapa envolvendo a coleta e a análise de dados compreende a realização de experimentos com dados reais. O presente trabalho possui dentre seus objetivos específicos a proposição de novas formas para avaliar grupos de pesquisadores, bem como comparar os dispositivos propostos no trabalho com as formas de avaliação consolidadas na literatura. Desta forma, foram coletados dados que permitam a experimentação das novas formas de avaliação propostas. A coleta desses dados tem por objetivo realizar experimentos em diferentes agrupamentos de autores existentes no mundo real. Para isso foram reunidos dados de publicações e

os grupos foram formados a partir da informação de países e universidades.

Os dados foram coletados em 10 de abril de 2023. São resultantes de uma pesquisa realizada na base de artigos científicos *Web Of Science*. Para extração das informações foi selecionado o Tópico *Information Science & Library Science* com controle de data dos últimos 10 anos, ou seja, do período entre 10/04/2013 e 10/04/2023. Foram recuperados 121.222 resultados. Mais detalhes sobre o processo de coleta dos dados estão presentes no APÊNDICE A - Coleta dos dados.

Ao trazer a tabela de artigos para um banco de dados relacional, foram realizadas verificações acerca da qualidade dos dados. Cada registro na tabela extraída da *Web Of Science* e importada para o PostgreSQL[9] representa um artigo com ao menos uma citação. Cada um desses artigos pode ter múltiplos autores, que por sua vez podem possuir mais de uma afiliação. Desta forma, cada artigo pode estar relacionado a múltiplas instituições e países.

Iniciamos o tratamento dos dados com as informações dos nomes dos Autores. A forma original que representa a listagem com os nomes dos autores em cada artigo respeita a representação SOBRENOME, NOME; [SOBRENOME, NOME]; (...). Criou-se uma tabela com nome e sobrenome de cada Autor, da forma como estavam originalmente descritos. Ou seja, cada registro desta tabela representa um registro diferente de SOBRENOME, NOME. Para a transformação de uma coluna única em vários registros diferentes, utilizamos a função SPLIT_PART do PostgreSQL. Esta função separa as diferentes afiliações listadas para cada um dos artigos.

Ainda na tabela recém-criada com os Autores, criamos uma segunda coluna representando o nome dos autores na forma NOME SOBRENOME. A utilização desta representação foi necessária porque alguns autores possuem múltiplos sobrenomes e nomes compostos e nem sempre a divisão SOBRENOME, NOME ocorre da mesma forma para um mesmo autor. Por exemplo, o autor Luiz Augusto Silva pode ser representado na forma: Silva, Luiz Augusto ou Augusto Silva, Luiz. A tabela visou representar cada autor de forma única e um identificador único foi associado a cada um dos autores.

Criamos também uma tabela para representar o relacionamento entre os artigos e os autores, considerando as diferentes formas de grafias possíveis dos nomes dos autores. O relacionamento possui um conjunto de pares únicos de

[9] https://www.postgresql.org/

identificador de Artigo e identificador de Autor. O número máximo de autores em único artigo, identificado no conjunto de dados analisado, ficou com 99 autores.

Observando a coluna que designa a afiliação dos autores observamos o seguinte exemplo:

[Centobelli, Piera; Oropallo, Eugenio] Univ Naples Federico II, Dept Ind Engn, Ple Tecchio 80, I-80125 Naples, Italy; [Cerchione, Roberto] Univ Naples Parthenope, Dept Engn, Ctr Direz Napoli Isola C4, I-80143 Naples, Italy; [Del Vecchio, Pasquale; Secundo, Giustina] Univ LUM Giuseppe Degennaro, Dept Management Finance & Technol, SS 100, Km 18, Casamassima Bari, Italy; [Cerchione, Roberto] Univ Naples Parthenope, Dept Engn, Ctr Direz Napoli Isola C4, I-80143 Naples, Italy

Neste exemplo temos uma listagem com quatro afiliações diferentes. A primeira delas corresponde aos autores Piera Centobelli e Eugênio Oropallo, da Universidade de Nápoles Federico II, com uma identificação referente a um departamento específico, endereço e, por último, a informação referente ao país que corresponde àquela afiliação. De forma mais organizada, o mesmo exemplo pode ser visto como uma listagem:

- [Centobelli, Piera; Oropallo, *Eugenio] Univ Naples Federico II, Dept Ind Engn, Ple Tecchio 80, I-80125 Naples*, **Italy**;
- [Cerchione, Roberto] *Univ Naples Parthenope, Dept Engn, Ctr Direz Napoli Isola C4, I-80143 Naples*, **Italy**;
- [Del Vecchio, Pasquale; Secundo, Giustina] *Univ LUM Giuseppe Degennaro, Dept Management Finance & Technol, SS 100, Km 18, Casamassima Bari,* **Italy**;
- [Cerchione, Roberto] *Univ Naples Parthenope, Dept Engn, Ctr Direz Napoli Isola C4, I-80143 Naples*, **Italy**

Na coluna afiliação, apresentada acima na forma original, tal como foi exportada da *Web of Science*, observamos informações detalhadas da afiliação para cada autor (ou lista de autores): endereço, universidade, e, em alguns casos, o departamento. O país associado sempre está presente e possui nome padronizado, sendo, portanto, de fácil normalização. Como o intuito da análise é observar os grupos

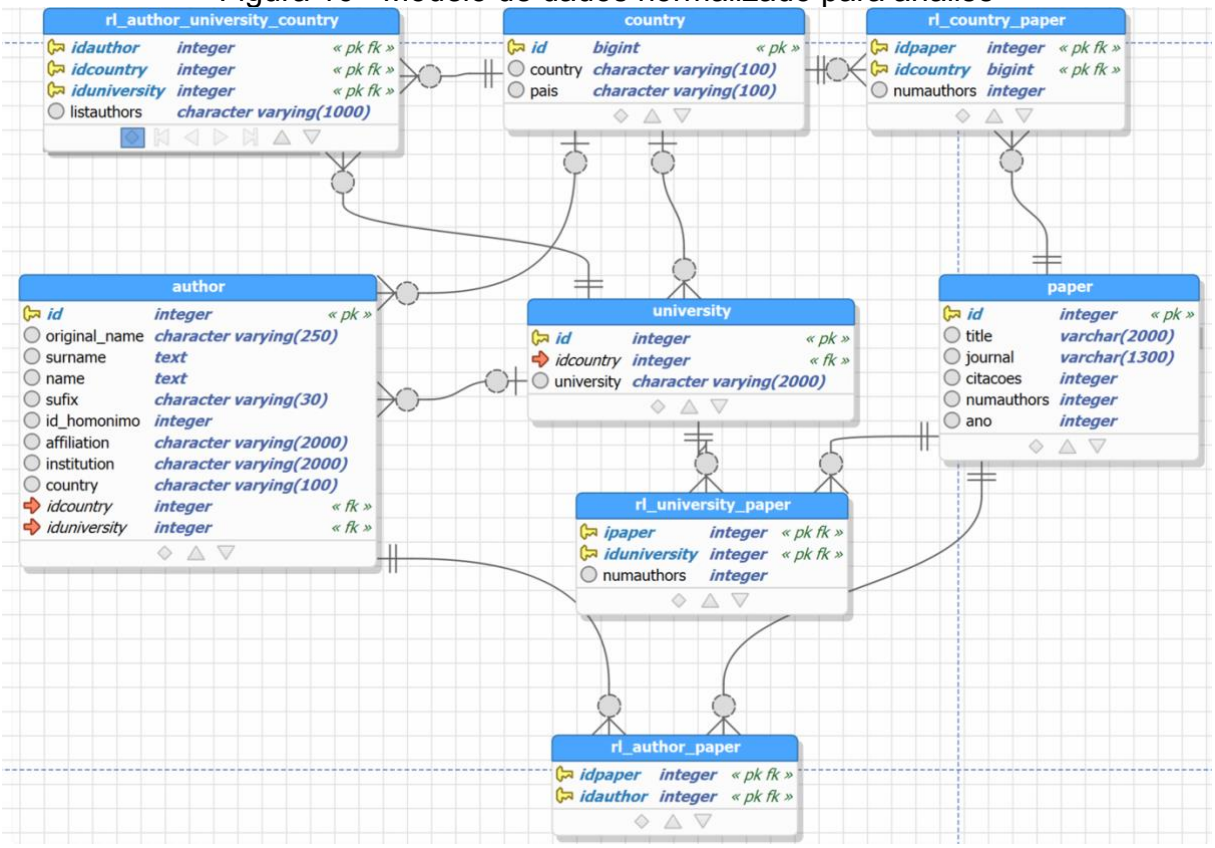
de autores houve a criação de relacionamentos entre universidades e artigos, e entre países e artigos.

O modelo de dados após normalização é apresentado na Figura 16. A tabela *Paper* representa os artigos importados, incluindo algumas colunas originais e informações sumarizadas, como o número de autores, por exemplo. Além delas, o modelo apresenta Author, University e Country, bem como os respectivos relacionamentos.

Todo o código em SQL referente ao processo de normalização está disponível no Apêndice B. Este mesmo Apêndice contém também as consultas em SQL e a criação de *views* feitas para facilitar a análise dos dados.

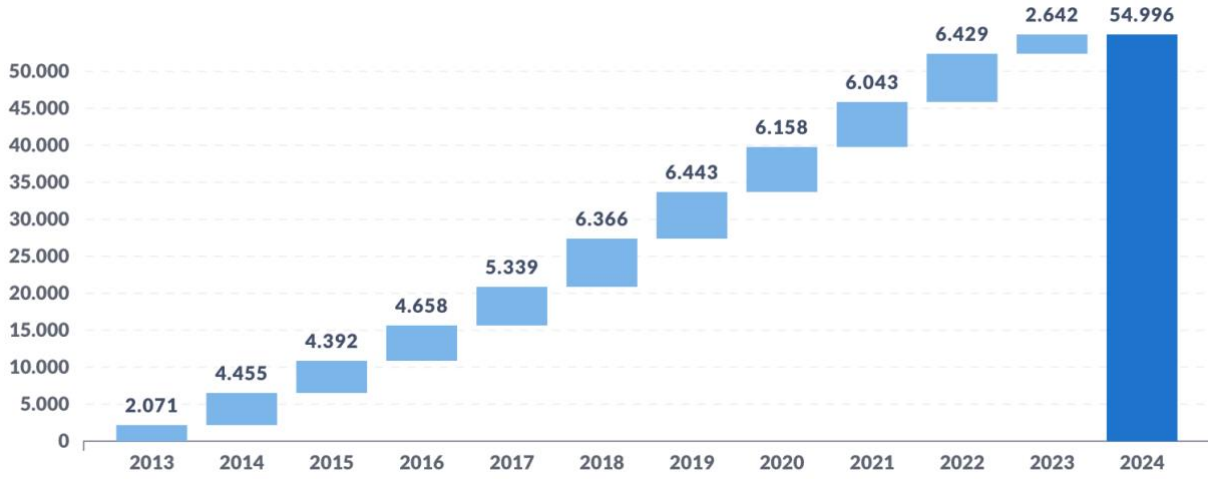
A partir da normalização dos atributos multivalorados de Author, incluindo a informação da coluna Affiliation, foi possível identificar que no conjunto de dados analisados existem artigos de 171 países diferentes. Um total de 54.998 diferentes artigos foram importados, uma vez que somente os artigos que possuem citações foram recuperados. Dois artigos não vieram com o valor do campo Ano preenchido. A média de autores em cada artigo foi de 3,33. A média de citações por artigo foi de 13,84.

Figura 16 - Modelo de dados normalizado para análise



Fonte: dados da pesquisa (2023).

Figura 17 - Gráfico Cascata com total de artigos por Ano entre 10 de abril de 2013 e 10 de abril de 2023



Fonte: dados da pesquisa (2023).

A Figura 17 apresenta um gráfico em cascata com o total de artigos por ano. É possível observar um crescimento anual constante no total de artigos. Verifica-se que entre 2014 e 2016 a média é de 4501 artigos por ano. Em 2017 esse número já passa de 5 mil artigos e, a partir de 2018, ultrapassa o número de 6 mil artigos ao ano. No ano de 2023 os dados se limitam até o dia 10 de abril. Por este motivo, apresentam valor menor. A análise dos dados e aplicação está disponível no capítulo 6. A seguir, apresenta-se o resultado da Revisão Sistemática de Literatura.

4 ESTADO DA ARTE

Neste capítulo é apresentada a sumarização dos 38 trabalhos correlatos, identificados na Revisão Sistemática de Literatura, que considera o cenário dos últimos quinze anos. Para cada um dos artigos foram identificados 4 aspectos principais: i) Quais indicadores foram utilizados para avaliar a produtividade do grupo em questão? ii) Qual grupo foi analisado (país, instituição etc.)? Se a análise realizada foi com foco genérico de propor uma medida para avaliar a produtividade de grupos ou se era um estudo de caso visando avaliar a produtividade de um grupo específico.

Este capítulo é dividido em 4 subseções: A primeira apresenta os trabalhos sobre estudos bibliométricos de áreas do conhecimento específicas com indicadores por países e instituições. A segunda seção apresenta trabalhos que focam na análise de produtividade através de redes de coautoria. A terceira é o grupo de trabalhos mais correlatos à proposta deste trabalho, que propõem alguma forma de medir a produtividade em grupos de pesquisadores. Na última subseção é apresentada uma sumarização dos resultados, considerando cada questão de pesquisa definida.

4.1 ESTUDOS DE CASO EM PAÍSES E INSTITUIÇÕES

Alguns trabalhos apresentam uma análise da produção bibliográfica de uma área de pesquisa em específico. Neles, é bastante comum que se apresente o total de produções de determinada temática em um país, sendo essa uma medida que pode ser aplicada na análise de grupos mais genéricos. Um exemplo é o trabalho que realiza uma análise da produção e colaboração na área de Viticultura por país e universidades (ALEIXANDRE et al., 2013b). Em outro trabalho semelhante, foi conduzida uma avaliação da colaboração da Universidade Politécnica de Valencia – UPV (CODINA-CANET, 2012). Dentre os resultados, os autores apresentam que a UPV mantém relações de coautoria com 665 instituições de 60 países, sendo os Estados Unidos o principal país colaborador em número de instituições. A produção em colaboração científica foi analisada de forma inter-regional, intrassetorial, intersetorial e intrarregional. Ou seja, houve análise e sugestão de indicadores para um grupo de autores. Os autores da UPV foram analisados como um grande grupo, e essa análise tomou como base o número de artigos científicos produzidos na instituição.

Uma metodologia para computar a produtividade territorial dos pesquisadores foi proposta por Abramo e D'Angelo (2015). Este trabalho analisou a produção científica de províncias italianas, por meio da utilização de indicadores globais por província e por área de estudo, tais como: total de pesquisadores, total de universidades, total de artigos e total de campos de pesquisa. Foi proposta uma medida de produtividade para cada autor, denominada *Fractional Scientific Strength* (FSS) ou Força Científica Fracionada, que considera a quantidade de anos desde a data de publicação, o total de citações e a contribuição do autor em cada artigo (considerando área de estudo e posição do nome de cada autor no artigo). Esta medida é então normalizada (FSS^N) para avaliar a produtividade nas províncias.

Em 2015 foi analisada a colaboração científica como procedimento para a análise de um domínio, aplicando-se esta teoria para a área de Zootecnia nas universidades UFV (Universidade Federal de Viçosa), UNESP (Universidade Estadual Paulista) e USP (Universidade de São Paulo) (ROSAS, 2015). Este trabalho utilizou como indicadores o número de citações (com ou sem colaboração, considerando ainda colaboração internacional), número de artigos e grau na rede de coautoria. Ao final, os autores identificaram que os trabalhos com colaboração internacional tendem a ter um maior impacto que os demais trabalhos.

Um estudo bibliométrico sobre a conferência finlandesa *Koli Calling* foi realizado por Simon (2016). O trabalhou apresenta análises sobre os principais autores, a colaboração entre autores e países, e, ainda, a entrada de novos autores na comunidade. Concluiu-se que a comunidade se encontra com um núcleo sólido de autores, mas que está aberta para a entrada de novos autores. O padrão de produtividade autoral adere à lei de Lotka, que descreve a frequência de publicação de autores.

Em 2017 foi realizada uma análise de processos de produção de novos conhecimentos em regiões-chave do mundo e na Rússia (GOLICHENKO; MALKOVA, 2017). Neste trabalho foram adotados métodos para analisar importação e exportação de conhecimentos, apresentando esses valores por país. Realizou-se uma estimativa sobre a qualidade do conhecimento produzido e foram analisados os fatores que influenciam na qualidade do novo conhecimento. Neste mesmo ano foi realizado um estudo sobre domínio e liderança em atividades de pesquisa (GONZALEZ-ALCAIDE *et al.,* 2017). O artigo apresenta número de documentos e percentagem de colaboração científica por país. A ordem das assinaturas dos autores e endereços

para correspondência em publicações científicas são usadas como variáveis para analisar as interações entre países entre: muito alto, alto, médio e baixo desenvolvimento humano. Os autores indicam que a colaboração entre países de desenvolvimento humano diferenciado se reflete por meio de ordem de autoria e designação como autores correspondentes em publicações científicas.

Diaz-Cardenas e sua equipe (2017) examinaram as características da produção científica em instituições no estado de Puebla - México. Os autores identificaram as instituições mais produtivas, e apresentaram dados como frequência e porcentagem de artigos citados por instituição. A análise foi focada nos 102 artigos mais citados do período analisado (2008-2012). Seguindo a tendência, os artigos com múltiplos autores (ao menos três) obtiveram o maior impacto em citações. O trabalho também apresentou uma análise sobre diferenças no fator de impacto em citações entre diferentes áreas da ciência.

Em 2018, sob uma nova perspectiva de dados baseada em concessões (ou doações), foi possível identificar os parceiros internacionais de colaboração em pesquisa da China (YUAN *et al.*, 2018). Usando dados da Fundação Nacional de Ciências Naturais da China (NSFC) entre a China e 75 países, de 2006 a 2016, o estudo examinou os parceiros de colaboração em três aspectos: 1) atividade geral de colaboração, 2) esforço relativo de pesquisa e 3) grupos de colaboração dos países. O estudo demonstrou que a China obteve uma preferência crescente por colaborar com Austrália, Holanda e Espanha, em detrimento de participações com Japão, Alemanha e Suécia. Naquela altura, o G7 e a Ásia-Pacífico possuíam mais de 75% de todas as colaborações, enquanto os BRICS e *The Belt and Road Initiative* (BRI) se mostravam relativamente fracos em colaboração de pesquisa, mesmo com a China mantendo a tendência crescente de colaboração com países desses blocos.

Em Belgrado, também foi realizada uma pesquisa sobre o desempenho intrainstitucional, na forma de um estudo de caso, por meio de uma análise quantitativa-qualitativa aprofundada do desempenho das instituições líderes da Universidade de Belgrado (UB) (PILCEVIC; JEREMIC; VUJOSEVIC, 2018). Essa pesquisa apurou diferentes faculdades, como Medicina e Física, por exemplo, e os autores empregaram a mediana da AVG_JIF_PERCENTILE como métrica para avaliar cada faculdade. A $JIF_{PERCENTILE}$ é calculada de acordo com o número de artigos de determinada área, juntamente com um *ranking* descendente previamente determinado.

As coautorias internacionais do Brasil em estudos métricos da informação, bem como os seus canais de comunicação foram abordados por Meschini, Alves E Oliveira (2018). Neste estudo os autores procuraram identificar com quais países o Brasil estabelece mais intensamente suas relações de coautoria em Estudos Métricos da Informação (EMI) e quais são seus principais canais de comunicação. O intuito do estudo foi avaliar a inserção da produção científica do Brasil no mundo, bem como os principais canais de comunicação utilizados nos EMI durante o período de 2011 a 2016. Ao final, apresentam-se indicadores de total de artigos por país e média de citações por artigo.

Um estudo de caso sobre artigos altamente citados em revistas de Biblioteconomia e Ciência da Informação foi apresentado em 2019 (SAHOO *et al*., 2019). Este trabalho apresenta indicadores comuns como total de artigos e citações por países e universidades/institutos de pesquisa. Como resultados, a Holanda é destaque dentre os países estudados, com a primeira colocação no *ranking* de autores, contando com 17,2% dos autores. Esse destaque é atribuído ao pioneirismo da universidade holandesa *Leiden University* com maior número de contribuições.

Outro estudo de caso, apresentado por Çakir (2019), considerou universidades turcas e seus trabalhos em conjunto com o CERN (antigo acrônimo para *Conseil Européen pour la Recherche Nucléaire*, atual *European Organization for Nuclear Research*). Os autores buscavam identificar o impacto de multiautorias nos *rankings* universitários, analisando a influência de artigos de vários autores nas metodologias de classificação. O estudo revelou que a presença de artigos do CERN com vários autores tem um impacto significativo em todos os indicadores de classificação bibliométrica. Os autores argumentam que, consequentemente, as posições de classificação das instituições em todo o mundo foram afetadas.

Uma análise sociométrica das relações entre os autores, coautores e instituições de ensino no grupo de trabalho 4 (GT4) do ENANCIB[10] entre 2003 e 2018 foi apresentada por Silva *et al.* (2019). A autora realizou análises de redes sociais utilizando o software UCINET[11], sendo possível identificar o aumento da quantidade de publicações entre 2003 e 2018 dos temas em Gestão da Informação (GI) e Gestão do Conhecimento (GC). Neste estudo, foi observada uma congruência entre as

[10] Encontro Nacional de Pesquisa em Ciência da Informação. http://enancib.ibict.br/
[11] http://www.analytictech.com/archive/ucinet.htm

universidades e pesquisadores mais produtivos no GT4. O estudo apresentou um total de artigos e percentagem por instituição.

Em 2020, foi realizado um estudo sobre a incidência de fatores sociais na pesquisa científica, aplicando análise cienciométrica a países estratégicos relacionados ao México (LANCHO-BARRANTES; CANTU-ORTIZ, 2020). O objetivo foi analisar como a sociedade influencia a pesquisa, na contramão do que é usualmente proposto. No estudo foram apresentadas medidas por país: colaboração nacional, internacional e institucional, além de propor medidas de sedentarismo e transitoriedade em ciência. Indica também a relação entre população, investimento em educação superior e gasto interno bruto em Pesquisa e Desenvolvimento, relacionados ao número de citações. Outro estudo de caso avaliou a influência da instituição de correspondência do autor no impacto científico da colaboração em instituições brasileiras (GRACIO *et al.*, 2020). As publicações foram divididas entre colaborações nacionais e estrangeiras. Neste trabalho são utilizadas diversas medidas relacionadas ao número de documentos e impacto de publicações em colaboração nacional e estrangeira. O estudo concluiu que o indicador de autor correspondente fornece informações relevantes para instituições brasileiras em colaboração internacional.

Também em 2020, foram avaliados os efeitos das colaborações de pesquisa no impacto científico, conforme medido pelo indicador *Field Weighted Citation Impact* (CICERO; MALGARINI, 2020). São comparadas diferentes formas de colaborações, inclusive em países diferentes, mas, sem propor uma medida para análise de grupos de pesquisa. As colaborações são distinguidas de acordo com a origem geográfica: nacional, internacional, institucional; bem como de acordo com a afiliação dos pesquisadores: acadêmicas ou corporativas. Os autores evidenciaram o aspecto positivo das colaborações internacionais e nacionais no impacto da pesquisa, sendo o impacto das colaborações internacionais mais evidente e independente do campo considerado.

Uma análise bibliométrica sobre “Biblioteconomia e Ciências da Informação” com base na *Web of Science* foi realizada por Thompson (2020). Foram demonstradas métricas por país, como número de institutos e porcentagem do total de artigos recuperados. Dentre os resultados demonstrou-se que as pesquisas em Biblioteconomia e Ciência da Informação são fortes na colaboração uma com a outra, mas as colaborações internacionais e interdisciplinares ainda são baixas na amostra

analisada (2007-2016). Nesta análise, mais uma vez se destaca que a maior contagem de citações se dá nas publicações com coautoria. Outro estudo bibliométrico foi proposto na área de Indicadores Sociais (KUMAR *et al.*, 2021). A pesquisa demonstrou resultados por país e instituição, como total de publicações, total de publicações com citação, total de citações, dentre outros. Neste cenário, a quantidade de autores ficou evidenciada, mais uma vez, como um dos aspectos que influencia na contagem de citações.

Uma avaliação da produção de publicações em nível de país foi realizada em 2021 (MORENO-DELGADO; GORRAIZ; REPISO, 2021). O foco desta avaliação foi na comunicação de campo de pesquisa, sendo utilizado o Fator de Impacto de Garfield. Neste trabalho foi proposta uma medida de impacto a nível nacional, que aplicou o Fator de Impacto de Garfield para países ao invés de periódicos. Ainda em 2021, foi proposta uma ferramenta de avaliação bibliométrica para análise de contribuições e atividade científica em Ciências Sociais e Humanas (GREGORIO-CHAVIANO *et al.*, 2021). A ferramenta proposta (*Dialnet Métricas*[12]) apresenta indicadores para avaliar o impacto da pesquisa em diferentes níveis, permitindo análises contextualizadas em diferentes níveis, podendo chegar a áreas e universidades.

Outro trabalho analisa o impacto das métricas de pesquisa nos indicadores para *rankings* de universidades na Índia (PAKKAN *et al.*, 2021). Dentre os indicadores pode-se destacar: Quantidade de Produção Acadêmica, Qualidade de Publicação Citada, Qualidade de publicação não citada, Colaboração Internacional e Nacional, dentre outros, aplicáveis a grupos. A distribuição geográfica e a colaboração internacional de publicações científicas latino-americanas e caribenhas sobre tuberculose no PubMed foi estudada por Torres-Pascual, Sánchez-Pérez E Àvila-Castells (2021). Este estudo apresentou medidas condensadas por país, como número de artigos com colaboração e de citações, dentre outras.

Uma análise bibliométrica com foco em coautorias nas publicações brasileiras sobre medicina regenerativa foi realizada por Acero e Klein (2021). Neste estudo, o total de publicações com autores brasileiros, a percentagem de publicações com autores brasileiros e o número de citações desses artigos foi demonstrado por país.

[12] https://dialnet.unirioja.es/metricas/

O estudo evidenciou o crescimento significativo de coautorias com autores brasileiros na última década.

Quadro 8 – Síntese dos estudos de caso em países e instituições recuperados na RSL

Artigo	**Indicadores do grupo**	**Grupo analisado**	**Genérico ou análise de um conjunto específico**
(ALEIXANDRE *et al*., 2013)	Total de artigos, total de colaborações	País, universidade	Específico: Viticultura
(CODINA-CANET, 2012)	Total de artigos, total de colaborações	Universidade	Específico: Universidade Politécnica de Valencia
(ABRAMO; D'ANGELO, 2015)	*Fractional Scientific Strength* (FSS) e FSS^N	Províncias	Específico: províncias italianas
(ROSAS, 2015)	Total de citações	Universidades	Específico: UFV, UNESP e USP
(SIMON, 2016)	Total de Colaborações, entrada de novos autores	Conferência	Específico: Koli Calling
(GOLICHENKO; MALKOVA, 2017)	Importação e exportação de conhecimento	País	Específico: Rússia
(GONZALEZ-ALCAIDE *et al*., 2017)	Total de artigos, total de colaborações	País	Genérico
(FELIPE DIAZ-CARDENAS *et al*., 2017)	Total de artigos, total de citações	Estado	Específico: Puebla (México)
(YUAN *et al*., 2018)	Total de colaborações	País	Específico: China
(PILCEVIC; JEREMIC; VUJOSEVIC, 2018)	$JIF_{PERCENTILE}$ total de artigos e *ranking* descendente	Universidade	Específico: setores da Universidade de Belgrado
(MESCHINI, 2018)	Total de colaborações	País	Específico: Brasil

(SAHOO *et al*., 2019)	Total de artigos, total de citações	País, Universidade e Instituto de Pesquisa	Específico: área de Ciência da Informação
(CAKIR *et al*., 2019)	Posição em *Rankings*	Universidade	Específico: CERN para Universidades Turcas
(SILVA, 2019)	Redes sociais baseadas em total de artigos	Evento	Específico: ENANCIB
(LANCHO-BARRANTES; CANTU-ORTIZ, 2020),	Colaboração nacional, internacional e institucional	País	Específico: México
(CICERO; MALGARINI, 2020)	*Field Weighted Citation Impact*	País	Genérico por país
(THOMPSON *et al*., 2020),	Total de artigos, total de institutos de pesquisa, colaboração	País	Específico: área de Ciência da Informação
(KUMAR *et al*., 2021)	Total de artigos, total de citações	País e instituição	Específico: indicadores institucionais
(MORENO-DELGADO; GORRAIZ; REPISO, 2021)	Fator de Impacto de Garfield para países	País	Genérico por país
(GREGORIO-CHAVIANO *et al*., 2021)	Impacto da pesquisa	Áreas e Universidades	Genérico para áreas e universidades
(PAKKAN *et al*., 2021)	Total de artigos, total de citações	Universidades	Específico: Índia
(TORRES-PASCUAL; SÁNCHEZ-PÉREZ; ÁVILA-CASTELLS, 2021)	Total de artigos, total de citações	País	Específico: Publicações científicas latino-americanas e caribenhas sobre tuberculose no PubMed
(ACERO; KLEIN, 2021)	Total de artigos	País	Específico: Brasil na área de medicina regenerativa

Fonte: Elaborado pela Autora

Todos os trabalhos apresentados nessa seção se encaixam nos critérios de busca como os que aplicam métricas existentes para avaliar a produtividade de grupos de pesquisadores - Quadro 8. Número de artigos e citações foram as métricas mais utilizadas nos artigos recuperados. Também foi possível identificar abordagens que aplicam métricas comuns a outros contextos (como revistas ou áreas de pesquisa) para avaliar grupos de pesquisadores. O Fator de Impacto de Garfield e JIF, medidas comumente associadas a jornais e revistas de publicação científica, foram utilizadas para medir a produção de universidades e países. Outra medida aplicada a grupos é a *Field Weighted Citation Impact*, utilizada para avaliar o impacto das citações para uma determinada área de pesquisa, aplicada no contexto de grupo de autores a países inteiros.

É possível notar uma massiva utilização de indicadores totalizadores: total de artigos, total de citações e total de colaborações. Ao utilizar o total de artigos de uma instituição nem sempre fica claro se esta medida foi coletada através do campo afiliação ou se foi gerada a partir de uma lista atualizada dos autores. Portanto, mesmo no contexto de países e instituições, não há uma metodologia estabelecida para compor métricas para produtividade desses grupos, o que dificulta a comparação entre artigos que avaliam países e instituições diferentes.

4.2 ANÁLISE EM REDES DE COAUTORIA

Além das abordagens que consideram números e percentuais de artigos, citações e autores, há outro ramo bem comum relativo à análise de grupos de coautores: as redes de coautoria. Em 2012 foi proposta uma abordagem híbrida para analisar a similaridade de conteúdo e similaridade de redes de coautoria em diferentes domínios (YANG; TANG, 2012). Os autores definiram que uma alta similaridade entre redes de coautoria de diferentes domínios é um resultado de grupos colaborativos que participam de trabalhos científicos em ambos os domínios. No que concerne a similaridade de conteúdo, esta indica o quanto dois domínios são semelhantes com base nos títulos e/ou resumos de suas publicações. Os autores propuseram as

medidas de similaridade utilizando um vetor de termos TF-IDF[13] para comparar duas redes de autoria.

Vanz (2013) mapeou redes colaborativas em estudos métricos de Ciência e Tecnologia. Neste trabalho foram apresentadas algumas medidas clássicas de redes de coautoria, como densidade e medidas de centralidade. A autora descreveu pesquisas empíricas aplicadas em redes de coautoria e suas descobertas, como a propriedade de conexão preferencial, o nível de agrupamento e o modelo sem escala. Alves (2014) propôs uma rede de coautoria institucional para área da Ciência da Informação com foco em parcerias interinstitucionais em Ciência da Informação que se formam por meio da rede de coautoria. Uma comparação entre indicadores de rede de coautoria (indicadores de centralidade de grau, centralidade de intermediação e de proximidade da rede institucional) foi apresentada, além dos conceitos CAPES atribuídos a cada instituição. O autor identificou a existência de correlação positiva entre os indicadores de rede e os conceitos CAPES.

Em 2016, os programas de estímulo governamental da China e da Coreia do Sul foram analisados a partir da avaliação bibliométrica extraída de redes de coautoria normalizadas (PARK; YOON; LEYDESDORFF, 2016). As análises foram realizadas a partir de dados extraídos da *Web Of Science* utilizando grafos com publicações em coautoria entre China e Coreia do Sul. Foram utilizados indicadores baseados em grafos (média, dependência etc.), com a aplicação de métodos de contagem inteira e fracionada. Considerou-se diferentes métodos de contagem de artigos, como os trabalhos com coautoria e aqueles que envolvem diferentes institutos e países. Os autores afirmam que o aumento da colaboração internacional pode levar a um menor número de publicação, utilizando contagem fracionada. Concluem que a contagem padrão (não fracionada) não é apropriada para a avaliação do estímulo de colaboração.

O desempenho científico de artigos de educação física na China foi analisado também em 2021 (ZHANG *et al.*, 2021a). A pesquisa tomou como base a perspectiva da análise de redes sociais. Foi analisada a influência da centralidade do grau de autor, bem como o *índice L*, proposto no artigo. O *índice L* corresponde ao número da rede que possui ao menos L autores (nós) cujos graus são pelo menos L nós

[13] *Term Frequency – Inverse Document Frequency*, podendo ser traduzido para Frequência do Termo – Frequência Inversa do Documento. É uma medida que indica a importância de um termo em um documento em relação a uma coleção de documentos.

conectados a um nó da rede. Ou seja, atribui um número ao conjunto de coautores, desde que estejam conectados à rede. Os autores compararam o *índice L* com a centralidade do grau e desempenho científico. O estudo evidencia que a centralidade do grau está positivamente correlacionada com o número de artigos, média de citações por artigo e *índice h*, destaca também que o *índice L* está positivamente correlacionado com o número de artigos, média de citações por artigo e *índice h*.

Também em 2021, efetuou-se uma análise de redes de colaboração científica em três conferências de educação em computação (SIGCSE *Technical Symposium*, ITiCSE e ICER) (ZHANG *et al*., 2021b). Este trabalho apresenta dados de total e percentagem de artigos sobre a temática por país. Foi analisada a localização geográfica dos autores e modelada uma rede de colaboração científica para cada conferência. O estudo evidenciou que a comunidade está aberta a novos autores e que o número de autores e nível de colaboração estão crescendo.

Quadro 9 - Análise em redes de coautoria recuperadas na RSL

Artigo	**Indicadores do grupo**	**Grupo analisado**	**Genérico ou análise de um conjunto específico**
(YANG; TANG, 2012)	Similaridade de redes de coautoria em diferentes domínios	Áreas	Genérico por diferentes áreas
(VANZ, 2013)	Densidade e medidas de centralidade	Institucional	Específico: estudos métricos de Ciência e Tecnologia
(ALVES, 2014)	Centralidade de grau, centralidade de intermediação e de proximidade da rede institucional	Institucional	Específico: coautoria institucional em Ciência da Informação
(PARK; YOON; LEYDESDORFF, 2016)	Indicadores baseados em grafos (média, dependência etc.). Métodos de contagem inteira e fracionada	País	Específico: China e Coréia
(ZHANG *et al*., 2021a)	*Índice L*	País e área	Específico: Educação física na China
(ZHANG *et al*., 2021b)	Total e percentagem de artigos sobre a temática por país	País	Específico: conferências de educação em computação (SIGCSE *Technical Symposium*, ITiCSE e ICER)

Fonte: Elaborado pela Autora

Uma síntese dos trabalhos recuperados na RSL que utilizam Análise de redes de coautoria é apresentada no Quadro 9. A análise de redes de coautoria é muito útil para analisar grupos de coautores e é um recurso muito utilizado. Análises visuais, como redes de relacionamento, são muito úteis para um universo e análise pontuais. Quando o objetivo é analisar grandes volumes de dados as redes podem se tornar muito complexas e uma tendência pode não ficar muito visível. Por isso é importante apresentar indicadores numéricos e após análise desses indicadores podemos utilizar as redes e outras formas visuais de análise.

4.3 INDICADORES OU METODOLOGIAS PARA AVALIAR GRUPOS DE PESQUISADORES

Bornmann *et al.* (2016a) propuseram um aplicativo baseado na regressão logística multinível bayesiana (BMLR) para a identificação de instituições que colaboram com sucesso. Neste trabalho, foi realizada uma análise de como uma instituição de referência colaborou com o sucesso e com quais outras instituições ela foi mais bem-sucedida. A taxa de *best paper* foi usada como indicador para avaliar o sucesso da colaboração de uma instituição. Essa taxa considera a proporção de artigos que pertencem aos 10% dos artigos mais citados nas áreas temáticas correspondentes e anos de publicação. Apesar de realizar uma análise mais visual, propõe uma taxa de colaboração entre instituições, mas não avalia instituições isoladamente.

Uma categorização de argumentos para métodos de contagem para indicadores de publicação e citação foi proposta por Gauffriau (GAUFFRIAU, 2017). Neste trabalho o problema na contagem de artigos e citações é explorado, sendo abordado o problema de calcular indicadores em artigos com múltiplos autores. A autora identificou os argumentos explícitos no texto para utilizar uma ou outra abordagem, identificando 4 grupos distintos de argumentos: i) os que são relacionados ao que um indicador mede; ii) os que abordam a aditividade de um método de contagem; iii) os que expõem razões pragmáticas para a escolha do método de contagem e iv) os que tratam da influência de um indicador na comunidade de pesquisa ou de como este é percebido pelos pesquisadores.

Em 2018, um estudo apontou relações estatísticas entre autor correspondente, coautoria internacional e impacto da citação nos sistemas nacionais

de pesquisa (DE MOYA-ANEGON *et al*., 2018). Neste trabalho os indicadores foram avaliados em diferentes países e instituições. Dentre os indicadores destacam-se: a) Liderança de pesquisa com base em autor correspondente; b) Colaboração internacional usando dados de coautoria internacional e c) Impacto da citação normalizada por campo. Os autores destacam a complexidade da relação entre autoria e indicadores baseados em citações, por refletir a fase de desenvolvimento científico de um país. Indicam ainda que deve haver distinção do efeito de um efeito de liderança real de um puramente estatístico considerando a contagem fracionada.

A colaboração científica a partir dos indicadores relacionais de coautoria foi abordada por Grácio (2018). Neste trabalho foram examinados os aspectos teóricos e conceituais acerca do método bibliométrico relacional. A autora utiliza um índice de colaboração IC, baseado no número de artigos em colaboração, que pode ser utilizado de forma direta para avaliar grupos de autores. Uma extensão do *índice h* específica para analisar coautoria é proposta no artigo “Bibliometria para trabalhos de colaboração” (tradução nossa) (ROSSI; STRUMIA; TORRE, 2019). Neste trabalho é realizada uma análise do número de autores em cada publicação propondo um *índice h* que não varia de acordo com o número de coautores. São apresentados argumentos teóricos e numéricos a favor da ponderação das contribuições individuais.

Uma nova abordagem de estudos bibliométricos considera gerações na composição de campos de pesquisa (PFRIEGER, 2021). Estima o impacto com base no registro de publicação, laços genealógicos e conexões colaborativas, usando nomes de autores e anos de publicação de artigos científicos relacionados a uma área de interesse. O trabalho não visa analisar grupos, mas apresenta indicadores que podem ser aplicados a grupos, como, por exemplo o CC (contagem de conexões colaborativas; calculada como a soma do número de coautores com o número de autores que listaram o autor em questão como coautor).

Uma estrutura de avaliação de produtividade orientada por dados para equipes de pesquisa colaborativa foi proposta por Wang *et al.* (2021). Descrever a produtividade para equipes colaborativas grandes e complexas que possuem muitos indicadores relacionados à produtividade é uma tarefa complexa, sendo foco deste trabalho analisar equipes de pesquisa de grande escala. Os autores propõem uma estrutura orientada a dados para processar, analisar e entender os indicadores relacionados à produtividade, composta por três etapas: 1) Descrição das equipes colaborativas com extração de indicadores; 2) Identificação dos principais indicadores

de produtividade; 3) Previsão de produtividade para verificar a utilidade dos indicadores.

Quadro 10 - Indicadores ou metodologias para avaliar grupos de pesquisadores

Artigo	**Indicadores do grupo**	**Grupo analisado**	**Genérico ou análise de um conjunto específico**
(BORNMANN *et al*., 2016)	Regressão logística multinível bayesiana (BMLR) para a identificação de instituições que colaboram com sucesso	Instituições	Específica: instituição de referência
(GAUFFRIAU, 2017)	Contagem de artigos e citações em artigos com múltiplos autores	Diverso	Diverso
(DE MOYA-ANEGON *et al*., 2018)	Liderança de pesquisa com base em autor correspondente; Colaboração internacional usando dados de coautoria internacional e Impacto da citação normalizada por campo	Países e Instituições	Genérico
(GRÁCIO, 2018)	Índice de colaboração IC	Grupos de autores	Genérico
(ROSSI; STRUMIA; TORRE, 2019)	Coautoria	Grupos de autores	Genérico
(PFRIEGER, 2021)	Gerações na composição de campos de pesquisa	Não visa analisar grupos, mas apresenta indicadores que podem ser aplicados a grupos	Genérico
(WANG *et al.,* 2021)	Produtividade do grupo	Grupos de autores	Genérico
(FASSIN, 2021)	Metodologia f^2	Universidades	Específico: Universidades na China

Fonte: Elaborado pela Autora

Um trabalho propôs estudar o progresso da China na pesquisa em gestão acadêmica internacional, com base na metodologia f^2. A metodologia f^2 baseia-se em uma classificação mais refinada das publicações em categorias de citações, com foco

no percentil 10% e no h-core (FASSIN, 2021a). O estudo analisou se o progresso das universidades chinesas na pesquisa em gestão acadêmica está em consonância com o progresso geral da ciência na China.

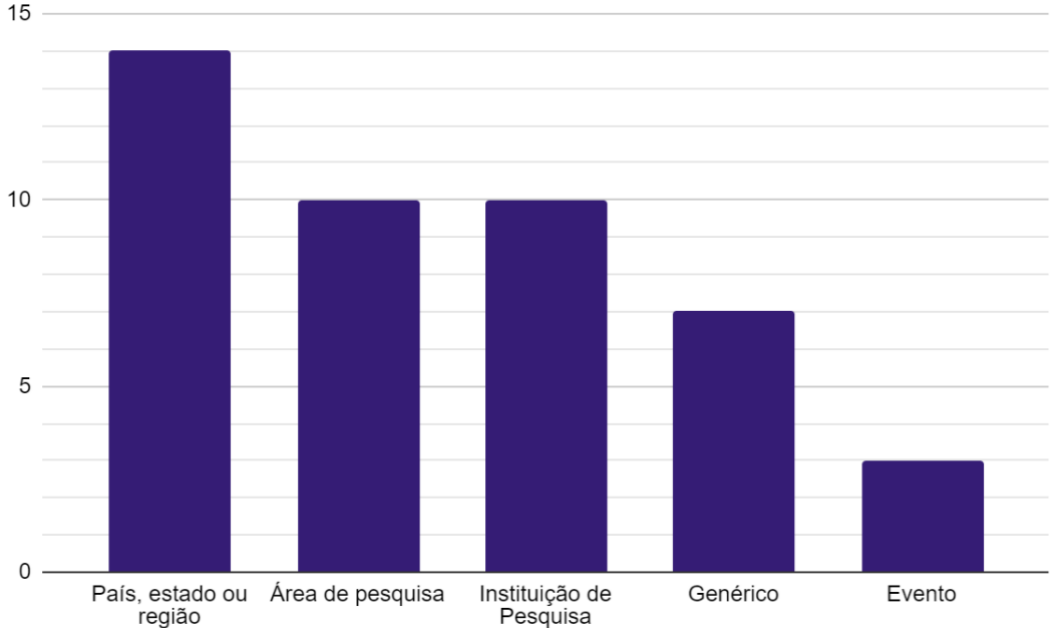
Os estudos apresentados nessa seção abordam de forma geral a avaliação de grupos de pesquisadores e estão sumarizados no Quadro 10. Alguns analisam características específicas na metodologia como a variação do *índice h* quando ocorrem coautorias (GAUFFRIAU, 2017; ROSSI; STRUMIA; TORRE, 2019). Outros trabalhos propõem indicadores ou metodologias para avaliação de grupos (BORNMANN et al., 2016b; FASSIN, 2021b; GRÁCIO, 2018; ZHANG et al., 2021a). Cabe destacar que os trabalhos que abordam esta problemática são bastantes recentes.

4.4 SÍNTESE DOS RESULTADOS DA RSL

Ao retomarmos as perguntas de pesquisa, temos: *RQ1: Como tem sido realizada a avaliação da produtividade dos grupos de pesquisadores nos últimos anos?* Com base nos trabalhos selecionados, foi possível identificar que a avaliação da produtividade dos grupos de pesquisadores ocorre de formas variadas, por exemplo, em agrupamentos por países e instituições (Figura 18). Em sua grande maioria, os trabalhos focam em representar e analisar casos específicos, seja em países, estados, regiões ou universidades e demais instituições de pesquisa. Existem ainda trabalhos que focam em Áreas e eventos de pesquisa específicos ao avaliar um país ou instituição.

São poucos os trabalhos (7) que propõem uma forma de avaliação ou indicador Genérico (Figura 18). Estes trabalhos compreendem, em maioria, propostas de indicadores para avaliação de grupos, sejam esses grupos países ou instituições, mas sem aplicar em um caso específico (DE MOYA-ANEGON et al., 2018; GONZALEZ-ALCAIDE et al., 2017; GRÁCIO, 2018; PFRIEGER, 2021; ROSSI; STRUMIA; TORRE, 2019). Destaca-se também um estudo qualitativo sobre os argumentos utilizados para utilizar uma outra forma de agregação (GAUFFRIAU, 2017). Existe também a proposta de um *framework* baseado em dados preexistentes para identificar os indicadores que melhor representam a produção de um grupo (WANG et al., 2021).

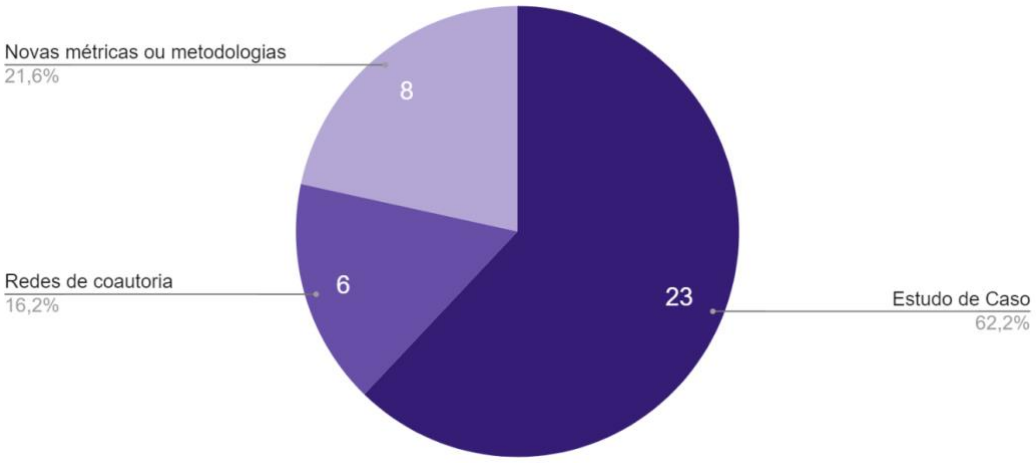
Figura 18- Total de Trabalhos por grupo analisado



Fonte: Elaborado pela Autora

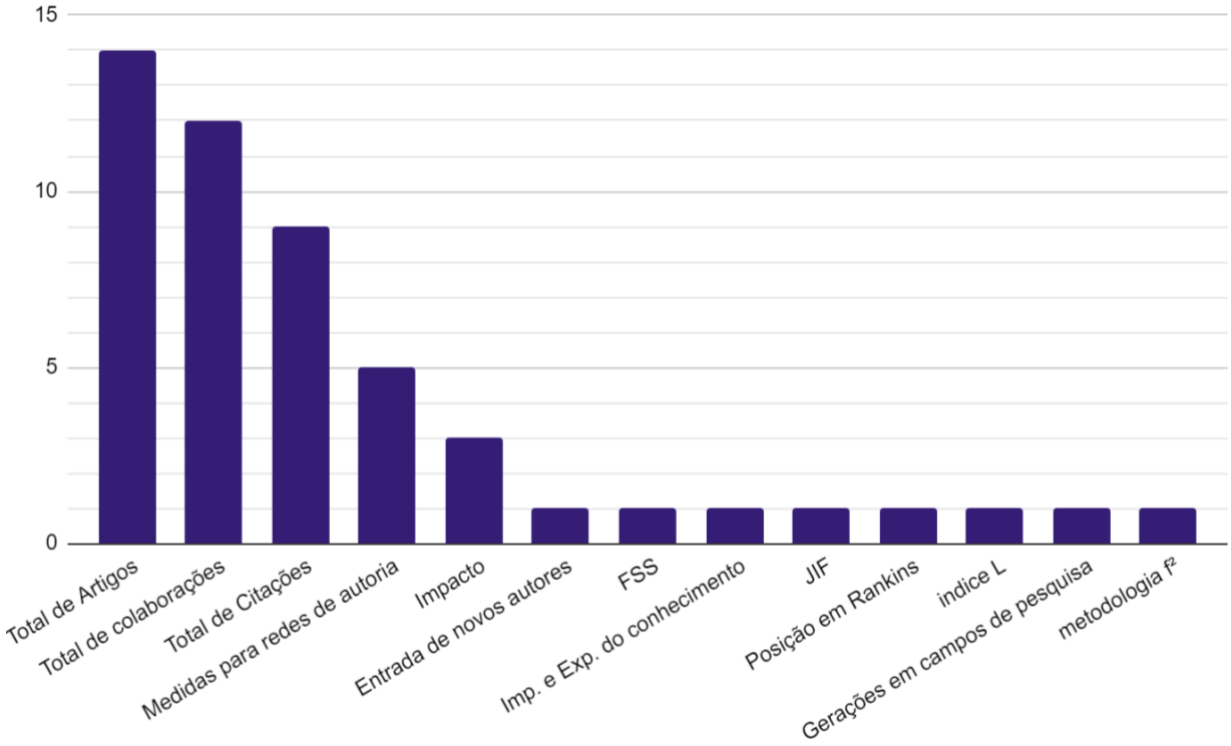
Ainda sobre a forma como as pesquisas sobre avaliação de produtividade em grupos têm sido realizadas é trazida uma sumarização dos trabalhos encontrados (Figura 19). Nesta figura há uma sumarização das tabelas 3, 4 e 5, que trazem os detalhes de cada publicação. Os estudos de caso compreendem a maioria do grupo analisado, com 62,2%. As análises realizadas em redes de coautoria correspondem a 16,2% dos estudos recuperados. Já as propostas de novas métricas e metodologias para avaliar grupos de pesquisadores correspondem somente a 21,6% dos trabalhos.

Figura 19 - Proporção das categorias de trabalhos recuperados na Revisão Sistemática da Literatura



Fonte: Elaborado pela Autora

Figura 20 - Indicadores identificados na Revisão Sistemática da Literatura



Fonte: Elaborado pela Autora

Acerca da segunda questão de pesquisa: *RQ2: Quais indicadores de produtividade científica têm sido utilizados para avaliar esses grupos de pesquisadores?* O resultado é apresentado na Figura 20, onde se verifica que os totais de artigos, citações e colaborações são os indicadores mais frequentemente utilizados para avaliar grupos de pesquisadores. O impacto, comumente relacionado ao número de citações, também é utilizado de forma recorrente. Cabe destacar que em um mesmo artigo mais de um indicador pode ser empregado, desta forma, cada indicador em cada artigo foi contabilizado para composição do gráfico ilustrado na Figura 20. entre os trabalhos recuperados, existem também propostas de outros indicadores.

Sobre a terceira questão *RQ3: Existe uma metodologia consolidada para a aplicação de indicadores de produtividade em grupos de pesquisadores?* Com base nos artigos analisados, é possível verificar que existem diversos índices diferentes (Figura 20) e que há uma certa consolidação acerca da utilização de indicadores totalizadores: artigos, colaborações e citações. Foi identificado ainda um *framework* proposto por Wang (2021) que define em linhas gerais os passos para identificar os indicadores que melhor representem a produtividade de um grupo de pesquisa, mas

não especifica como deve acontecer a avaliação nem os critérios necessários para avaliar o grupo.

Considerando o levantamento realizado podemos afirmar, portanto, que não há uma forma consolidada de se avaliar grupos de pesquisadores. A maioria dos trabalhos são estudos de caso, existindo poucos trabalhos que propõem indicadores ou outras formas de avaliação de maneira genérica, sendo bastante aplicado o uso de contadores. O próximo capítulo apresenta a proposta do presente trabalho: um *framework* para análise de grupos de pesquisadores.

5 PROPOSTA

Conforme apresentado anteriormente, existem alguns índices para avaliar um grupo de pesquisadores, com ênfase no *índice* h_1 e no *índice* h_2 (MITRA, 2006), índices h sucessivos (SCHUBERT, 2007), CP-*index* (ALTMANN; ABBASI; HWANG, 2009), e a razão h_2/h_1 – (ROUSSEAU; YANG; YUE, 2010). Todos esses índices dependem de uma entrada definida como um conjunto de artigos científicos. No entanto, a forma como os artigos são contabilizados em cada estratégia pode variar: usar os artigos para calcular cada índice do pesquisador separadamente ou organizar um grande conjunto de artigos para calcular um índice global.

Neste capítulo apresentamos o *IN-GROUP* uma abordagem desenvolvida no escopo desta tese para contabilizar os artigos na aplicação do índice h, útil na avaliação de grupos de pesquisadores *(*SANTOS; DUTRA, 2022). O *framework F-GROUP*, tema central desta tese, é apresentado na seção 5.2, seguido de cenários de aplicação na seção 5.3, evidenciando e exemplificando os passos necessários para avaliar a produção científica em grupos de pesquisadores.

5.1 FORMAS DE AGREGAÇÃO NA AVALIAÇÃO DE GRUPOS: APLICAÇÃO DO ÍNDICE H

Em um primeiro momento o problema da agregação e seleção de artigos será ilustrado usando como exemplo o índice h. De forma resumida, os trabalhos que aplicaram o *índice h* para grupos de pesquisadores utilizam três abordagens principais:

- *Todos*: dado um conjunto de artigos representando todos os membros do grupo, esta abordagem usa a mesma ideia do *índice h*, ou seja, número de artigos mais citados com um total de citações igual ou superior ao número de artigos (VAN RAAN, 2006);
- *Sucessiva*: dado o *índice h* para cada pesquisador de um determinado grupo (índice h individual), esta abordagem considera o número de principais autores com um í*ndice h* maior ou igual ao tamanho do conjunto de autores principais. (SCHUBERT, 2007);
- *Média ou mediana*: dado o *índice h* para cada pesquisador de um determinado grupo (índice h individual), esta abordagem calcula a média

ou mediana do *índice h* para esse grupo (KHAN et al., 2013; MUGNAINI; PACKER; MENEGHINI, 2008; RAD et al., 2010).

As abordagens *Sucessiva* e *Média* do *índice h* são aplicadas com base no *índice h* previamente calculado. Na abordagem *Sucessiva* nem todos os colaboradores são representados, uma vez que apenas alguns pesquisadores com os maiores valores de índice h são considerados. Por outro lado, a abordagem *Média* pode ser influenciada por apenas um grande pesquisador, especialmente em grupos menores, desta forma pode-se utilizar a mediana para atenuar este problema. Além disso, para impulsionar o *índice h* global, um departamento ou instituição pode contratar pesquisadores com índices h de valor já elevado. Desta forma, não estaria representado o trabalho de cada pesquisador dentro do grupo avaliado, mas, sobretudo, os trabalhos anteriores de alguns dos pesquisadores dentro do grupo.

Outro problema ao calcular o índice global com base no índice de cada pesquisador individualmente é que grupos que costumam publicar juntos podem melhorar seu índice de forma uniforme. Por exemplo, dez artigos publicados pelo mesmo grupo de autores com dez citações cada um pode impulsionar significativamente um índice global, com base apenas nesses dez artigos. O mesmo conjunto de artigos é, então, contabilizado várias vezes ao calcular o *índice h* para cada pesquisador, e é contado muitas vezes, beneficiando grupos que costumam publicar artigos com vários pesquisadores em colaboração.

Ao avaliar um grupo de pesquisadores, sendo este um departamento inteiro ou um grupo menor, é fundamental apresentar indicadores que representem o trabalho e esforço daquele grupo em conjunto. O cálculo deste número deve considerar o trabalho do grupo como um esforço de equipe e valorizar as contribuições de todos os pesquisadores.

5.1.1 *IN-GROUP*: Uma forma mais equânime de selecionar trabalhos ao aplicar variações do índice H na avaliação de grupos de pesquisadores

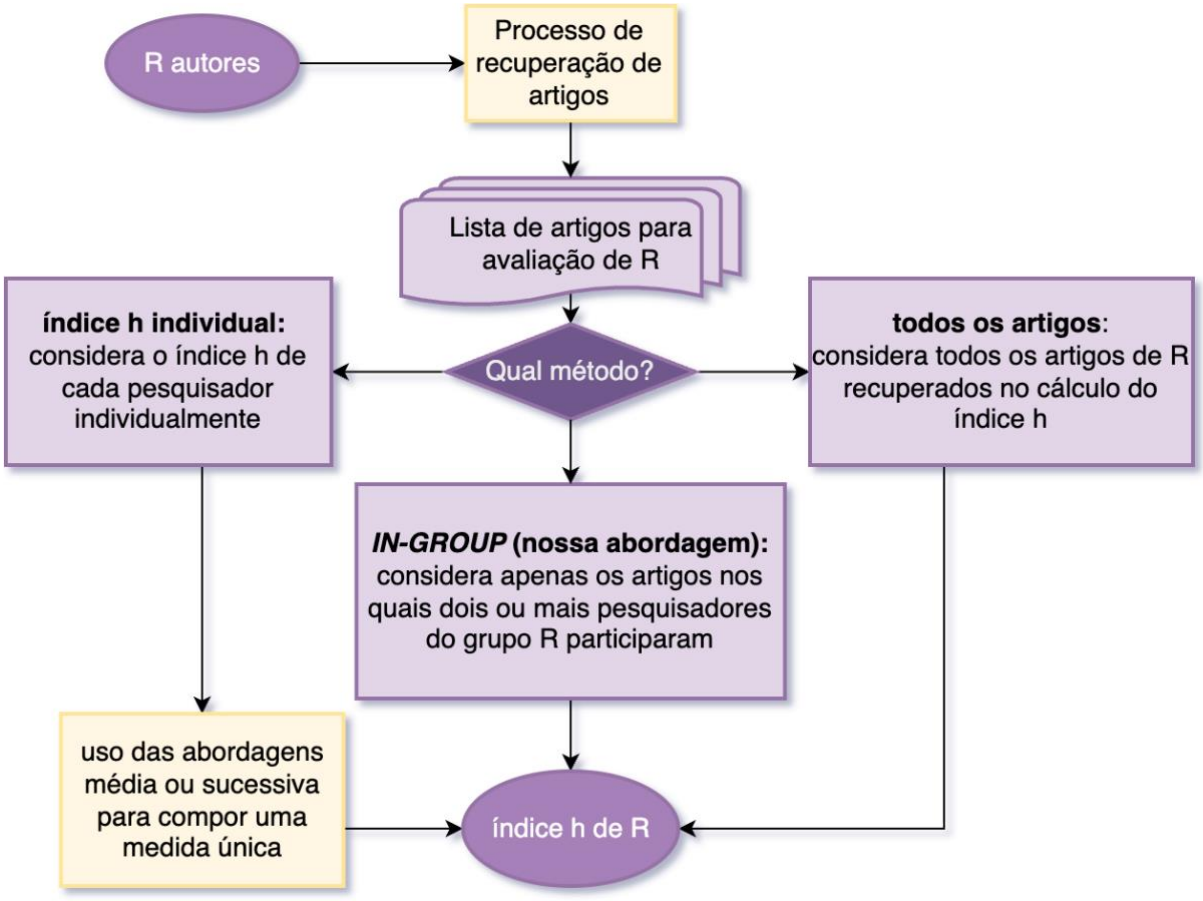
O primeiro aspecto a ser abordado quando o objetivo é avaliar um grupo de pesquisadores é definir o que é um grupo de pesquisadores. O número de pesquisadores em cada grupo é relevante? Deve-se considerar a proximidade dos pesquisadores em seus campos de trabalho? É possível comparar pesquisadores já estabelecidos com estudantes iniciantes na carreira acadêmica? Na abordagem *IN-*

GROUP, a definição de um grupo de pesquisadores é baseada na proposta por Wang (2021): duas ou mais pessoas trabalhando em conjunto ao compartilhar e comunicar suas produções coletivas.

A Figura 21 apresenta um fluxograma de um processo convencional para avaliação de grupos de pesquisadores, enriquecido com nossa abordagem *IN-GROUP*. A entrada é uma lista de autores representada pela letra R.

O primeiro passo é recuperar os artigos de autoria dos pesquisadores da lista R. Uma vez que esses artigos são recuperados, para calcular o *índice h*, é possível combinar seus dados de diferentes maneiras: i) considerando o *índice h* individualmente; ii) considerando todo o conjunto de artigos recuperados; ou iii) considerando apenas uma parte deste conjunto de artigos (nossa abordagem). É possível obter um único valor do *índice h* representando os pesquisadores da lista R através de todos esses métodos.

Figura 21 - Cálculo do índice h via seleção e agregação de artigos.



Fonte: Adaptado e traduzido de Santos e Dutra (2022).

Através do fluxograma representado na Figura 21 é possível identificar os principais aspectos na avaliação de grupos: a) A listagem de autores, representada por R; b) O processo de recuperação dos artigos desses autores; c) método de agregação/composição das medidas para cálculo do indicador. A abordagem *IN-GROUP* considera apenas os artigos nos quais dois ou mais pesquisadores do grupo R participam. A participação dos membros do grupo pode variar e ter diferentes graus. Dependendo do objetivo da avaliação, pode haver a necessidade de uma integração com o grupo ainda maior. Uma breve explicação sobre grau de colaboração interno na abordagem *IN-GROUP* é apresentada a seguir.

5.1.2 Grau de Colaboração Interno no *IN-GROUP*

Um dos primeiros problemas ao se trabalhar com a produtividade de grupos está relacionada a quais artigos do grupo devem ser selecionados. A abordagem *IN-GROUP* propõe que somente os artigos produzidos pelos membros de um grupo de forma conjunta devem ser considerados. Conforme descrito na Figura 21, no mínimo dois integrantes são necessários, mas nem sempre o valor mínimo de autores em cada artigo será igual a 2. Para alguns casos um valor maior de grau de colaboração é necessário. O Grau de Colaboração Interno na abordagem *IN-GROUP* é o número de autores mínimo para incluir um artigo na avaliação do grupo.

Dado o cenário de comparar grupos de pesquisadores com diferentes números de integrantes (tais como departamentos, países etc.), pode haver uma tendência em se atribuir valores maiores a grupos maiores. Uma equipe com muitos pesquisadores sempre teria uma vantagem utilizando o *IN-GROUP* em especial com valores de Grau de Colaboração mais baixos. Para equilibrar este valor, um grau de colaboração interno maior deve ser utilizado.

5.1.3 Cenário comparativo entre abordagens de agregação

Para facilitar a compreensão das diferentes formas de cálculo do índice h para grupos de pesquisadores, será apresentado o exemplo de um cenário fictício. Neste exemplo, vamos considerar um grupo de pesquisa com sete autores.

Para ilustrar como apenas adicionar ou remover um autor impacta no resultado em termos de seleção e agregação de artigos, consideramos dois

subconjuntos de autores, ilustrado no Quadro 11. O grupo R2 compreende todos os pesquisadores do grupo R1 mais um pesquisador, denominado X, com *índice h* no valor de 30.

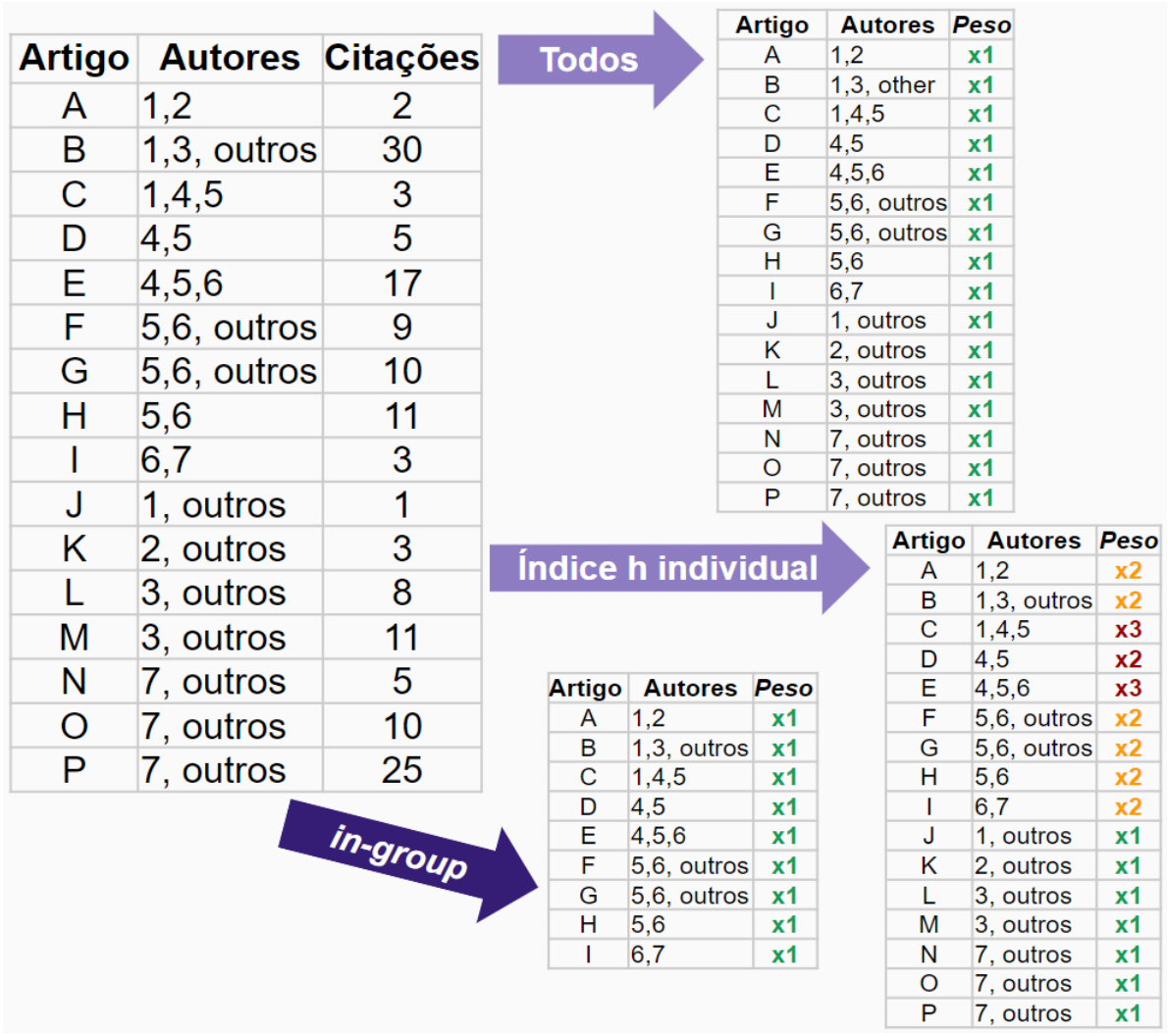
Considerando o grupo R1 = [1, 2, 3, 4, 5, 6, 7], é apresentado um conjunto de artigos ilustrados na tabela à esquerda da Figura 22. Neste exemplo, a tabela apresentada contém todos os trabalhos pré-selecionados, com restrições como intervalo de tempo e campo de pesquisa, entre outros, necessários de acordo com o objetivo da análise.

Quadro 11 - Grupos R1 e R2

Grupo	Lista de Autores
R1	1, 2, 3, 4, 5, 6, 7
R2	1, 2, 3, 4, 5, 6, 7, *X*

Fonte: Elaborado pela autora

Figura 22 - Opções de cálculo de índice h para o Grupo R1



Fonte: Adaptado e traduzido de Santos e Dutra (2022).

A Figura 22 mostra como seria a seleção dos artigos com base em trabalhos anteriores. Na abordagem *todos*, todo os artigos de um determinado grupo são considerados no cômputo do *índice h*. Nesse caso, cada artigo tem peso *x1*, pois é considerado apenas uma vez. A abordagem "*Índice h individual*" representa as abordagens sucessiva e média, onde os índices h são calculados individualmente para cada pesquisador. Em uma segunda etapa, um índice global baseado no *índice h* anterior é calculado, fazendo com que um mesmo artigo possa ser considerado mais de uma vez.

Com base na Figura 22, pode-se destacar duas questões fundamentais: a primeira é que o uso de artigos escritos por qualquer membro do grupo não representa o trabalho de todo o grupo, mas sim a produção de cada membro separadamente. Outra questão diz respeito às abordagens que primeiro calculam o *índice h* para cada pesquisador separadamente, potencializando a avaliação global do grupo ao contabilizar os mesmos artigos mais de uma vez. No exemplo, o artigo C é usado para calcular o *índice h* para pesquisadores 1, 4 e 5. Consequentemente, o artigo C poderia aumentar significativamente este índice no grupo de pesquisa avaliado.

A abordagem *IN-GROUP* foi proposta para superar esses problemas. Nesta abordagem, um artigo deve ter dois ou mais autores do mesmo grupo de pesquisadores para ser selecionado. Neste exemplo foi utilizado um Grau de Colaboração Interno = 2. Além disso, o índice do grupo deve ser calculado com base em uma avaliação dos artigos produzidos pelo grupo, não no *índice h* de cada pesquisador individualmente. Assim, selecionamos os principais artigos que representam a contribuição de um grupo como um esforço de equipe. Esta forma de seleção supera o problema de se contar o mesmo artigo mais de uma vez e normaliza a situação relacionada à disparidade de produção dos pesquisadores de um mesmo grupo. Esta seleção também é útil para coletar outras métricas, como contagem de artigos e citações. Também pode ser usada como entrada para o Indicador Crown, por exemplo.

O índice h de cada autor foi calculado de acordo com o total de citações apresentado na Figura 22, e baseado apenas na lista de trabalhos presentes na mesma figura. O valor do índice h para cada autor do exemplo é apresentado na Tabela 4.

Tabela 4 - Índice h de cada autor no exemplo

Autores	índice h do autor
1	2
2	2
3	3
4	3
5	6
6	4
7	4
X	30

Fonte: Elaborado pela autora

Os índices h para os grupos R1 e R2 calculados em cada uma das diferentes abordagens são apresentados no Quadro 12. Os resultados obtidos variam consideravelmente ao adicionar apenas um pesquisador ao grupo R1.

Quadro 12 - Índices h R1 e R2 calculados por diferentes abordagens.

Abordagem	Descrição	*índice h* de R1	*índice h* de R2
Todos os artigos	Dados todos os artigos nos quais qualquer pesquisador do grupo é autor: maior número natural de *n* artigos com um número de citações igual a *n* cada.	8	30
Sucessiva	Dado o *índice h* de cada membro do grupo: o maior número natural de *n* autores com um *índice h* igual a *n* cada.	3	4
Média	Dado o *índice h* de cada membro do grupo: a média do *índice h* do grupo.	3.43	6.25
IN-GROUP	**Dados todos os artigos nos quais dois ou mais pesquisadores do grupo são autores: maior número natural de *n* artigos com um número de citações maior ou igual a *n* cada.**	**5**	**5**

Fonte: Adaptado e traduzido de Santos e Dutra (2022).

Como R2 compreende todos os pesquisadores do grupo R1 mais um pesquisador X com um *índice h* de valor 30, os resultados na abordagem "*todos os artigos"* variam de 8 para 30. Nas abordagens *sucessiva* e *média* os valores também sofreram variação ao adicionar um novo pesquisador, na média o valor quase dobra ao adicionar apenas um pesquisador. A única abordagem que não varia neste exemplo é a abordagem *IN-GROUP*, pois apenas os trabalhos desenvolvidos em conjunto por membros do grupo são considerados. Somente quando o novo membro X começar a produzir em conjunto com o grupo que a avaliação considerará seus trabalhos.

Desta forma, as principais diferenças entre a abordagem aqui proposta e as demais abordagens são:

1) A abordagem *IN-GROUP* não teve variação do *índice h* nos grupos R1 e R2 do dado exemplo.
2) Ao invés de deixar o resultado pouco representativo, como na abordagem sucessiva, onde apenas os melhores pesquisadores são contabilizados, a abordagem *IN-GROUP* reconhece apenas um artigo cuja coautoria represente ao menos dois membros do grupo.
3) Ao invés de computar o mesmo artigo várias vezes com diferentes autores, como nas abordagens Sucessiva e Média, a abordagem *IN-GROUP* considera cada artigo apenas uma vez.

A abordagem *IN-GROUP* também pode ser aplicada a grupos maiores, como departamentos, instituições e até mesmo países. Esta nova abordagem de agregação de artigos para avaliação de grupos de pesquisadores é demonstrada como uma alternativa útil para selecionar e agrupar artigos, aplicando métricas baseadas no *índice h* para grupos de pesquisa para superar as limitações existentes (SANTOS; DUTRA, 2022).

Na próxima seção apresentamos o *F-GROUP*, um *framework* para aplicação de indicadores de produtividade científica em grupos de pesquisadores, com ênfase na colaboração.

5.2 *FRAMEWORK F-GROUP*

O problema da avaliação de grupos de pesquisadores envolve diversos fatores. Desde a definição do grupo a ser avaliado até os objetivos da avaliação

passando pelas formas de agregação, vários aspectos devem ser considerados. Existem diferentes formas de contabilizar e agregar artigos científicos: Todas (*all papers*) (VAN RAAN, 2006); *Sucessiva* (SCHUBERT, 2007); *Média ou Mediana*: (KHAN et al., 2013; MUGNAINI; PACKER; MENEGHINI, 2008; RAD et al., 2010); *IN-GROUP* (SANTOS; DUTRA, 2022). Deve-se ainda considerar aspectos relacionados à área de atuação, janela temporal e autocitação, dentre outros fatores.

Os indicadores para avaliação de um grupo são variados. Alguns indicadores foram propostos ao longo dos anos para avaliar os mais diversos grupos de pesquisadores, como países, instituições, departamentos, entre outros (ALTMANN; ABBASI; HWANG, 2009; GRACIO et al., 2020; KHAN et al., 2013; SCHUBERT, 2007). Dentre esses indicadores, os totalizadores são os mais utilizados: total de artigos, total de citações e total de colaborações (entre diferentes países e instituições).

No entanto, mesmo ao utilizar indicadores mais simples, como totais de artigos e citações, algumas questões sobre a seleção dos artigos ao se avaliar o grupo continuam em aberto. Dentre essas questões, podemos destacar as identificadas no decorrer deste trabalho:

a) Devem ser considerados todos os artigos de um autor que trabalhou em várias instituições, ou apenas os que foram publicados durante o período de trabalho em sua instituição atual?
b) Como balancear esses vínculos ao analisar a produção de um pesquisador quando o intuito é avaliar o trabalho de um grupo?
c) O número de autores em cada artigo deverá ser considerado?
d) Devemos considerar todas as citações em que um dos autores faz parte do grupo? Ou devemos nos concentrar apenas nos artigos em que todos os membros do grupo participam?

A combinação entre as perguntas anteriores com a falta de padronização e informação sobre a seleção dos artigos para a avaliação de grupos são fatores que motivaram a criação do *framework F-GROUP*, proposto neste trabalho.

Um *framework* é definido como uma “estrutura que possibilita o desenvolvimento de algo sobre sua base inicial” (MACEDO; SOUZA, 2023, p. 3). Os autores destacam ainda a natureza normativa da estrutura dos *frameworks*, tornando resultados de pesquisa mais rigorosos e permitindo a sua generalização. Portanto, o *framework F-GROUP*, aqui proposto, visa facilitar uma padronização na avaliação de

grupos de pesquisadores, visando permitir a futura comparação entre diferentes trabalhos que avaliam esses grupos.

O *framework F-GROUP* para avaliação de grupos de pesquisadores possui 3 fases: I) Identificação de Parâmetros Iniciais, II) Seleção de Artigos e III) Definição de Indicadores, Figura 23. A fase I) Parâmetros Iniciais foi dividida em três etapas: (1) Descrição do(s) Grupo, (2) Objetivos da Avaliação e (3) Fontes de Dados. Enquanto a fase II) Seleção de Artigos engloba as (4) Formas de contagem e agregação e os (5) Aspectos representativos, a fase III) Definição de Indicadores é composta pela etapa (6) Listagem de Indicadores.

São destacados no *framework* o Grau de Colaboração Interno e a abordagem de agregação *IN-GROUP*, apresentados por Santos e Dutra (2022). Para construção deste *framework*, levou-se em consideração todos os problemas relacionados à medição da produtividade para grupos de pesquisadores. Em especial, o problema relacionado à falta de comparabilidade entre diferentes trabalhos.

Figura 23 – Framework *F-GROUP*

Fonte: Elaborado pela autora

Critérios relacionados a quais artigos devem ser considerados ao se avaliar a produção de um grupo podem variar de acordo com os objetivos da avaliação. A afiliação, a data em que o artigo foi submetido ou publicado, a área, o número de

autores, a fonte dos dados, dentre outros, podem impactar significativamente nos resultados da avaliação. Enquanto algumas bases de dados de artigos científicos permitem a exportação com mais detalhes, incluindo informações como afiliação e país dos autores, outras são mais restritas, não fornecendo sequer aplicações para exportação dos dados. Esses detalhes da avaliação, quando explicitados de forma padronizada, permitem uma melhor comparação entre diferentes trabalhos, e consequentemente, entre diferentes grupos de autores. Em seguida, cada fase do *framework F-GROUP* é descrita com mais detalhes.

5.2.1 I) Identificação de Parâmetros iniciais

A fase de Identificação de Parâmetros Iniciais compreende os aspectos iniciais da avaliação que será realizada. É nesta fase que o grupo é caracterizado, os objetivos da avaliação são definidos e as fontes de dados são apresentadas.

Na etapa 1 (Descrição do(s) Grupo(s)), os grupos que serão avaliados são caracterizados. Caso sejam poucos grupos, pode-se nomeá-los. Em caso de múltiplos grupos, é importante explicitar a quantidade dos grupos que serão avaliados. Nesta etapa deve-se descrever que tipo de grupo está sendo analisado, podendo ser um grupo de países, instituições ou departamentos. O grupo pode ainda ser um laboratório de pesquisa, um grupo de pesquisa, ou outra configuração qualquer que reúna dois ou mais pesquisadores que precisem sem avaliados.

Os objetivos da avaliação, etapa 2, podem ser múltiplos, tais como:

i. buscar uma comparação entre grupos, ou seja, analisar dois ou mais grupos existentes e comparar as diferentes métricas advindas de cada grupo (ROSAS, 2015; ZHANG et al., 2021a).
ii. buscar uma ordenação através de uma métrica única ou composição de métricas (AUSLOOS, 2013; ÇAKIR et al., 2019).
iii. analisar a produtividade a partir do financiamento dispendido às pesquisas realizadas (WANG et al., 2021).
iv. verificar a similaridade entre dois grupos a partir dos indicadores analisados.
v. realizar a seleção de grupos que atendam a um determinado conjunto de critérios.

A explicitação dos objetivos é importante para que os estudos com objetivos similares possam ser comparados. Além disso, os objetivos auxiliam na delimitação dos indicadores que serão definidos ao final da aplicação do *F-GROUP.*

A etapa 3 (Fontes de Dados) visa explicitar quais são as fontes dos dados que serão utilizadas na avaliação do Grupo. Podem ser bases científicas internacionais podem ainda ser bases específicas de países ou de áreas de pesquisa. Nesta etapa é também importante deixar claro quais foram os tratamentos realizados na base de dados, se a base original foi incrementada com dados de outras bases ou se o nome dos autores passou por algum tratamento nos dados. Todos os tratamentos realizados na base de dados devem estar descritos em detalhes.

Estas três primeiras etapas são de suma importância para permitir a comparação entre diferentes trabalhos. Definir os grupos que serão avaliados, os objetivos da avaliação e as fontes de dados permitem que trabalhos como revisões sistemáticas, e outros trabalhos que sumarizam a pesquisa científica, possam ser realizados de forma mais segura e eficaz.

5.2.2 II) Seleção dos artigos

A fase de seleção de artigos engloba duas etapas: Formas de contagem e agregação (4) e Aspectos representativos (5). Esta fase visa dirimir o problema da falta de padronização e explicitação na forma de seleção, agregação e contagem dos artigos.

As formas de contagem e agregação de artigos em avaliações de grupos de autores vêm sendo discutidas recentemente (GAUFFRIAU, 2021; SANTOS; DUTRA, 2022). A forma sucessiva é exclusiva para a aplicação de índice h e suas variações. Enquanto isso, as abordagens *IN-GROUP* (SANTOS; DUTRA, 2022), *Média* e *Todos os artigos* são aplicáveis a outros índices, como contadores de artigos e citações, por exemplo.

O foco do *framework F-GROUP* está, portanto, na definição da forma como os artigos serão selecionados ao se avaliar um grupo. As formas de agregação devem estar bem delimitadas e especificadas para permitir a comparação entre diferentes trabalhos que avaliam grupos de autores.

A etapa 5 engloba os aspectos representativos do grupo, que podem variar de acordo com o objetivo da análise pretendida e com as fontes dos dados. Pode-se

definir uma janela temporal dos trabalhos que serão analisados de forma a atender o objetivo da análise. Outro aspecto que merece atenção é se a quantidade de autores será levada em consideração, ou seja, se somente grupos com o mesmo número de participantes serão comparados. Existe ainda a possibilidade de delimitar os artigos de acordo com a área que está sendo investigada, para quando o objetivo for selecionar a produção relacionada a um assunto específico. Adicionalmente, ao se avaliar um país, instituição ou até um departamento, é necessário definir se serão utilizados somente os artigos que foram produzidos no local que está sendo avaliado.

O Grau de colaboração interno, aqui proposto, também deve ser definido, uma vez que a análise pode exigir um envolvimento de um número maior ou menor dos autores nas produções bibliográficas. Outro aspecto relevante é se serão ou não considerados os valores de citação onde houve autocitação. Ainda nessa etapa, outros critérios podem ser utilizados de acordo com os objetivos da avaliação, como restringir os artigos de um determinado conjunto de revistas ou conferências, dentre outros critérios necessários.

No decorrer do trabalho algumas questões foram levantadas, dentre elas: O número de autores em cada artigo deve ser considerado? Deve-se considerar todos os artigos publicados por um autor que trabalhou em várias instituições ou apenas naquelas publicadas durante o período no qual esta pessoa trabalha em sua instituição atual? Como balancear esses vínculos ao analisar a produção de um pesquisador quando o intuito é avaliar o trabalho de um grupo? Deve-se considerar a proximidade dos pesquisadores em seus campos de trabalho? Devemos considerar todas as citações em que um dos autores faz parte do grupo? Ou devemos nos concentrar apenas nos artigos em que todos os membros do grupo participam? De forma mais ampla: Quais artigos devem ser selecionados para que os indicadores representem a produção científica do grupo?

Na Fase II do Framework temos indicativos de como responder essas questões. No entanto, não são questões fáceis ou diretas. Muitas dessas perguntas vão depender do que foi definido na Primeira Fase: Identificação dos Parâmetros Iniciais. Estas questões estão diretamente ligadas aos grupos que serão analisados, aos objetivos da avaliação e às fontes de dados utilizadas. Por exemplo, ao avaliar uma instituição, com o objetivo de analisar a produção de artigos ao longo dos anos, pode-se utilizar somente os artigos produzidos pelos autores enquanto fazem parte da instituição. No entanto, se o objetivo é analisar as redes colaborações dos seus

pesquisadores, pode ser mais interessante incluir os artigos que foram produzidos por seus autores, mesmo quando não faziam parte da instituição.

5.2.3 III) Definição de indicadores

A última fase é a que especifica os indicadores que serão utilizados para avaliação dos grupos. Caso mais de uma forma de agregação seja utilizada, deve-se especificar quais as diferentes formas de agregação utilizadas para cada indicador. Esta fase é a que finaliza os aspectos que envolvem planejamento de uma avaliação de produção científica em grupos de pesquisadores.

Diferentes indicadores podem ser utilizados. Desde os mais simples, tal como o total de citações e de artigos, ou como o de colaborações, que pode representar o total de colaborações entre diferentes países, instituições ou outros autores. Diferentes indicadores, como, por exemplo, medidas de acesso aos artigos na *Internet*, também podem ser utilizados. Pode-se, ainda, utilizar os indicadores baseados em índice h para grupos, por exemplo: os índices h_1 e h_2, a abordagem sucessiva, o CP-*index* e a taxa h_2/h_1. Pode-se também utilizar o indicador Crown e demais abordagens FSS (ABRAMO; D'ANGELO, 2015), $JIF_{PERCENTILE}$ (PILCEVIC; JEREMIC; VUJOSEVIC, 2018), *Field Weighted Citation Impact* (CICERO; MALGARINI, 2020), Fator de Impacto de Garfield (MORENO-DELGADO; GORRAIZ; REPISO, 2021), além da metodologia f^2 (FASSIN, 2021a).

Ao aplicar as 3 fases do *framework*, é possível delimitar a avaliação de um grupo de pesquisadores. A aplicação do *framework* é útil tanto para o planejamento de uma avaliação, quanto para a comparação de diferentes estudos envolvendo grupos de autores. A seguir, são apresentados cenários de aplicação do *framework* proposto, de forma a explorar sua aplicabilidade em cenários comuns, conhecidos e identificados na revisão sistemática da literatura.

5.3 CENÁRIOS DE APLICAÇÃO

Com o intuito de facilitar a compreensão da aplicabilidade do *framework F-GROUP* serão apresentados três cenários de uso. No Cenário A, temos um edital para disponibilizar recursos de forma que seja desenvolvido um novo Projeto de Pesquisa. No Cenário B, temos uma análise da colaboração interna entre pesquisadores de um mesmo país com base em trabalhos da *Web Of Science*. No Cenário C, o *framework*

F-GROUP e a abordagem *IN-GROUP* são apresentados como forma de verificação de uma política de interdisciplinaridade entre departamentos de uma universidade.

5.3.1 Cenário Fictício A – Edital de Projeto de Pesquisa.

Neste cenário, a Fundação para Avanço da Pesquisa (fictícia) abriu um edital para financiar um projeto de pesquisa no valor de 3 milhões de dólares para pesquisas relacionadas à erradicação do vírus Ebola. De acordo com o *framework F-GROUP*, na primeira fase temos a identificação dos parâmetros iniciais: Descrição dos Grupos, Objetivos da avaliação e Fontes de dados, Quadro 13. Neste cenário foi definido, por fins de simplificação, que os grupos devem ter 3 pesquisadores. O objetivo é avaliar os grupos, organizando-os em uma lista para selecionar os primeiros de acordo com um indicador.

Quadro 13 – Aplicação do *framework F-GROUP* no Cenário Fictício A

I) Identificação de Parâmetros Iniciais	1.Descrição do(s) grupo(s)	Grupos de pesquisa com 3 pesquisadores.
	2.Objetivos da avaliação	Ordenar: o grupo melhor avaliado será selecionado. Identificar um único grupo de 3 pesquisadores que possuam pesquisas relevantes e que trabalhem em conjunto. Com pesquisa bem estabelecida na área.
	3.Fontes de Dados	PubMed
II) Seleção dos Artigos	4.Formas de contagem e agregação	*IN-GROUP*
	5.Aspectos Representativos	Janela temporal: 10 anos Número de autores: 3 ou mais Área: Medicina – vírus Ebola Grau de Colaboração Interno: 3 Autocitação: não avaliado
III) Definição de Indicadores	6.Listagem de Indicadores	índice h com abordagem *IN-GROUP*

Fonte: Elaborado pela autora

Na fase de seleção de artigos, a Fundação para Avanço da Pesquisa, neste edital gostaria de privilegiar uma equipe que já possua pesquisas relevantes, seja atuante na área e que produza bem em conjunto. Por este motivo a abordagem *IN-GROUP* será utilizada como forma de contagem e agregação dos artigos.

A etapa de aspectos representativos permite avaliar os parâmetros para considerar os trabalhos que serão utilizados na avaliação do grupo: janela temporal, se o número de autores de cada trabalho será considerado, se somente os trabalhos realizados em uma área específica serão cobertos. Nesta etapa o grau de colaboração interno também deve ser definido. Neste cenário, somente os trabalhos onde todos os 3 integrantes do grupo participaram deverão ser considerados. A questão da autocitação não será considerada. Os itens que englobam os aspectos representativos estão descritos no Quadro 13. Como indicador, será utilizado somente o índice h.

Neste edital foi registrado interesse de 3 grupos (Quadro 14). Os Grupos inscritos possuem composições diferentes. Em comum, são todos autores atuantes em pesquisas relacionadas ao vírus Ebola e com números altos de publicações e citações. O Grupo A é composto por pesquisadores com um número elevado de produção científica, mas que nunca trabalharam em conjunto, formaram a equipe para trabalhar no edital, é o grupo com o maior valor de índice H na média. O Grupo B é formado por uma dupla de pesquisadores que iniciaram um trabalho recente, possuem algumas publicações em conjunto e chamaram o maior pesquisador da área para compor a equipe, a fim de aumentar os indicadores gerais e terem mais chance no processo seletivo. O Grupo C é formado por pesquisadores que atuam em conjunto há vários anos, possuindo uma robusta pesquisa na área em conjunto.

Para este edital, 3 grupos de pesquisadores se inscreveram:

Quadro 14 - Cenário A - Grupos avaliados

Grupo	Autores	Características
A	Smith, Johnson e Williams	Pesquisadores que nunca publicaram em conjunto.
B	Silva, Oliveira e Souza	Maior pesquisador na área e dupla com trabalho recente.
C	Wang, Li e Zhang	Pesquisadores com várias pesquisas em conjunto ao longo dos anos.

Fonte: Elaborado pela autora

Para estes grupos, foram calculados os valores de índice h considerando as diferentes formas de contagem e agregação: Sucessiva, Média, Todos os artigos e a abordagem *IN-GROUP* na Tabela 5:

Tabela 5 - Resultados das diferentes abordagens do Cenário A

Grupo	Sucessiva	Média	Todos os artigos	*IN-GROUP*
A	15	18.6	32	**0**
B	10	22.4	36	**1**
C	12	18.4	20	**12**

Fonte: elaborado pela autora.

Neste cenário podemos observar que se o trabalho em conjunto dos pesquisadores não fosse considerado, muito provavelmente os Grupos A ou B seriam escolhidos no edital. O Grupo A possui valores maiores em quase todas as formas de cálculo do índice h, no entanto, nunca trabalhou em conjunto.

Existem cenários para os quais o objetivo é reunir um grupo seleto de pesquisadores e incentivar que trabalhem em conjunto. Para tais cenários, as demais abordagens já atenderiam. No entanto, ao considerar os trabalhos em conjunto, requisito do edital em questão, a abordagem *IN-GROUP* é a única que considera os trabalhos em conjunto destes pesquisadores. Ao se utilizar esta estratégia, é ainda possível atenuar casos em que um pesquisador renomado é adicionado ao grupo para aumentar os indicadores gerais. O cenário com grupos de três pesquisadores foi feito para facilitar a compreensão. O *framework F-GROUP* e a abordagem *IN-GROUP* podem ser utilizados para grupos maiores e com diferentes valores de Grau de Colaboração Interno.

5.3.2 Cenário Fictício B – Comparação entre países acerca da colaboração interna de seus pesquisadores

No Cenário B o objetivo é comparar a colaboração interna para dois países diferentes: Índia e China. Neste cenário serão fornecidas duas listagens, contendo os nomes de todos os pesquisadores de cada país. Será utilizada a abordagem *IN-GROUP*, onde apenas os trabalhos que contenham três ou mais autores de cada país serão considerados, ou seja, Índice de Colaboração = 3.

O Quadro 15 apresenta a aplicação do *framework F-GROUP*. Para cálculo dos indicadores foi realizada uma busca na *Web Of Science*, na qual se recuperou todos os artigos dos autores que constam em cada lista.

Quadro 15 - Aplicação do *framework F-GROUP* no Cenário Fictício B

I) Identificação de Parâmetros Iniciais	1.Descrição do(s) grupo(s)	Serão dois grupos. O primeiro grupo é composto por todos os pesquisadores da China e o segundo grupo por todos os pesquisadores da Índia.
	2.Objetivos da avaliação	Comparar países. Identificar qual país possui o maior valor de índice h.
	3.Fontes de Dados	*Web Of Science*
II) Seleção dos Artigos	4.Formas de contagem e agregação	*IN-GROUP*
	5.Aspectos Representativos	Janela temporal: 10 anos Número de autores: 3 ou mais Área: Sem definição de área, toda a produção dos pesquisadores será considerada. Grau de Colaboração Interno: 3 Autocitação: não avaliado Adicional: autores com dupla nacionalidade serão considerados para o cálculo dos indicadores nos dois países.
III) Definição de Indicadores	6.Listagem de Indicadores	índice h com abordagem *IN-GROUP*

Fonte: Elaborado pela autora

Em seguida os valores de índice h com a abordagem IN-GROUP devem ser calculados, considerando os aspectos representativos e formas de agregação definidas. Ou seja, apenas artigos dos últimos 10 anos foram considerados e apenas os artigos onde 3 ou mais autores do mesmo país foram considerados.

Este é um cenário que foi trazido para demonstrar o poder de sintetização da do índice h com abordagem *IN-GROUP*. Um único valor é capaz de representar a colaboração interna de um país inteiro, podendo ser comparado com outros países. Valores com dados reais utilizando o *F-GROUP* para comparação entre países serão apresentados no próximo capítulo.

5.3.3 Cenário Fictício C – Universidade U

A Universidade U deseja analisar a colaboração entre os seus departamentos ao longo dos anos. A atual reitora foi eleita com uma proposta de aumentar a interdisciplinaridade dentro da universidade e gostaria de analisar a evolução deste quesito. O Quadro 16 apresenta uma proposta de avaliação deste cenário utilizando o *framework F-GROUP*.

Quadro 16 - Aplicação do *framework F-GROUP* no Cenário Fictício A

I) Identificação de Parâmetros Iniciais	1.Descrição do(s) grupo(s)	Docentes pesquisadores da Universidade U
	2.Objetivos da avaliação	Acompanhar a colaboração interna entre departamentos ao longo dos anos.
	3.Fontes de Dados	*Web Of Science*
II) Seleção dos Artigos	4.Formas de contagem e agregação	*IN-GROUP* Serão contabilizados somente os artigos onde docentes de 2 departamentos diferentes colaboraram.
	5.Aspectos Representativos	Janela temporal: 10 anos Número de autores: 2 ou mais Área: Sem definição de área, toda a produção dos pesquisadores será considerada. Grau de Colaboração Interno: 2 Autocitação: não avaliado
III) Definição de Indicadores	6.Listagem de Indicadores	índice h com abordagem *IN-GROUP*

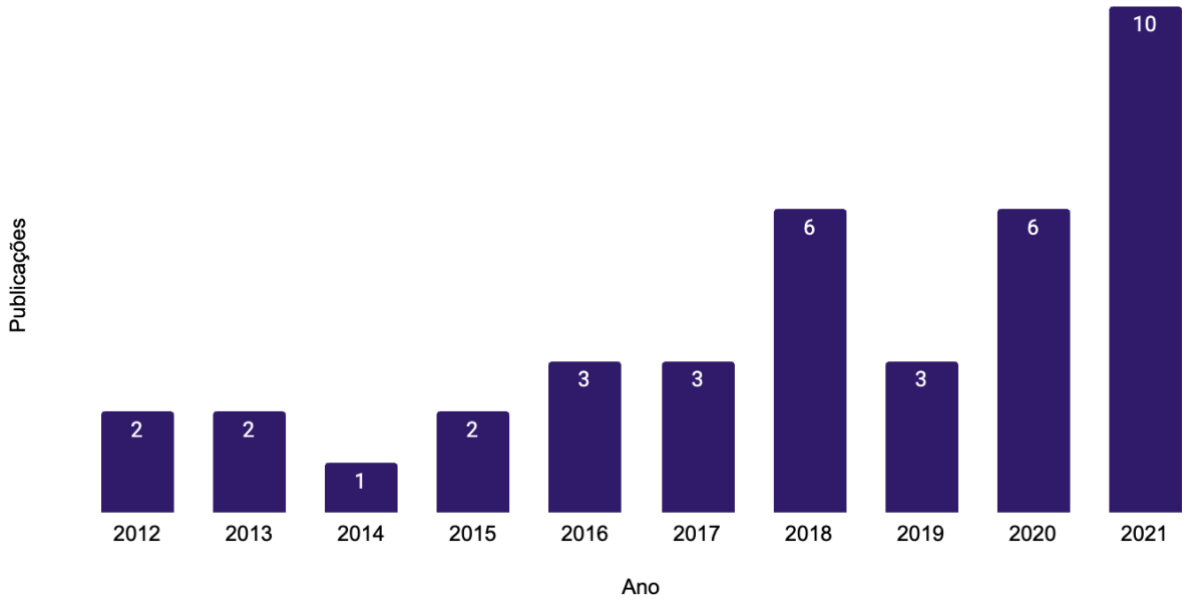
Fonte: Elaborado pela autora

A título de exemplo, vamos supor que o indicador de colaboração entre departamentos aumentou. Um exemplo de gráfico que demonstraria uma evolução da colaboração é apresentado na Figura 24. Considerando que neste cenário fictício: a) A Universidade U possui 50 departamentos; b) os departamentos têm em média 20 docentes; c) a produção média dos docentes é de 8 artigos por ano; e d) a média de produção em colaboração com docentes de outros departamentos é de 3% ao ano; estimamos que o valor da colaboração interna entre departamentos com índice h *IN-GROUP* da Universidade U é atualmente de 120. Lembrando que cada artigo é

considerado de forma única na abordagem *IN-GROUP*. Portanto, os artigos produzidos por docentes de dois departamentos diferentes serão considerados apenas uma vez para cada departamento.

No cenário fictício proposto, foi possível identificar que as ações propostas na Universidade U contribuíram para que a colaboração entre os departamentos fosse acelerada. No cenário fictício C também foi considerada uma queda ocorrida em 2020 devido à Pandemia de Covid-19.

Figura 24 - Índice h *IN-GROUP* da Universidade U - Cenário FICTÍCIO C



Fonte: elaborado pela autora

Através do cenário proposto foi possível utilizar o *framework F-GROUP* para delimitar a avaliação. Adicionalmente, fica demonstrada a utilidade da abordagem *IN-GROUP* para análise da colaboração interna para instituições de ensino e pesquisa.

A aplicação em cenários-exemplo é útil para a demonstração das diferentes situações em que o *F-GROUP* pode ser utilizado. Na próxima seção são apresentadas avaliações com dados reais, extraídos da *Web Of Science*, aplicando o framework *F-GROUP* e a abordagem *IN-GROUP* para avaliação de grupos de Países e Universidades na área de Ciência da Informação.

6 ANÁLISE E DISCUSSÃO DOS RESULTADOS

O último objetivo específico deste trabalho é comparar os dispositivos propostos com as formas de avaliação consolidadas na literatura. Com este intuito, utilizaremos dados de grupos reais da produção científica da área de Ciência da Informação e Biblioteconomia. Os dados analisados neste capítulo foram coletados conforme disposto no APÊNDICE A - Coleta dos dados. O tratamento dos dados ocorreu conforme descrito na Seção 3.4 EXTRAÇÃO E TRATAMENTO DE DADOS PARA EXPERIMENTOS COM DADOS REAIS e APÊNDICE B – Código em SQL para Tratamento dos Dados.

Primeiramente são realizadas análises acerca do Número de Autores (6.1) e de Citações (6.2). Uma avaliação aplicando o *framework F-GROUP* para análise da produção científica de países é apresentada na seção 6.3. Na seção 6.4, utilizando a mesma base de dados, há uma avaliação das principais universidades com produção científica em Ciência da Informação nos últimos 10 anos. Histogramas com as primeiras universidades são analisados de acordo com diferentes métricas e indicadores. Os detalhes das avaliações dos diferentes grupos são apresentados conforme o *framework F-GROUP*. O índice h é calculado utilizando a abordagem *IN-GROUP*. Na última seção apresentamos um resumo comparativo entre as formas de avaliação consolidadas na literatura e a combinação de IN-GROUP e F-GROUP.

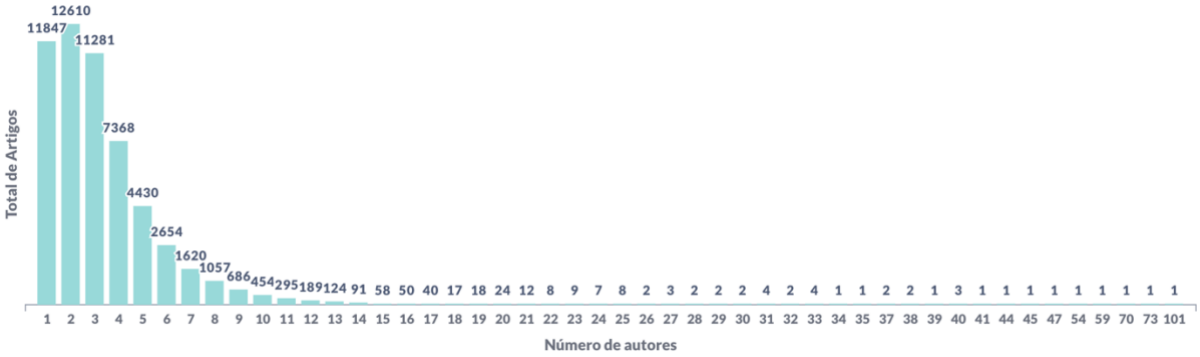
6.1 ANÁLISE DO NÚMERO DE AUTORES

Após a normalização dos autores, foi possível criar uma coluna na tabela de artigos contento o número de autores. A criação de colunas com valores sumarizados não é comum na modelagem de dados relacional, uma vez que este valor pode ser obtido através de uma consulta com *joins*. No entanto, optou-se por criar esta coluna para facilitar a geração de relatórios e gráficos.

O total de artigos para cada tamanho de grupo é apresentado na Figura 25. É possível observar uma concentração em artigos com 1, 2 ou 3 autores. O número de artigos com grupos maiores de autores reduz de forma bastante significativa a partir de 4 autores por artigo. O número máximo de autores por artigo foi um registro com 99 autores. O total de autores por artigo apresenta uma curva com distribuição

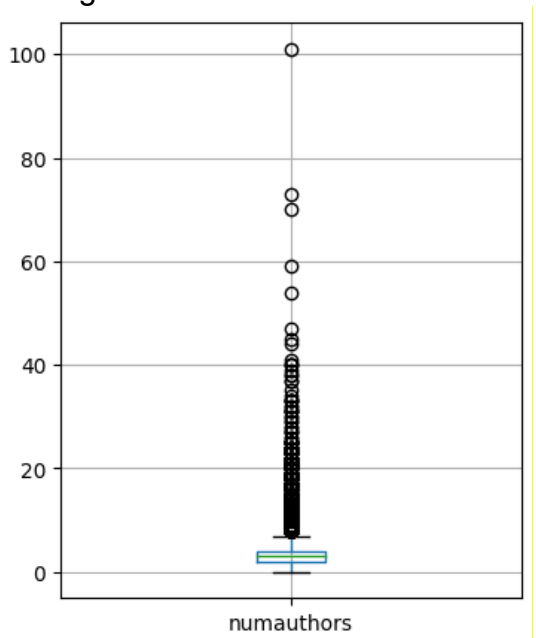
assimétrica positiva, onde os dados estão concentrados após a média, com uma curva alongada à direita.

Figura 25 - Total de artigos com mesmo Número de Autores



Fonte: dados da pesquisa (2023).

Figura 26 - Diagrama de Caixa com Número de Autores



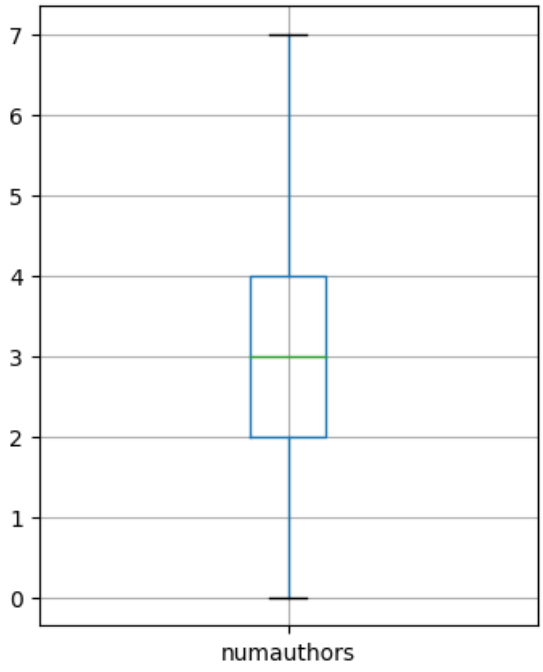
Fonte: dados da pesquisa (2023).

Quando ocorre assimetria, a média pode não ser a melhor forma de representar os dados (BARBETTA; REIS; BORNIA, 2004). Para analisar melhor a distribuição do número de autores no conjunto de dados analisado criamos um diagrama de caixa (também chamado de *boxplot*), Figura 26. Neste diagrama, o centro representa a mediana, a linha superior do retângulo representa o primeiro quartil, e a linha inferior o terceiro quartil. Os limites inferior e superior são representados pelos traços abaixo e acima da caixa, respectivamente. Os demais valores são ditos registros

discrepantes e são representados por pequenos círculos. Como foi possível observar, tanto no diagrama de caixa quanto no histograma, diversos registros discrepantes foram registrados. Dentro da caixa fica localizado o conjunto de dados onde estão presentes 50% dos valores mais prováveis.

Para que pudéssemos visualizar melhor os valores, removemos, apenas para uma breve análise, os valores discrepantes (ou *outliers*) para produzir um novo diagrama de caixa. Neste conjunto de dados reduzido, o valor da mediana ficou em 3 autores, limite superior foram 4 autores, limite inferior de 2 autores por artigo. Os valores acima de 7 foram marcados como discrepantes, não sendo considerados. Foi verificado que somente 5% dos artigos possuem mais do que 7 autores.

Figura 27 - Diagrama de Caixa com Número de Autores (análise com remoção de valores discrepantes)



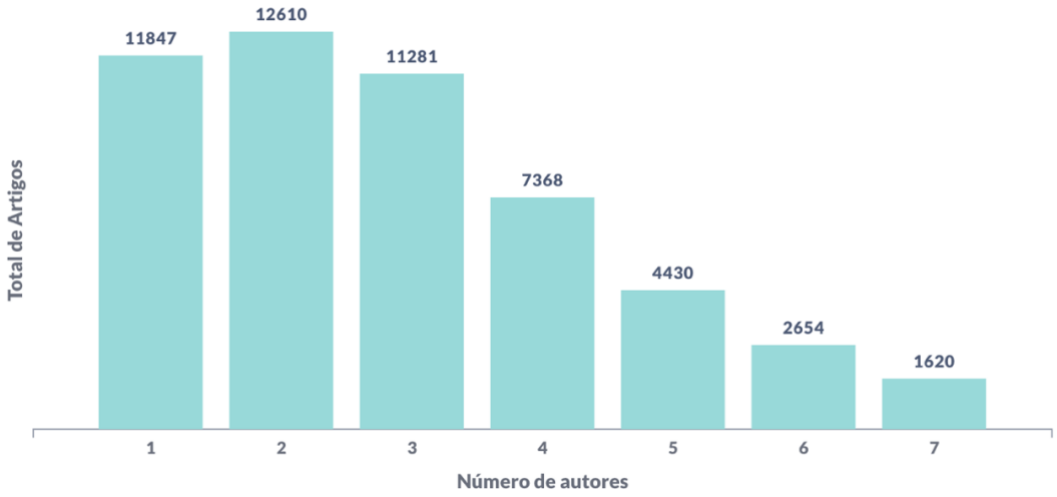
Fonte: dados da pesquisa (2023).

A moda do número de Autores é 2 (Figura 25), ou seja, o tamanho de grupo mais comum no conjunto de dados analisado são 2 autores. Já a mediana ficou em 3 autores (Figura 27), ou seja, ao selecionar o valor na metade da lista ordenada, temos um conjunto com 3 autores. A média do grupo, considerando os valores discrepantes foi de 3,33. Ao remover os valores discrepantes a média ficou em 2,89 autores. Ao verificar estes valores e o Diagrama de Caixa presente na Figura 27, podemos afirmar que mais de 50% dos artigos possui entre 2 e 4 autores. Constatamos, ainda, que a

grande maioria dos artigos (94,2%) possui entre 1 e 7 autores.

O número de autores é uma variável que apresenta alguns valores bastante discrepantes, o que, inicialmente, pode significar que seja um conjunto de dados que foge à distribuição normal. Analisando o diagrama de caixa da Figura 27, após a remoção dos valores discrepantes, a variável parece se adequar um pouco mais à curva normal. Para analisar a normalidade da variável Número de Autores apresentamos um histograma na Figura 28.

Figura 28 - Total de artigos com mesmo Número de Autores (análise com remoção de valores discrepantes)



Fonte: dados da pesquisa (2023).

Nesta seção analisamos a variável número de autores por meio de uma abordagem gráfica utilizando histogramas (Figura 25 e Figura 28) e diagramas de caixa (Figura 26 e Figura 27). Inicialmente o diagrama de caixa da Figura 27 (onde houve a remoção de valores discrepantes) sugeriu que os dados poderiam ter uma distribuição normal, considerando que mediana e quartis pareciam alinhados. No entanto uma análise a partir dos dados presentes no histograma da Figura 28 revelou outras nuances sobres esta variável.

Ao analisar o histograma sem valores discrepantes observamos que a distribuição dos valores não é simétrica, mas possui uma assimetria positiva. A maioria dos valores se concentra à direita da média. Esta assimetria indica que uma parcela dos artigos analisados possui um número significativamente maior do que a média observada do número de autores.

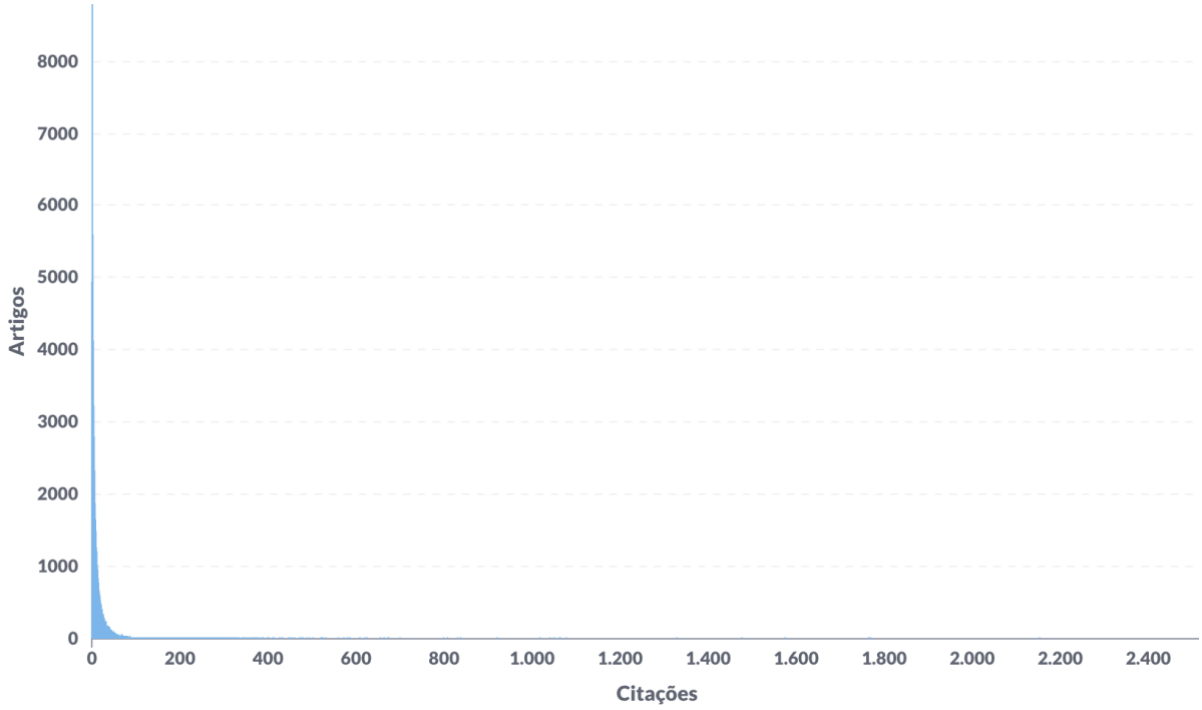
A assimetria positiva observada pode ser resultado de diferentes fatores. Dentre eles, destacamos os valores discrepantes com muitos autores (Figura 25). No

entanto, outros fatores podem estar relacionados à composição de grupos maiores de autores. Em alguns casos a pesquisa necessita de um número maior de colaborações. Remover ou manter os valores discrepantes pode influenciar as análises ao se comparar com outras variáveis. Por este motivo, a remoção de valores discrepantes foi utilizada somente para uma breve análise visual da variável número de autores. Nas demais análises, os valores discrepantes do número de autores serão mantidos.

6.2 ANÁLISE DO NÚMERO DE CITAÇÕES

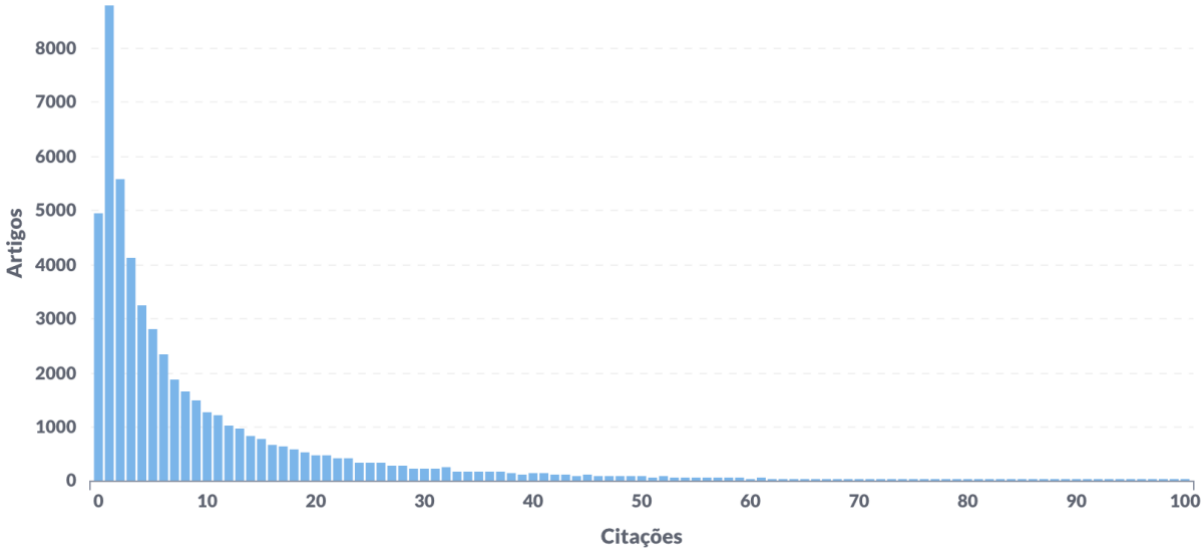
O número de citações analisado é o original extraído dos registros da *Web Of Science* em 10 de abril de 2023. De forma semelhante ao número de autores, o histograma com o total de citações apresenta uma curva com distribuição assimétrica positiva, onde os dados estão concentrados após a média, com uma curva alongada à direita, Figura 29. É possível observar uma concentração nos artigos com menos citações, mais próximo de 1, uma vez que os artigos com 0 citações não foram importados. Um segundo histograma foi trazido para que pudéssemos visualizar melhor a parte inicial do gráfico, sem os valores discrepantes, Figura 30. Há uma grande concentração de artigos que possuem até 10 citações. A moda é de 2 citações.

Figura 29 - Total de artigos por total de citações



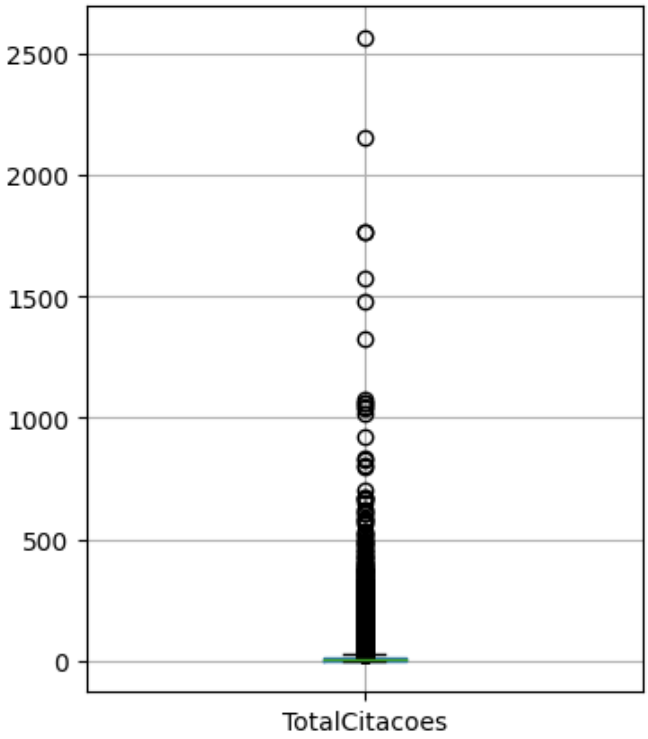
Fonte: dados da pesquisa (2023).

Figura 30 - Total de Artigos por Total de Citações sem valores discrepantes



Fonte: dados da pesquisa (2023).

Figura 31 - Diagrama de Caixa com Total de Citações

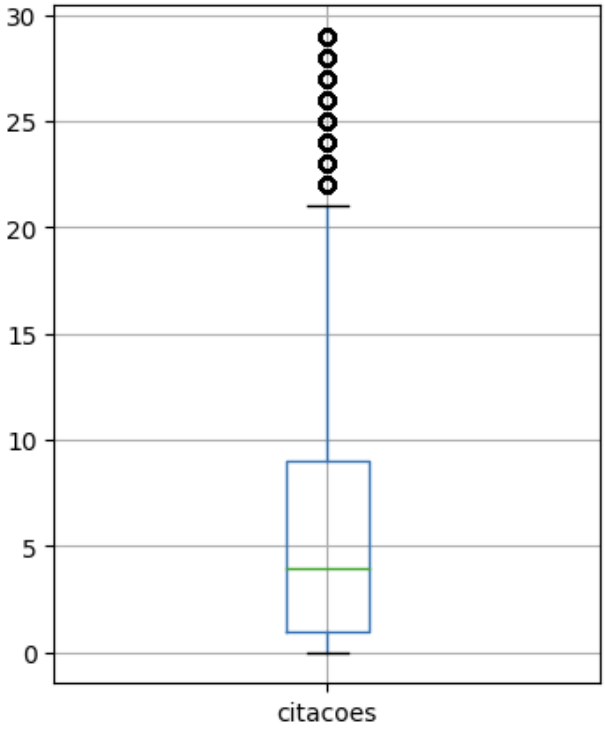


Fonte: dados da pesquisa (2023).

O Diagrama de Caixa das citações, apresentado na Figura 31, ficou pouco visível dado o grande número de valores de citações discrepantes. Neste conjunto a média ficou em 13,84 citações. Um segundo Diagrama de Caixa, removendo, para

uma breve análise, os valores discrepantes de citações é apresentado na Figura 32. Mesmo com a supressão de valores discrepantes de citações ficou clara a natureza, ainda, bastante assimétrica dos dados.

Figura 32 - Diagrama de Caixa com Total de Citações com análise da remoção de valores discrepantes do conjunto de dados inicial



Fonte: dados da pesquisa (2023).

O valor mediano observado ficou com 4 citações, limite superior de 21 citações, o limite inferior foi de 1 (uma) citação. Neste novo conjunto de valores, após a remoção dos valores discrepantes do conjunto inicial, os valores acima de 21 citações foram considerados discrepantes.

O número de citações do conjunto analisado está moderadamente concentrado entre 1 e 9 citações. A concentração de citações entre 1 e 9 foi de 58% dentre os artigos com mais de uma citação. Ao considerar o limite superior, 84% dos artigos possuem 21 ou menos citações dentro do conjunto completo, sem a remoção de valores discrepantes. Essa percentagem tende a ser ainda maior ao considerarmos que dos 121.222 artigos trazidos na busca, apenas 54.998 foram importados para a base de dados. Os 66.224 artigos restantes, com 0 citações, não foram importados, ao considerá-los, temos que 92,8% dos artigos possuem 21 citações ou menos.

Os valores discrepantes, acima de 30 citações na base de dados com 1 (uma) citação ou mais, constituem um conjunto de dados com bastante relevância para as análises, tendo sido removidos apenas para uma melhor visualização da concentração dos dados em diagramas de caixa e histograma. O número de citações é frequentemente utilizado para atribuir relevância a um trabalho. Em nossas futuras análises, todos os valores de citação acima de 1 (um) serão utilizados.

A seguir, serão apresentadas comparações entre as variáveis número de autores e citações. Essas comparações visam entender se há alguma relação entre essas variáveis para que possamos utilizá-las nas futuras análises de grupos de autores.

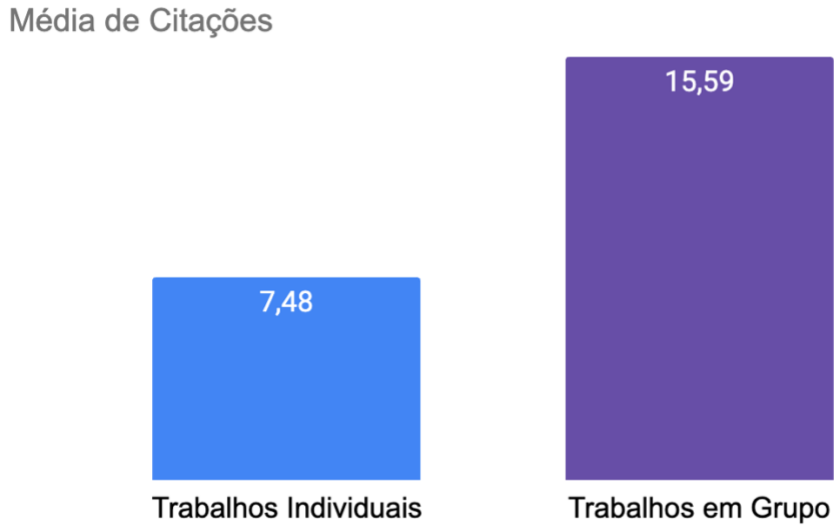
6.2.1 Citações em Trabalhos Individuais e em Grupo

Nesta seção vamos analisar a combinação das variáveis “Número de Autores” e “Citações”. Para esta análise, a variável quantitativa e contínua “Número de Autores” foi discretizada. Os trabalhos foram separados em dois grupos diferentes: o primeiro conjunto contém os artigos em que apenas um autor consta na lista de autoria, chamado de Trabalhos Individuais. O segundo conjunto continha todos os artigos em que dois ou mais autores constam na lista de autoria, chamado de “Trabalhos em Grupo”.

Na Figura 33 temos os valores das médias trazidas para uma primeira inferência sobre a influência do número de autores no número de citações. A média de citações dos trabalhos individuais foi de 7,48 citações. Enquanto a média de citações em trabalhos realizados em grupo quase dobrou, chegando a 15,59 citações. O valor do desvio padrão para os dois grupos também foi verificado. O desvio padrão das citações nos trabalhos em grupo foi de 41,54 e para os trabalhos individuais foi observado o valor de 22,75 citações.

Como abordado na seção anterior, a média pode não ser considerada o melhor valor para analisar dados assimétricos. No entanto, esta assimetria está presente nos dois grupos avaliados, elevando os valores para acima da mediana em ambas as observações.

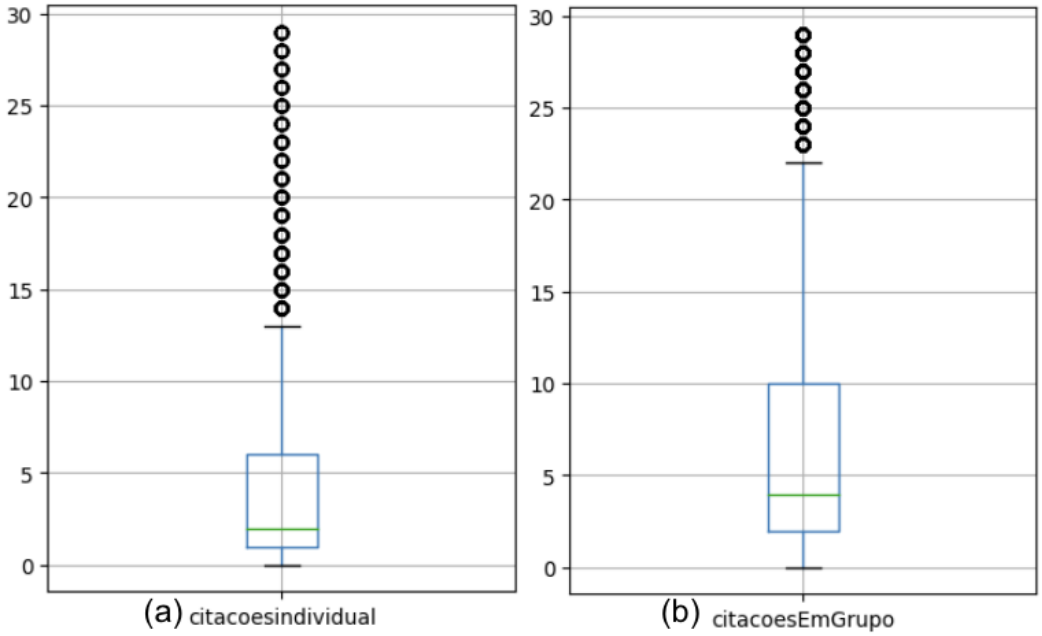
Figura 33 - Comparação entre a média das citações de trabalhos Individuais e trabalhos com dois ou mais autores



Fonte: dados da pesquisa (2023).

Os diagramas de caixa para os dois grupos são apresentados na Figura 34. Na Figura 34 (a) temos o diagrama de caixa para o grupo de trabalhos individuais, onde a mediana ficou em 2 citações, quartil superior com 6 citações e inferior em 1 citação. O diagrama de caixa dos artigos realizados em grupo Figura 34 (b) apresentou mediana de 4 citações, limite superior de 10 citações e limite inferior de 2 citações.

Figura 34 - Comparativo Diagramas de Caixa entre o número de citações em artigos com único (a) ou múltiplos autores (b)

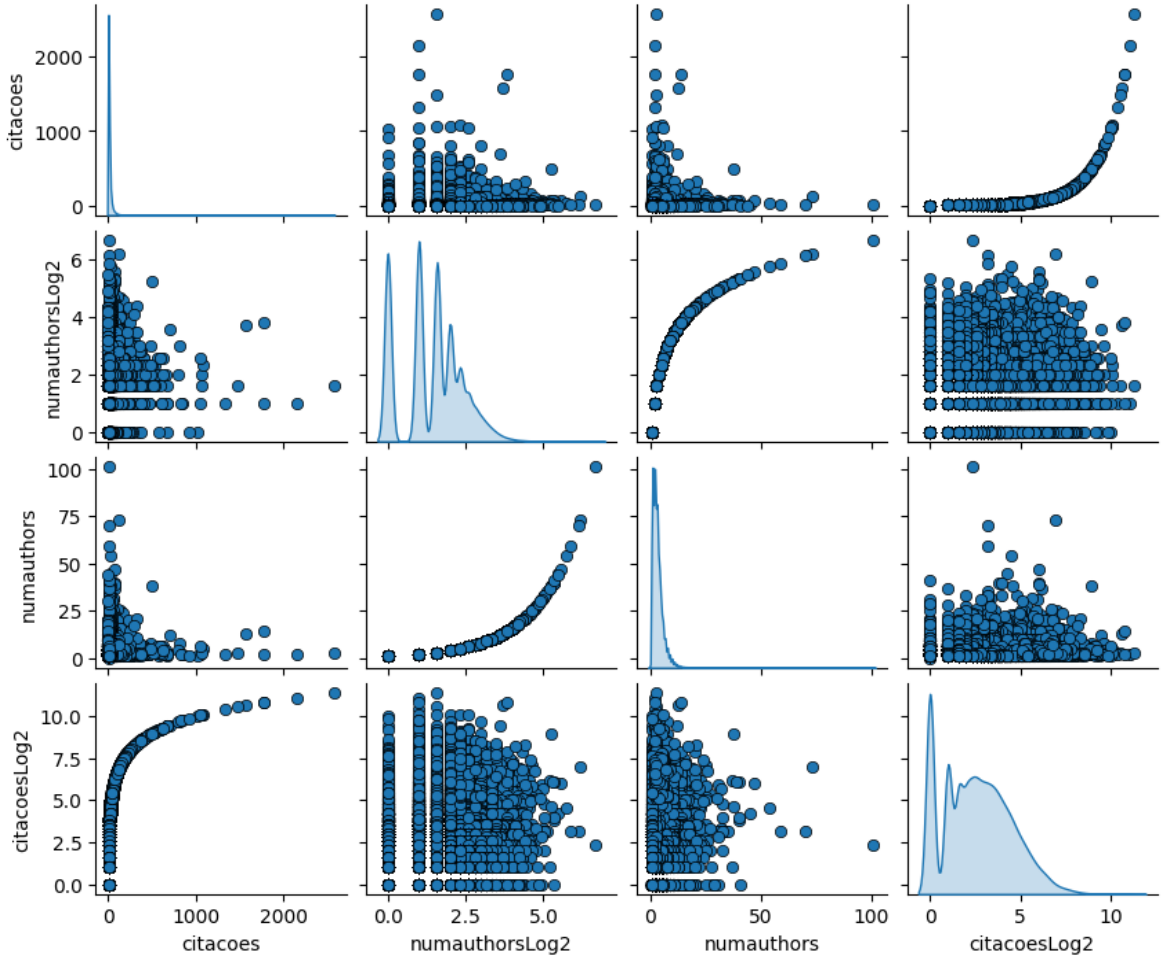


Fonte: dados da pesquisa (2023).

É possível afirmar, portanto, que o valor das citações está associado à realização ou não dos trabalhos em grupo. Os trabalhos em grupo apresentam média de citações maior, bem como mediana e distribuição dos dados de forma geral com valores superiores aos trabalhos realizados individualmente, conforme apresentado na Figura 34. Se a realização de trabalhos em grupo não tivesse associação com a variável de número de citações, os valores de média e mediana seriam mais próximos.

Considerando este indício de associação entre as variáveis “citações” e “número de Autores”, procedemos à análise de correlação entre elas. Como sabemos que as variáveis possuem assimetria, realizamos transformação em escala logarítmica como tentativa de normalizar os dados.

Figura 35 - Análise de Correlação entre número de autores e número de citações



Fonte: dados da pesquisa (2023).

Na Figura 35 temos as comparações entre as variáveis “citações” (citacoes), “Log2 do Número de Autores” (numauthorsLog2), Número de Autores (numauthors) e

Log 2 de citações (citacoesLog2). O primeiro gráfico da Figura 35 apresenta a distribuição da variável citações. Na diagonal temos gráficos de linha com as distribuições das demais variáveis, onde podemos observar novamente a assimetria das variáveis “Número de Autores” e “Citações”. As tentativas de normalização através de Log foram frustradas, conforme apresentados nos gráficos de distribuição das variáveis “Log2 do Número de Autores” (numauthorsLog2) e Log 2 de citações (citacoesLog2).

Ao observar as variáveis originais e suas transformações logarítmicas de forma gráfica (Figura 35), não foi possível encontrar correlações entre diferentes variáveis. As únicas correlações foram entre as variáveis originais e suas transformações logarítmicas. Ou seja, não foi possível encontrar uma correlação entre a variável número de autores e número de citações. Foram realizados testes estatísticos não-paramétricos, que evidenciaram que as amostras de citações para trabalhos em grupo e individuais possuem distribuição diferente, com valor-p próximo de zero, conforme consta no APÊNDICE C – Código em Python para Análise dos Dados.

Existem algumas explicações possíveis para essa ausência de correlação, mesmo após a média e diagramas de caixa indicarem a diferença no número de citações para trabalhos individuais e em grupo. Uma primeira explicação é a própria natureza dos dados, que são assimétricos, com muitos valores discrepantes. Não há também uma relação direta entre as citações e o número de autores, ou seja, aumentar ou diminuir o número de autores não garante um número maior de citações em um artigo. Podemos então afirmar, a partir dos dados apresentados, que não há uma correlação entre o número de autores e o número de citações.

No entanto, há um indício de associação entre o número de citações e a variável discreta que separou os trabalhos em grupo e individuais. Portanto, é possível afirmar que os trabalhos realizados em grupos tendem a ter uma maior média e mediana de citações em artigos analisados com o tópico de Biblioteconomia e Ciência da Informação nos últimos 10 anos.

Considerando a associação entre o número de artigos realizados em grupo e o número de citações, procedemos à análise utilizando o *framework F-GROUP*, aqui proposto. Primeiramente utilizamos países como grupos de autores. Em seguida, os artigos foram agrupados de acordo com as Universidades, dentre as demais instituições que constam na coluna de afiliações dos artigos.

6.3 AVALIAÇÃO DA PRODUÇÃO CIENTÍFICA EM CIÊNCIA DA INFORMAÇÃO POR PAÍS

No escopo deste trabalho propomos a forma de agregação *IN-GROUP* como uma alternativa às diferentes formas de calcular o índice h para grupos de autores. Propomos também um *framework F-GROUP*, para análise de grupos de autores. Este *framework* possui 3 fases: I) Definição dos Parâmetros Iniciais; II) Definição da forma de seleção dos Artigos; e III) Apresentação dos Indicadores que serão utilizados para realizar a avaliação dos grupos de autores.

Quadro 17- Aplicação do *framework F-GROUP* para análise da produção científica em Ciência da Informação em países entre 2013 e 2023

I) Identificação de Parâmetros Iniciais	1.Descrição do(s) grupo(s)	Serão criados 171 grupos. Os grupos são compostos por países que apresentaram algum artigo no campo da Ciência da Informação no período entre 2013 e 2023 com 1 (uma) ou mais citações.
	2.Objetivos da avaliação	Comparação entre países. Analisar os primeiros 15 países, utilizando diferentes indicadores, incluindo o índice h *IN-GROUP* considerando diferentes índices de colaboração.
	3.Fontes de Dados	*Web Of Science*
II) Seleção dos Artigos	4.Formas de contagem e agregação	*IN-GROUP*, *all papers*
	5.Aspectos Representativos	Janela temporal: 10 anos (10 de abril de 2013 a 10 de abril de 2023) Número de autores: 2 ou mais Grau de Colaboração Interno: 2, 4, 6, média de autores da mesma nacionalidade por país. Autocitação: não avaliado. Adicional: autores com dupla nacionalidade serão considerados para o cálculo dos indicadores nos dois países.
III) Definição de Indicadores	6.Listagem de Indicadores	índice h com abordagem *IN-GROUP*, índice h com abordagem *all papers*, total de citações, total de autores, média de autores por artigo, média de citações por artigo.

Fonte: elaborado pela autora

A aplicação do *framework F-GROUP* utilizando os dados coletados na *Web Of Science* para Ciência da Informação e Biblioteconomia é apresentada no Quadro 17. Nesta primeira avaliação os países são os grupos de autores avaliados.

Tabela 6 - Indicadores de Produção científica por país (Top 15)

País	Total de Artigos	Total de Autores	Total de Citações	Média de autores por artigo	Média de autores do mesmo país por artigo	Média de citações
EUA	16881	36072	265426	3,90	2,68	17,36
China	6738	13718	106123	5,48	2,82	16,17
Inglaterra	3882	9453	69439	3,69	1,93	19,82
Espanha	3150	6114	40082	3,27	2,32	13,96
Austrália	2694	6589	43484	3,86	2,10	17,31
Canadá	2501	6422	45786	3,78	2,13	20,11
Alemanha	2453	5613	37394	3,43	2,12	16,50
Índia	2291	4763	25146	3,18	2,34	11,70
Itália	1588	4183	22431	3,75	2,43	15,22
Holanda	1564	3955	37404	3,84	1,99	25,79
Coréia do Sul	1525	3282	25449	3,52	2,15	18,45
Taiwan	1386	2850	24235	3,66	2,17	19,47
Brasil	1372	3343	9537	3,48	2,70	7,27
França	1205	3535	18209	4,11	2,03	16,21
África do Sul	937	1717	7282	2,92	1,69	9,06

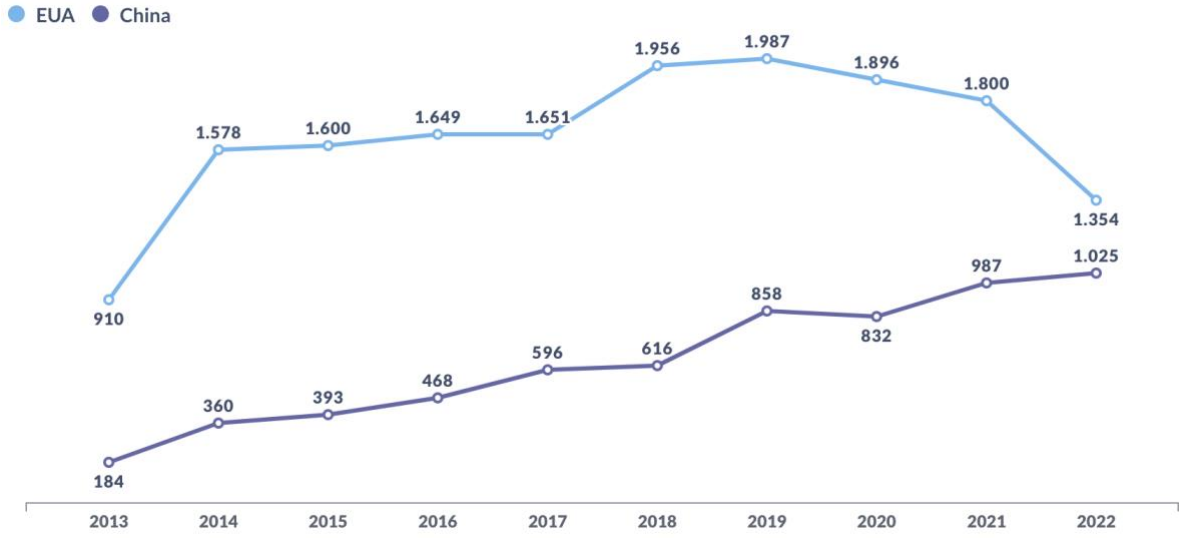
Fonte: dados da pesquisa (2023).

Conforme disposto no Quadro 17 o objetivo desta análise é a Comparação entre países, por meio da análise dos 15 primeiros, com a utilização de diferentes indicadores, como o índice h *IN-GROUP*, e considerando diferentes tamanhos de

grupos. Como subsídio para esta análise, serão apresentados também outros indicadores como totais e média de citações, autores e artigos.

Na Tabela 6 temos os 15 primeiros países ordenados por Total de Artigos. O primeiro país é os Estados Unidos, representado pela sigla EUA, que possui mais que o dobro de produção do segundo país (China). Ao comparar as citações dos primeiros dois países, os valores são ainda mais discrepantes: os EUA possuem quase 3 vezes mais citações que a China. No entanto, ao observar o gráfico da Figura 36 - Total de Artigos por ano EUA e ChinaFigura 36, observamos que a China está ficando a cada ano mais próxima da produção americana. Cabe uma ressalva de que a *Web of Science* é de origem americana, onde o inglês é a língua nativa. Ainda assim, a China segue com valores bastante elevados no número de citações.

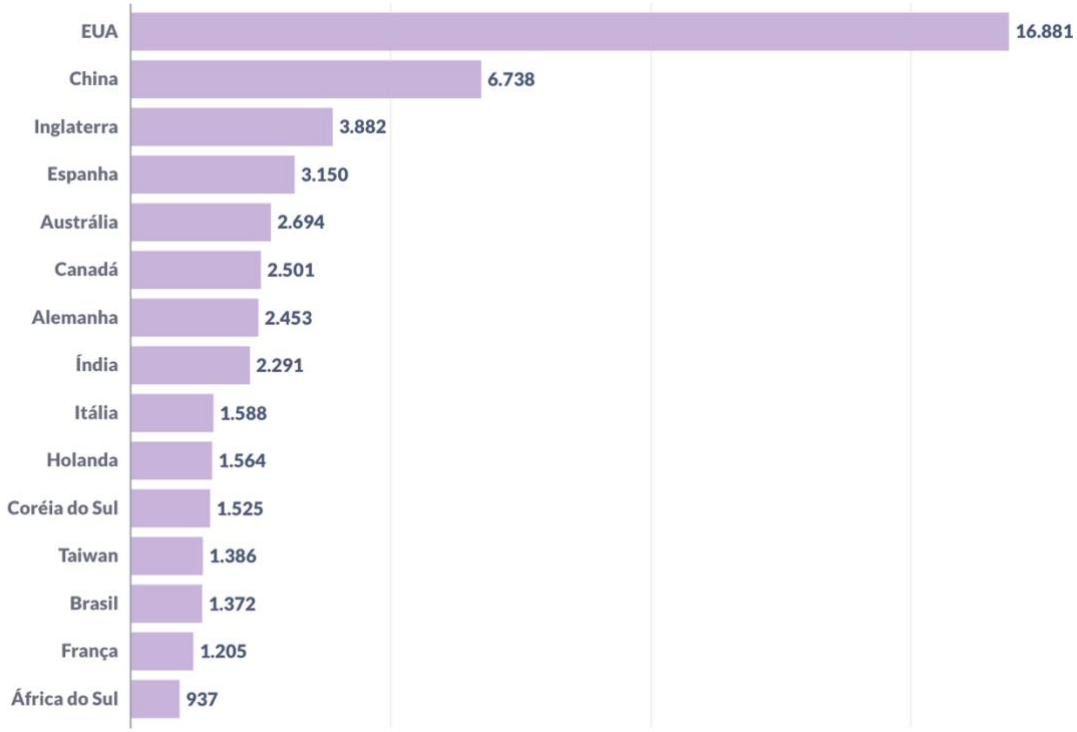
Figura 36 - Total de Artigos por ano EUA e China



Fonte: dados da pesquisa (2023).

Na Tabela 6 são apresentados os valores médios dos grupos de autores. Na coluna média de autores por artigo (coluna 5) é apresentado o valor da média de autores nos artigos que possuem algum autor daquele país. Na coluna seguinte, Média de autores do mesmo país por artigos, temos a média por artigo de autores cuja nacionalidade é do país que está sendo analisado. O último valor médio apresentado é o de citações, que apresenta a média de citações dos artigos onde algum membro é do país em questão. O Brasil apresenta o menor valor de média de citações no grupo, 7,27. Mesmo com nível similar em termos de produção e autores, isso ocorre devido às características da base de dados *Web of Science*, onde a maioria das produções são escritas na língua inglesa.

Figura 37 - Total de artigos por país Top 15 países



Fonte: dados da pesquisa (2023).

EUA e China se destacam dentre os demais países em termos de produção. Na Figura 37 trazemos os primeiros 15 países considerando a produção científica. Após EUA e China aparecem Inglaterra, seguida por Espanha com produção acima de 3 mil artigos nos últimos 10 anos. Em seguida temos Austrália, Canadá, Alemanha e Índia com produção em torno de 2.500 artigos no mesmo período. Um terceiro grupo com Itália, Holanda e Coreia do Sul publicaram pouco mais de 1500 artigos cada. Taiwan e Brasil publicaram pouco mais de 1300 artigos cada, seguidos por França com 1205 artigos e África do Sul com 937 artigos.

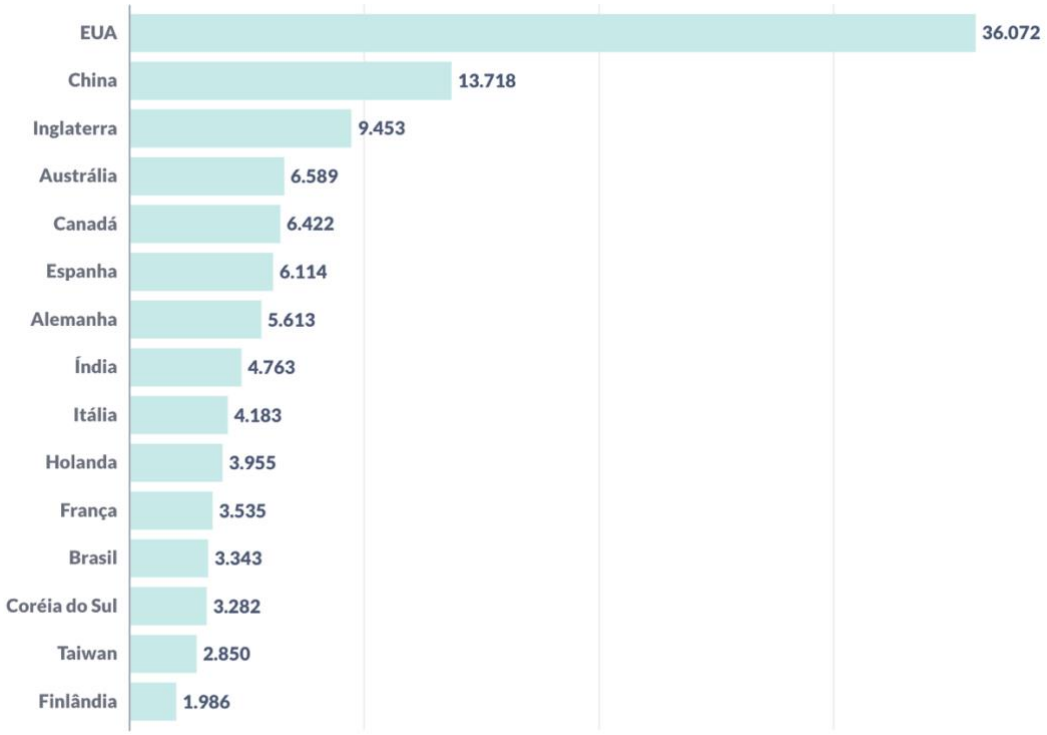
Em relação ao total de Autores (terceira coluna da Tabela 6), temos mais uma vez uma liderança americana, com 36072 diferentes autores. A China aparece em seguida com 13.718 autores. A Figura 38 ilustra os primeiros 15 países quando a ordenação é feita a partir do total de autores de cada país. O terceiro país, bem como no total de artigos, é a Inglaterra com 9.453 autores, seguida por Austrália, Canadá e Espanha, com pouco mais de 6 mil autores cada.

Os primeiros 14 países em publicação de artigos, também se mantiveram dentre os 14 primeiros no quesito número de autores. Mas a África do Sul que figurava

entre os 15 países com maior produção de artigos (Figura 37), foi substituída pela Finlândia quanto ao número de autores (Figura 38).

A Alemanha foi o país com o sétimo maior número de autores, um total de 5.613. Em seguida temos Índia e Itália com 4.763 e 4.183 autores respectivamente. Holanda, França, Brasil e Coréia do Sul podem ser agrupados dentre os países entre 3 e 4 mil autores. Taiwan teve 2850 autores elencados, seguido pela Finlândia com 1.986 autores.

Figura 38 - Total de autores por país Top 15 países



Fonte: dados da pesquisa (2023).

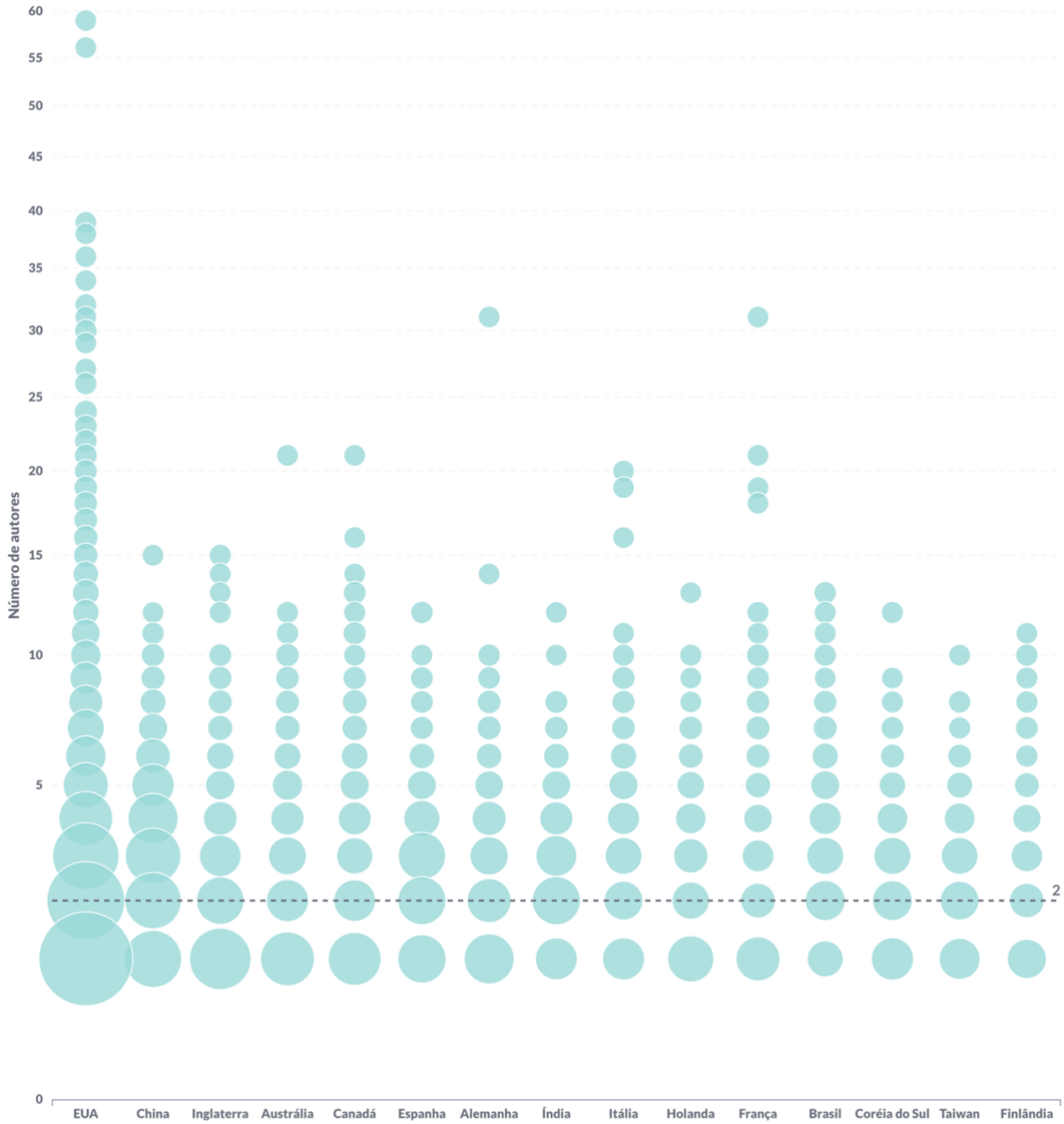
Apesar de EUA e China serem países populosos este não é o único fator a ser considerado ao analisar o número de autores. Índia e Brasil tiveram menos autores que Espanha e Inglaterra, que são países menos populosos, por exemplo. Aspectos socioeconômicos também podem influenciar no número de autores em cada país. Mesmo que o número de autores possa ser diretamente influenciado pelo total de habitantes de cada país, na amostra analisada não há uma relação direta quanto ao número de habitantes e de autores na área de Ciência da Informação. Enquanto a Índia é atualmente o país mais populoso do mundo[14], este país fica em oitavo no

[14] Em abril de 2023 a Índia ultrapassou a China como país de maior população, de acordo com a Organização das Nações Unidas. Disponível em <https://population.un.org/wpp/> acesso em 28 de Dezembro de 2023.

número de autores. É fato que a Índia assumiu a posição de país mais populoso do mundo recentemente, em 2023, mas já estava na segunda posição no *ranking* de países mais populosos há décadas.

A população americana, atualmente com 334 milhões de pessoas, não chega a um quarto da população dos países mais populosos do mundo que possuem populações na casa dos 1,4 bilhão cada. A Indonésia, quarto país mais populoso no mundo não possui tantos autores a ponto de estar dentre os 15 países com maior número de autores.

Figura 39 Dispersão do número de autores por país no Top 15 países por total de autores



Fonte: dados da pesquisa (2023).

O tamanho da população pode influenciar no total de autores em uma área de pesquisa, mas existem outros fatores a serem considerados. Em 2019 EUA e China investiram 612 e 514 bilhões de dólares em pesquisa e desenvolvimento, respectivamente (OECD, 2023). Esse investimento foi acompanhado por Japão, Alemanha e Coreia do Sul que investiram mais de 100 bilhões cada no mesmo ano. Ou seja, o tamanho da população combinado com o investimento em pesquisas ao longo do tempo, ajudam a explicar a posição da maioria dos países presentes no *ranking* do número de autores com pesquisas publicadas na área de Ciência da Informação e Biblioteconomia entre 2013 e 2023.

O total de autores dos 15 países com o maior número de autores na área de Ciência da Informação foi detalhado em um diagrama de dispersão, Figura 39. Neste diagrama apenas o número de autores de mesma nacionalidade é considerado. Por exemplo, caso um artigo tenha 5 autores: 2 autores dos EUA, 2 autores da China e 1 autor da Austrália, será contabilizado 1 artigo com dois autores nos EUA, 1 artigo com 2 autores para a China e 1 artigo com 1 autor na Austrália. Quanto maior o círculo, maior o número de artigos com o mesmo número de autores.

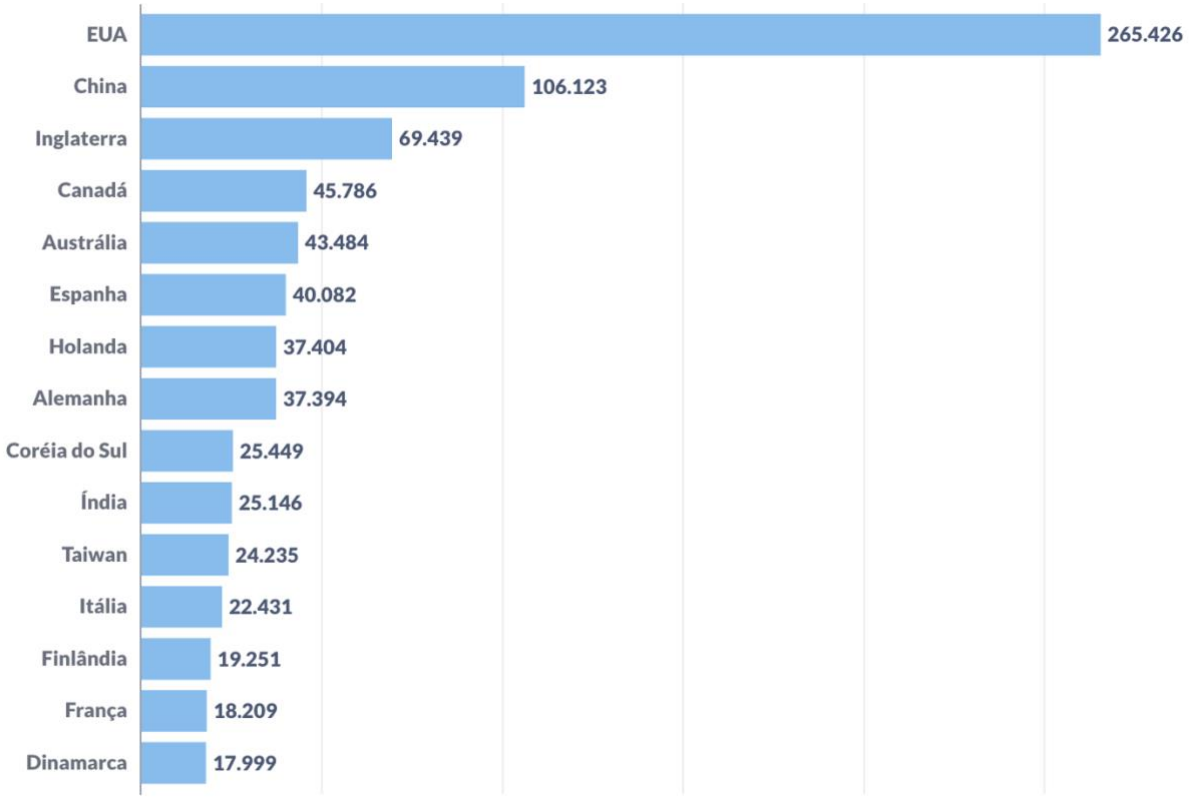
O círculo mais abaixo representa o número de artigos em cada país onde havia apenas 1 (um) autor de cada nacionalidade (Figura 39). A linha pontilhada 2 registra os círculos que representam o número de artigos com 2 autores, ou seja, os artigos na linha e acima são aqueles realizados em grupo de coautores de mesma nacionalidade.

Ao considerar USA, o primeiro país do gráfico, é possível verificar que grande parte dos artigos com autores americanos são produzidos com um único autor americano. O valor da moda deste grupo é, portanto, igual a 1. O mesmo comportamento é observado na Inglaterra, na Austrália, no Canadá, na Alemanha, na Itália, na Holanda , França, Coréia do Sul, Taiwan e Finlândia. Na China, o número de artigos com apenas 1 autor chinês é menor do que artigos com 2 ou 3 autores. Este comportamento é mais incomum, sendo visto somente na Espanha, Índia e Brasil, além da China. Pode indicar que esses países costumam trazer mais coatores de sua própria nacionalidade para colaboração.

Outro aspecto interessante que pode ser observado na Figura 39 é a dispersão do número de autores em cada país. Os EUA possuem como característica uma produção sustentada com número maior de autores americanos, chegando a 69

coautores. Os demais países possuem artigos com uma concentração abaixo de 15 autores de mesma nacionalidade por artigo.

Figura 40 - Total de Citações por Artigo Top 15 países



Fonte: dados da pesquisa (2023).

O total de citações por artigos também teve os EUA com liderança bastante vantajosa em relação aos demais países, foram 265.426 citações, quase o triplo do segundo país com mais citações, a China, que teve 106.123 citações (Figura 40 e quarta coluna da Tabela 6). A Inglaterra mais uma vez ocupou a terceira posição, com 69.493 citações no total. O Canadá ficou com a quarta posição com 45.786 citações. Canadá, Austrália e Espanha formam um grupo de países com mais de 40 mil citações cada. Holanda e Alemanha, ficam próximas com mais de 37 mil citações cada. Coreia do Sul, Índia, Taiwan e Itália obtiveram mais do que 20 mil citações cada, formando um quarto grupo de países com números de citações semelhantes. Finlândia e França aparecem mais próximas da Dinamarca, que figura pela primeira vez nesta análise dentre os 15 primeiros países, com 17.999 citações.

É esperado que países com um maior número de publicações tenham, consequentemente, um maior número de citações. Para os 3 primeiros países (EUA,

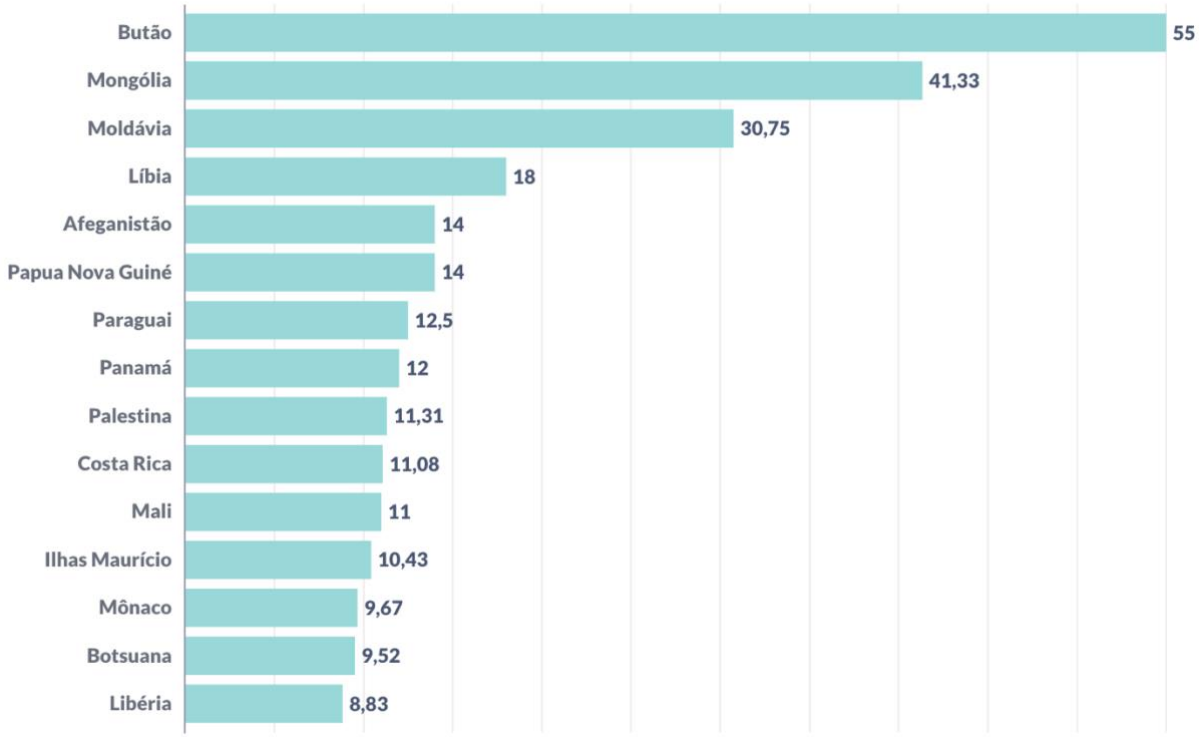
China e Inglaterra), o *ranking* se manteve igual no número de publicações. Mas o Canadá assumiu a quarta posição ultrapassando Espanha e Austrália, superiores em no número de artigos. Outro país que superou seus semelhantes em número de artigos com um maior número de citações foi a Holanda, que subiu 3 posições quando se considera o total de citações. Cabe ressaltar que a África do Sul, que figurava dentre os 15 países com maior produção, não conseguiu se manter na lista dos países com maior número de citações.

Na quinta coluna da Tabela 6, onde temos os países que mais publicaram em revistas e eventos indexados pela *Web of Science*, a média de autores por artigo nos países é em sua grande maioria um número um pouco maior que 3. As exceções são China e França, com 5,48 e 4,11 autores em média, respectivamente. Considerando que objetivo desta métrica é analisar o tamanho dos grupos envolvidos em cada artigo, a média de autores em cada um dos artigos foi obtida considerando-se a coluna "numAuthors", que armazena o número de autores para cada artigo. Não se trata, portanto, do número total de diferentes autores dividido pelo total de artigos em cada país.

A Figura 41 apresenta os 15 países com maior média de autores por artigo. Butão (*Bhutan*) com média de 55 autores por artigo é o primeiro da lista. Na base analisada constam somente 2 artigos com autores butaneses, um deles com 9 e outro com 101 autores. É importante destacar que esta média considera autores de diferentes nacionalidades, onde ao menos um deles é do país que está sendo analisado. Em seguida temos a Mongólia com média de 41,33, que além dos mesmos dois artigos compartilhados com autores butaneses possuem também um outro artigo com 14 autores.

A Moldávia é o terceiro país da lista com 30,75 autores em média, fazendo parte também do artigo com 101 autores dentre os 4 trabalhos com autores moldavos presentes na base de dados analisada. A Líbia também teve sua média de autores influenciada pela participação no artigo com 101 autores, a influência não foi tão grande quanto nos demais países por ter uma produção um pouco maior, com 7 artigos.

Figura 41 - Média de Autores por Artigo Top 15 países



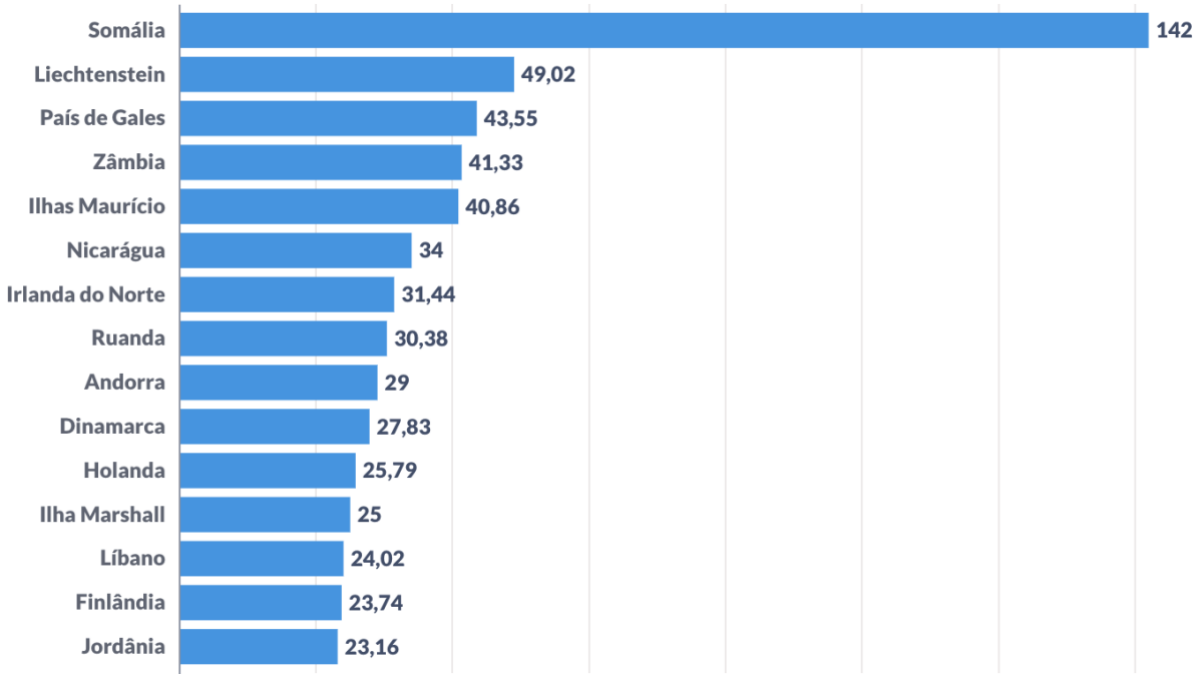
Fonte: dados da pesquisa (2023).

Já os afegãos ficaram com um média de 14 autores, referente a um único artigo, mesmo caso de Papua Nova Guiné, com apenas um artigo. Os países latino-americanos Paraguai e Panamá compõem a lista com média aproximada de 12 autores, referente a duas produções em cada país. Palestina e Costa Rica também participaram do artigo com 101 autores, mas suas médias ficaram reduzidas por possuem mais produções, 13 artigos cada.

Os demais países (Mali, Ilhas Maurício, Mônaco, Botswana e Libéria) possuem características semelhantes aos demais membros do grupo: uma produção tímida e com número de autores elevado. Os países no topo do *ranking* com maior média de autores por artigo, em geral, são países com um menor PIB per capita e que possuem uma produção científica em menor escala, participando de poucos artigos com um número elevado de autores, tendo os valores de média de autores aumentada por este motivo. Nenhum dos países com maior média de autores está presente nas demais listagens de número de citações ou número de publicações. Este fato corrobora a conclusão de que um número maior de autores não está diretamente relacionado a um maior número de citações, como já havia sido constatado na seção 6.2.

A média de citações por artigo em cada país é demonstrada na sexta coluna da Tabela 6. As médias de citações nos países que mais produziram nos últimos 10 anos na área de Biblioteconomia e Ciência da Informação variam de 7,27 (Brasil) a 25,79 (Holanda). Os EUA têm uma média de 17,36 citações por artigo e a China 16,17 citações por artigo. Quando ordenamos a lista de países pela média de citações (Figura 42), temos somente a Holanda que está entre os 15 países que mais produzem e entre os 15 países que possuem maior média de citações.

Figura 42 - Média de citações por artigo Top 15 países



Fonte: dados da pesquisa (2023).

A Somália possui somente um artigo na base de dados analisada, por isso está com seu valor de média de citações tão acima dos demais. Os demais países possuem no geral uma produção menor, com exceção da Holanda, e com bons valores de citação. O simples uso da média de citações mostra como esse valor pode ser deturpado quando se tenta sintetizar a produção acadêmica de um grupo. Grupos com menor produção, mas que participaram de artigos onde apenas um membro fazia parte, apresentaram esta métrica com valor bastante elevado. Enquanto isso, países com produção muito mais robusta não aparecem na listagem.

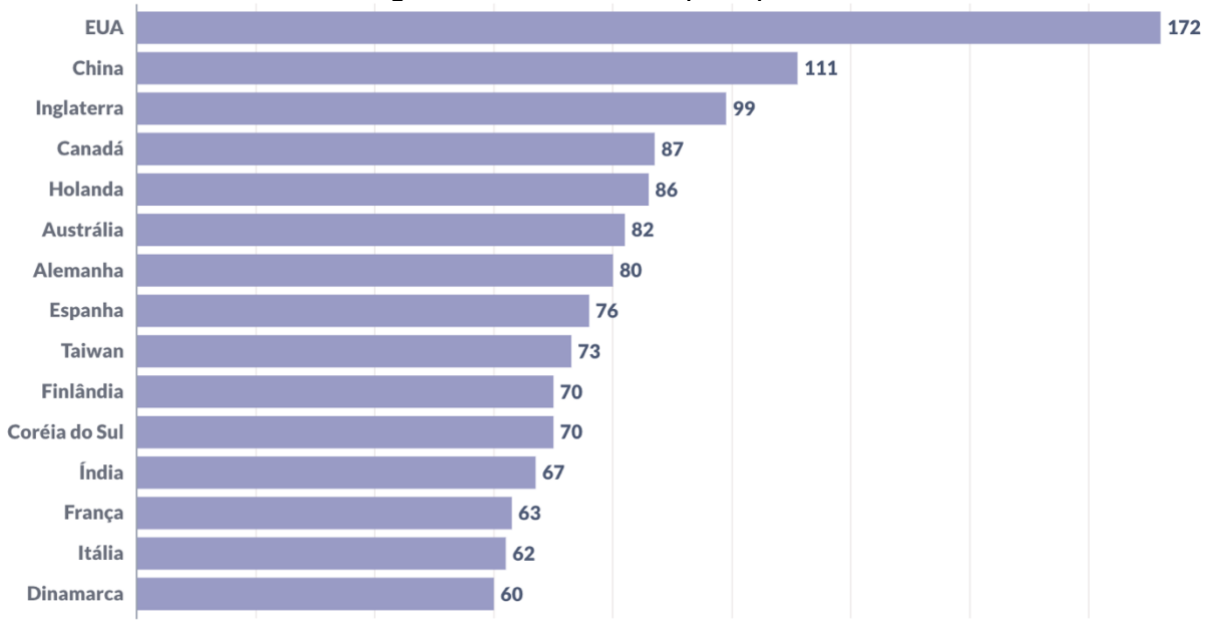
Os valores de média apresentados justificam a criação de indicadores mais consistentes como o índice h. Nem sempre a média é o valor que melhor resume a produção de um pesquisador e esta mesma ideia é valida quando estamos avaliando

grupos de pesquisadores. Na próxima seção são apresentados os valores em *ranking* dos países com maior valor de índice h com a abordagem *all Papers* e a abordagem *IN-GROUP*, aqui proposta.

6.3.1 Valor de Índice h e *IN-GROUP* por país

Os valores de média de autores e de citações apresentaram resultados bastante diferente dos valores de soma de autoria e citação. Estas diferenças ocorrem dada a natureza assimétrica dessas variáveis. Indicadores, como o índice h surgiram como uma alternativa à aplicação de médias, ao combinar o total de artigos com o número de citações (HIRSCH, 2010). No decorrer da tese apresentamos as limitações de aplicar o índice h para grupos de autores. Para fins de comparação, apresentamos o índice h calculado de maneira "*all papers*" onde todos os artigos são considerados para calcular um único indicador para o grupo (Figura 43).

Figura 43 - Índice h Top 15 países



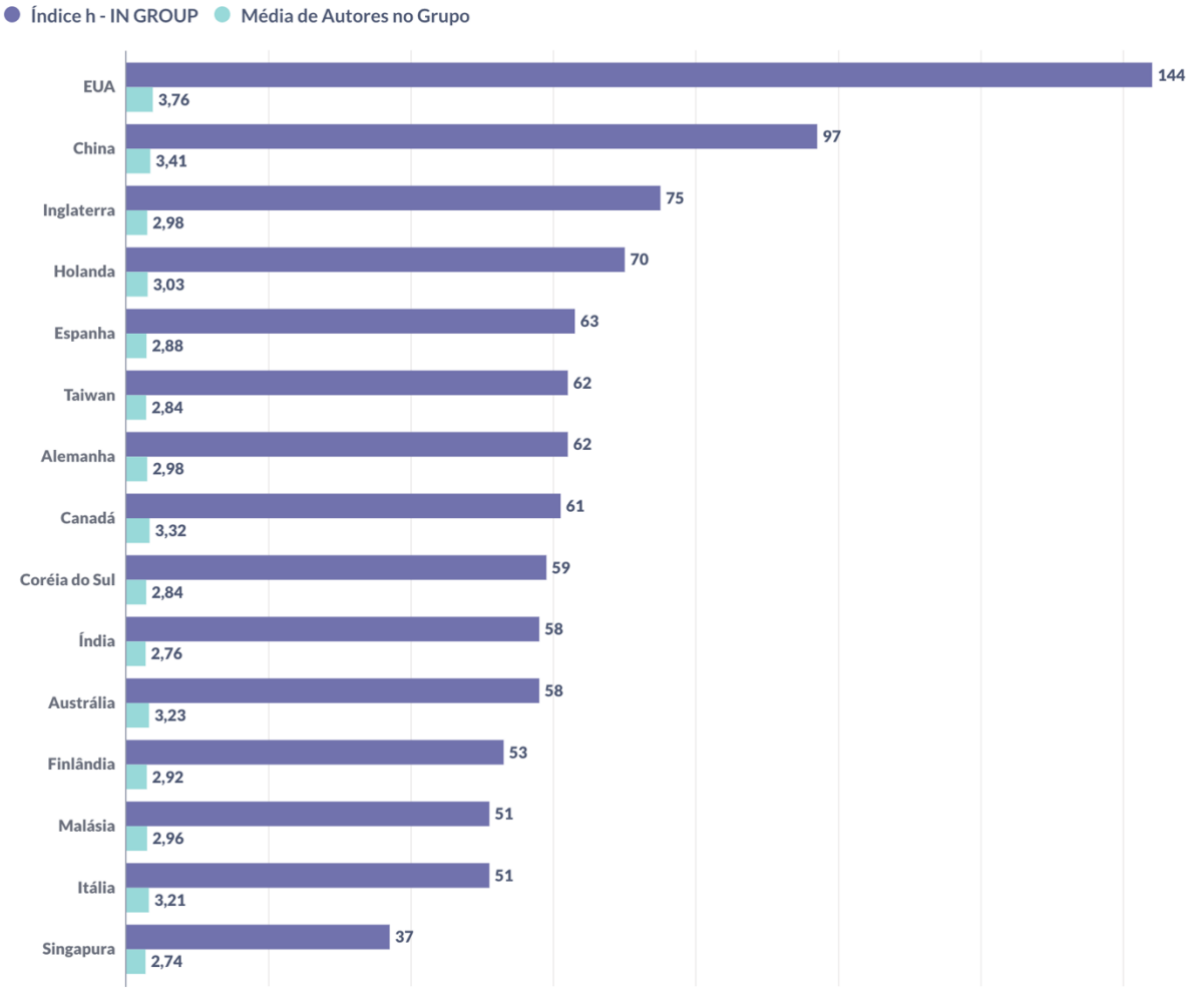
Fonte: dados da pesquisa (2023).

Considerando que os EUA possuem um maior número de artigos e de citações, seguido nessas mesmas medições pela China e Inglaterra, é esperado que o valor do índice h destes países também seja elevado. Em especial, Estados Unidos e China possuem valores distantes dos demais países. Esses países se destacaram em todas

as medidas em que foi considerado um quantitativo de soma: total de publicações, total de autores e total de citações.

O índice h quando calculado utilizando a abordagem *IN-GROUP*, varia de acordo com o Índice de Colaboração, que é o número mínimo de autores do grupo que cada artigo deve possuir para ser considerado. Como grande parte do conjunto de dados avaliado possui número de autores variando entre 2 e 6, os valores de índice h *IN-GROUP* foram calculados utilizando com esses diferentes valores de Índice de Colaboração, sendo 2, 4 e 6, respetivamente (Figura 44, Figura 45 e Figura 46).

Figura 44 - índice h *IN-GROUP* e média de Autores Top 15 países - Índice de Colaboração = 2

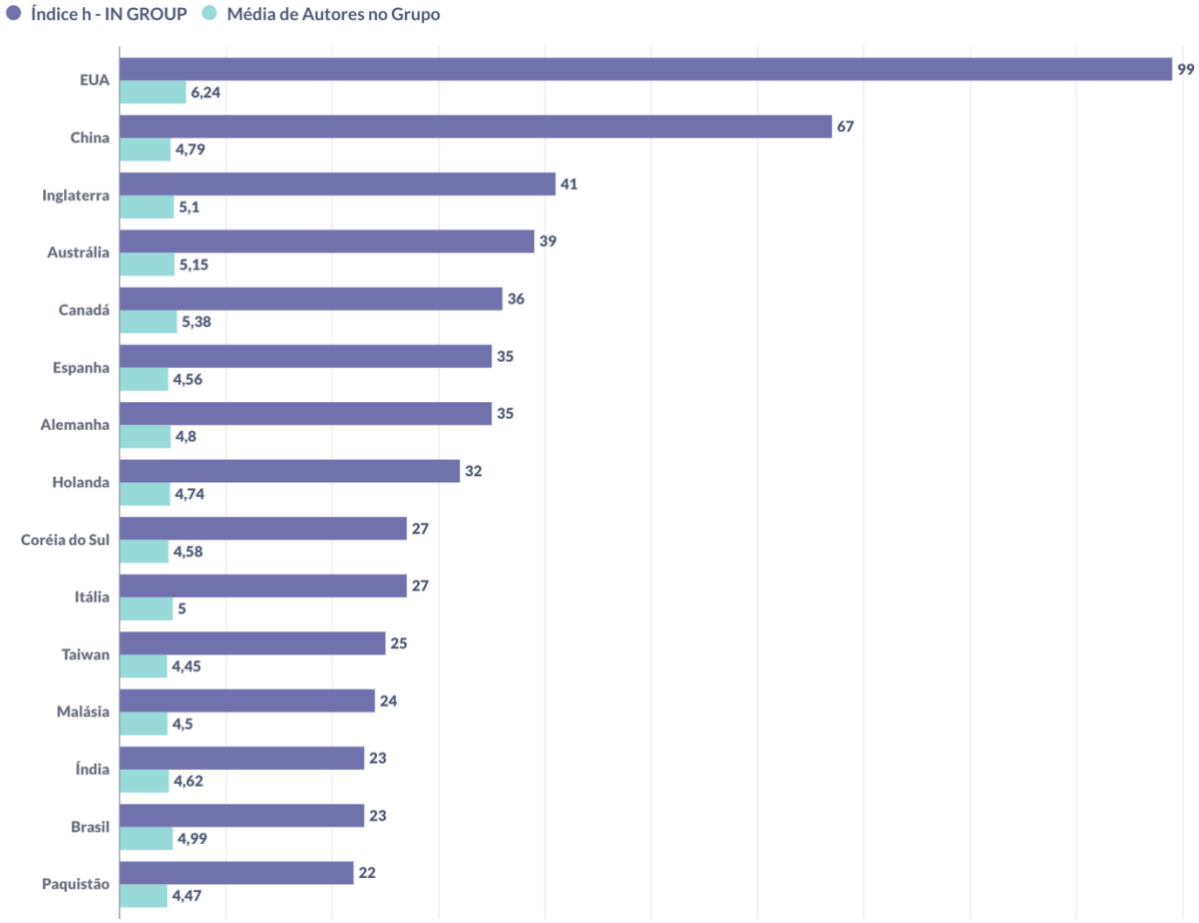


Fonte: dados da pesquisa (2023).

Este conjunto de gráficos apresenta em suas barras na cor roxa o valor do índice h calculado através da abordagem *IN-GROUP*. As barras na cor azul, representam a média de autores de mesma nacionalidade dos artigos utilizados para composição do índice h *IN-GROUP*. A média de autores pode variar à medida que o

Índice de Colaboração aumenta.

Figura 45- índice h *IN-GROUP* e Média de Autores Top 15 países – Índice de Colaboração = 4

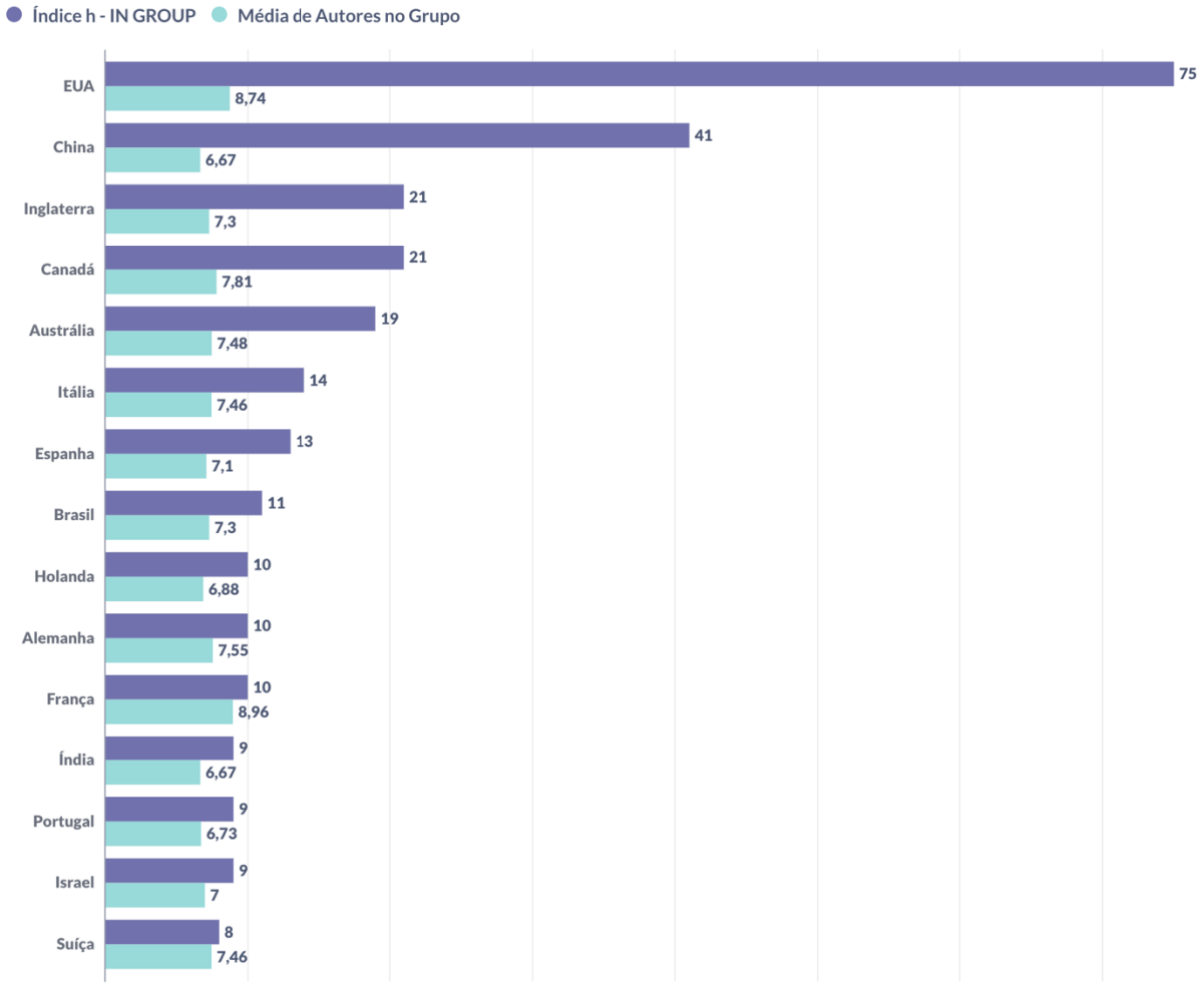


Fonte: dados da pesquisa (2023).

Quando se aumenta o grau de colaboração (Figura 44, Figura 45 e Figura 46), é possível observar que o valor do índice h tende a diminuir em todos os países observados. Estados Unidos e China permanecem em suas posições independente do valor do IC. A Inglaterra, terceiro país do *ranking* quando o Índice de Colaboração é 2 ou 4, é substituída pela Canadá quando o valor do IC é 6. Canadá, Austrália, Itália, Brasil e França sobem posições à medida que o valor do tamanho do grupo de coautores do mesmo país aumenta. Isso pode indicar que esses países tendem a publicar com coautores de mesma nacionalidade e em grupos maiores.

Enquanto Inglaterra, Holanda, Espanha, Taiwan, Alemanha, Coréia do Sul, Finlândia e Singapura vão perdendo posições à medida que o tamanho do grupo é aumentado. Isso pode indicar que a produção desses países tende a possuir menos autores por artigo, ou, ainda, pode indicar um número maior de colaborações externas.

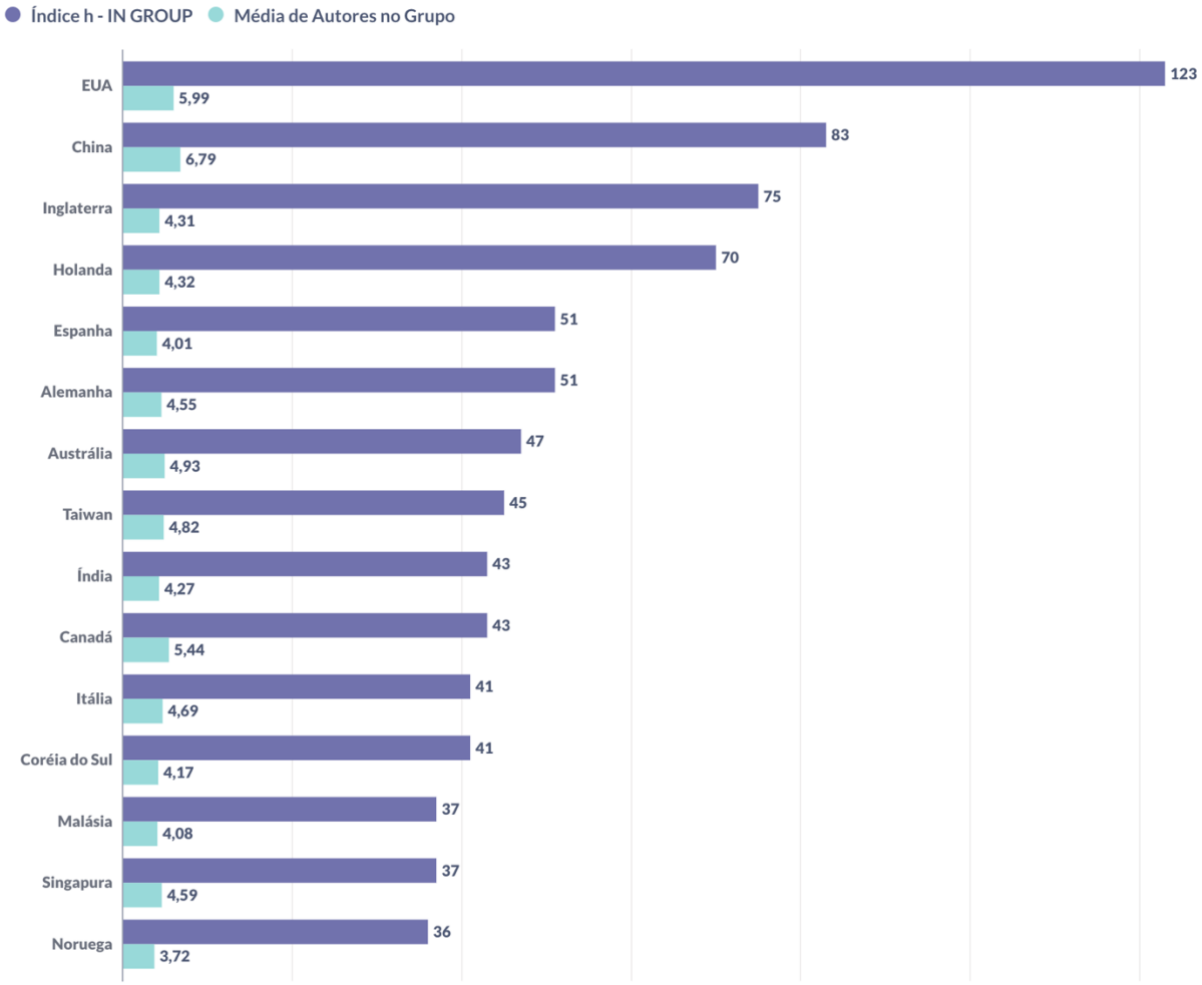
Figura 46- Índice h *IN-GROUP* e Média de Autores Top 15 países - Índice de Colaboração = 6



Fonte: dados da pesquisa (2023).

Considerando que o Grau de Colaboração pode influenciar em um *ranking* baseado no índice h com *IN-GROUP*, pode se tornar complicado definir qual o Índice de Colaboração deve ser usado. Em especial, quando os grupos analisados são tão diferentes uns dos outros. O que num país como os EUA é considerado um grupo médio pode ser diferente quando consideramos a China, que possui média de autores por artigo mais elevada. Para tentar equalizar a medida, considerando que diferentes grupos podem ter diferentes características, vamos considerar a média de autores em cada país como o Índice de Colaboração para calcular o índice h *IN-GROUP* de cada um dos países (Figura 47). Nesta configuração os EUA (USA) permanecem com valor elevado, mas um grupo formado por China, Inglaterra e Holanda formam um bloco mais coeso.

Figura 47 - índice h *IN-GROUP* e média de Autores Top 15 países - Índice de Colaboração = Média de autores do país



Fonte: dados da pesquisa (2023).

Os mesmos países que lideravam os *rankings* nos indicadores de totais de artigos, autores e citações também lideram o *ranking*, quando se afere os valores do índice h *IN-GROUP*. Ao utilizar a média de autores do mesmo grupo como Índice de Colaboração, a discrepância no grupo dos 15 primeiros países foi atenuada (Figura 47). Embora o *ranking* tenha se mantido, EUA e China acabaram ficando com valores mais próximos do que quando o valor do Índice de Colaboração era fixo. Essa diferença mais atenuada pode ocorrer por diferentes fatores, um deles é que a China é um país com média de autores maior (5,48) dentre os principais países analisados. No entanto, a diferença principal quando se utiliza a média foi uma queda no valor do índice para os EUA.

Como a média de autores por artigo dos EUA é 3,9, todos os artigos com 4 ou mais autores foram desconsiderados, e este deve ser o principal motivo do impacto

do índice. São necessárias mais avaliações com diferentes grupos para consolidarmos a sugestão de se utilizar a média como Índice de Colaboração.

Na próxima seção são apresentados os resultados dos experimentos para universidades.

6.4 AVALIAÇÃO DA PRODUÇÃO EM CIÊNCIA DA INFORMAÇÃO POR UNIVERSIDADES

Nesta nova avaliação, o mesmo conjunto de dados é analisado sob a perspectiva de grupos diferentes: Universidades. A aplicação do *F-GROUP* é apresentada no Quadro 18. Os parâmetros iniciais, detalhes de seleção dos artigos e os indiciadores definidos estão presentes.

Diferentemente do que ocorreu com relação aos países, as universidades não possuem nomenclatura padronizada. Foram necessários diversos tratamentos nos dados para que pudéssemos realizar esta análise. Os processos de normalização aplicados estão disponíveis no e APÊNDICE B – Código em SQL para Tratamento dos Dados. O objetivo desta análise é identificar as principais universidades com produção acadêmica relevante na área de Ciência da Informação nos últimos 10 anos (Quadro 18). Para isso utilizou-se o indicador índice h *IN-GROUP* considerando diferentes tamanhos de grupos de autores. São apresentadas as 15 primeiras universidades com maior valor para diferentes indicadores. Como subsídio para esta análise, são apresentados também outros indicadores como totais e média de citações, autores e artigos.

Após o tratamento dos dados, 20567 diferentes universidades/instituições diferentes foram identificadas. Para fins de simplificação, padronizamos o nome para universidades, uma vez que estas são a grande maioria de instituições identificadas na base de dados.

Quadro 18 - Aplicação do *framework F-GROUP* para análise da produção científica em Ciência da Informação em instituições entre 2013 e 2023

I) Identificação de Parâmetros Iniciais	1.Descrição do(s) grupo(s)	São 20.567 universidades. Os grupos são compostos por instituições que possuem algum artigo no campo da Ciência da Informação e Biblioteconomia na *Web of Science* no período entre abril de 2013 e abril de 2023 com uma ou mais citações.
	2.Objetivos da avaliação	Comparação entre instituições. Identificar aquelas que possuem produção mais relevante na área de Ciência da Informação e Biblioteconomia ao longo dos últimos 10 anos.
	3.Fontes de Dados	*Web Of Science*
II) Seleção dos Artigos	4.Formas de contagem e agregação	*IN-GROUP* e *all papers*
	5.Aspectos Representativos	Janela temporal: 10 anos (10 de abril de 2013 a 10 de abril de 2023) Número de autores: 2 ou mais Grau de Colaboração Interno: 2, 3, média e mediana de autores por instituição. Autocitação: não avaliado Adicional: autores com mais de uma afiliação serão considerados nos indicadores de todas as instituições vinculadas.
III) Definição de Indicadores	6.Listagem de Indicadores	índice h com abordagem *IN-GROUP*, índice h com abordagem total de artigos (*all papers*), total de citações, total de autores, média de autores por artigo, média de citações por artigo.

Fonte: elaborado pela autora

Na Tabela 7 temos uma lista com as 15 primeiras universidades ordenadas por Total de Artigos. Os nomes das universidades foram trazidos conforme constam na base da *Web Of Science*. A primeira Universidade é de Wuhan, na China. A presença massiva de Universidades dos EUA e China não surpreende, pois, esses países lideram de forma absoluta a produção quando analisamos a produção de artigos por país em Ciência da Informação e Biblioteconomia nos últimos dez anos. Existem 4 Universidades que se destacam das demais no total de artigos: Wuhan Univ (China),

Univ Texas (EUA), Univ Hong Kong (China), Univ Calif (EUA), possuem mais de 700 artigos cada, Figura 48.

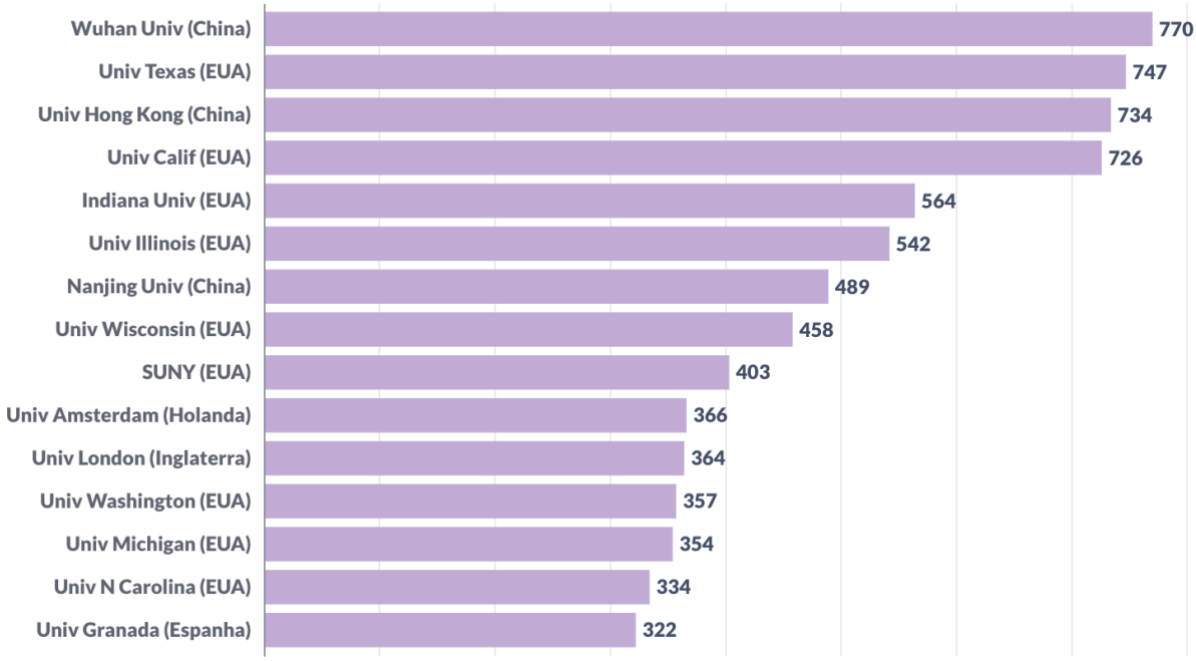
Tabela 7 - Indicadores de Produção científica por Universidade (Top 15)

Universidade	País	Total de Artigos	Total de Autores	Total de Citações	Média de autores por artigo	Média de Autores da Univ. por artigo	Média de Citações por Artigo
Wuhan Univ	China	770	1890	11138	6,14	2,29	14,55
Univ Texas	EUA	747	2608	16948	5,68	1,6	23,29
Univ Hong Kong	China	734	1843	17121	5,16	1,53	23,72
Univ Calif	EUA	726	2986	13311	5,31	1,87	20,04
Indiana Univ	EUA	564	1679	9545	5,16	1,7	17,65
Univ Illinois	EUA	542	1499	7082	4,24	1,63	14,36
Nanjing Univ	China	489	1278	5588	5,92	2,05	11,65
Univ Wisconsin	EUA	458	1496	9312	5,24	1,69	21,61
SUNY	EUA	403	1159	7696	4,56	1,45	20,22
Univ Amsterdam	Holanda	366	899	8734	3,99	1,57	24,45
Univ London	Inglaterra	364	1065	6567	4,36	1,11	19,6
Univ Washington	EUA	357	1605	5646	6	1,74	16,57
Univ Michigan	EUA	354	1530	7593	5,56	1,75	22,03
Univ N Carolina	EUA	334	1295	5685	5,01	1,68	21,9
Univ Granada	Espanha	322	545	6971	3,73	1,92	22,22

Fonte: dados da pesquisa (2023).

EUA e China se destacam dentre os demais países em termos de produção, e, ao analisar as Universidades, esse destaque se mantém. A Figura 48 apresenta as primeiras 15 universidades, considerando o total de artigos. A lista é composta por 9 universidades americanas, 3 chinesas, 1 holandesa, 1 inglesa e 1 espanhola.

Figura 48 - Total de artigos por Universidade Top 15 universidades



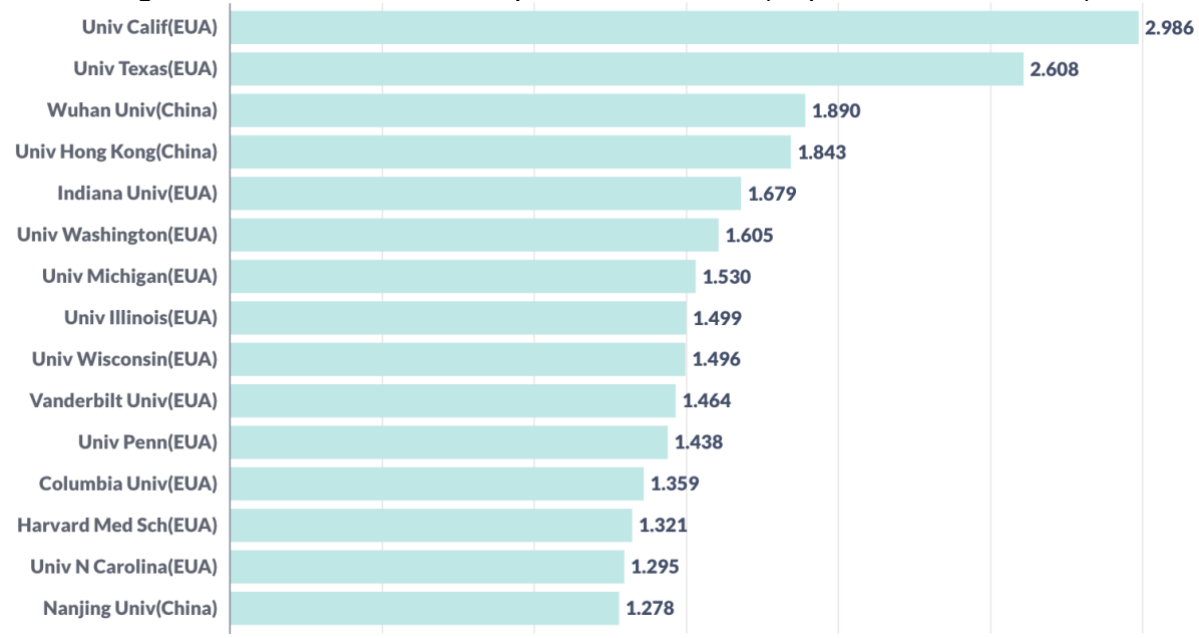
Fonte: dados da pesquisa (2023).

Quando se verifica o total de Autores (quarta coluna da Tabela 7), com mais detalhes na Figura 49, observa-se uma liderança americana acentuada, com as 2 primeiras universidades desse país (e um total de 12) entre o top 15 de universidades com o maior número de autores. O restante deste *ranking* é composto por 3 universidade chinesas, segundo país com o maior número de autores.

A Universidade da Califórnia é a primeira colocada no *ranking*, com 2.986 autores, seguida pela Universidade do Texas, com 2.608 autores. Em seguida, em um patamar consideravelmente menor, temos duas universidades chinesas: Wuhan e Hong Kong, com pouco mais de 1.800 autores cada. As demais universidades americanas aparecem em seguida na ordem: Indiana (1.679), Washington (1.605), Michigan (1.530), Illinois (1.499), Wisconsin (1.496), Vanderbilt (1.464), Pensilvânia (1.438), Columbia (1.359), Harvard (1.321) e Carolina do Norte (1.295). A última universidade a compor o top 15 é a chinesa Nanjing, com 1.278 autores.

A predominância americana no número de autores por universidade era esperada, dado o número acentuado de autores americanos identificados na análise por país. Os EUA possuem mais de 36 mil autores na base de dados analisada; o segundo colocado, a China, não chega a 14 mil autores.

Figura 49 - Total de autores por universidade (Top 15 Universidades)

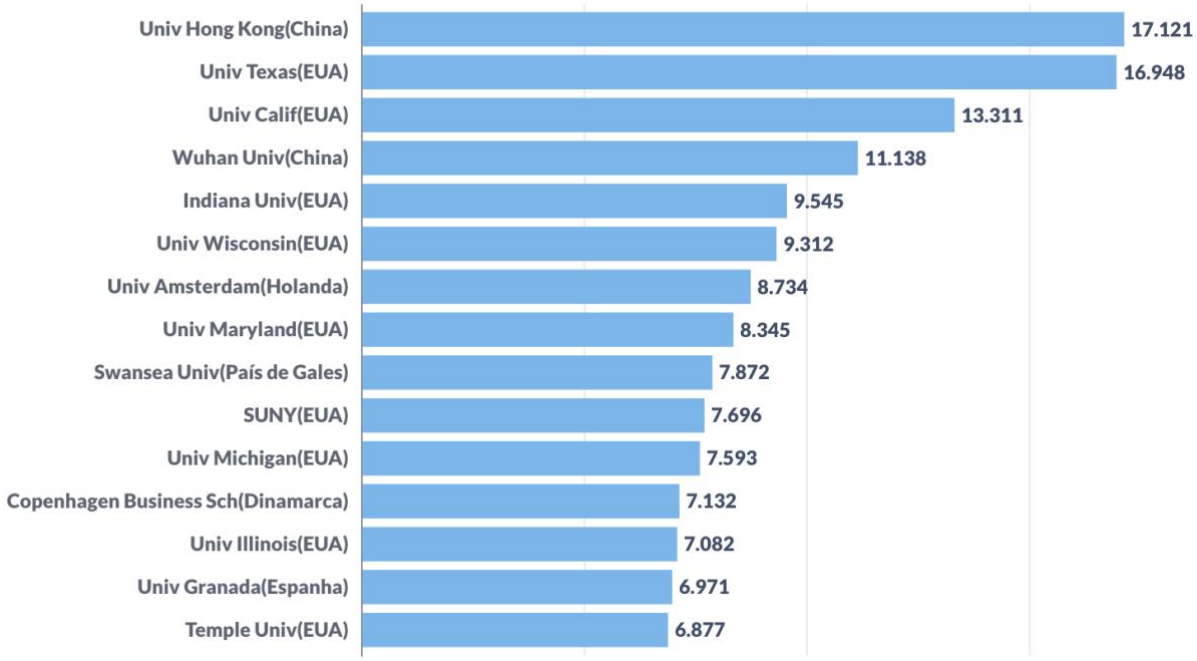


Fonte: dados da pesquisa (2023).

O total de citações por universidade apresenta a Universidade de Hong Kong na liderança do *ranking*, com 17.121 citações. Em seguida, com número de citações bastante expressivo, temos a Universidade do Texas, com 16.948 citações. Compõem este *ranking* de total de citações 8 instituições americanas, 3 chinesas, além de uma universidade na Holanda, País de Gales, Dinamarca e Espanha (uma de cada país).

É esperado que universidades com um maior número de publicações tenham, consequentemente, um maior número de citações. No entanto, o *ranking* de Citações se apresentou com diferenças significativas em relação à Figura 48 (*Ranking* de Total de Artigos). A Universidade da Califórnia teve o maior número de artigos, mas ocupou somente a terceira posição no *ranking* de citações. A Universidade do Illinois caiu 5 posições quando comparada ao *ranking* de produção de artigos. As universidades de Swansea e de Granada entraram na listagem de citações, mesmo não compondo o *ranking* das principais universidades quanto a produção de artigos.

Figura 50 - Total de Citações Top 15 Universidades

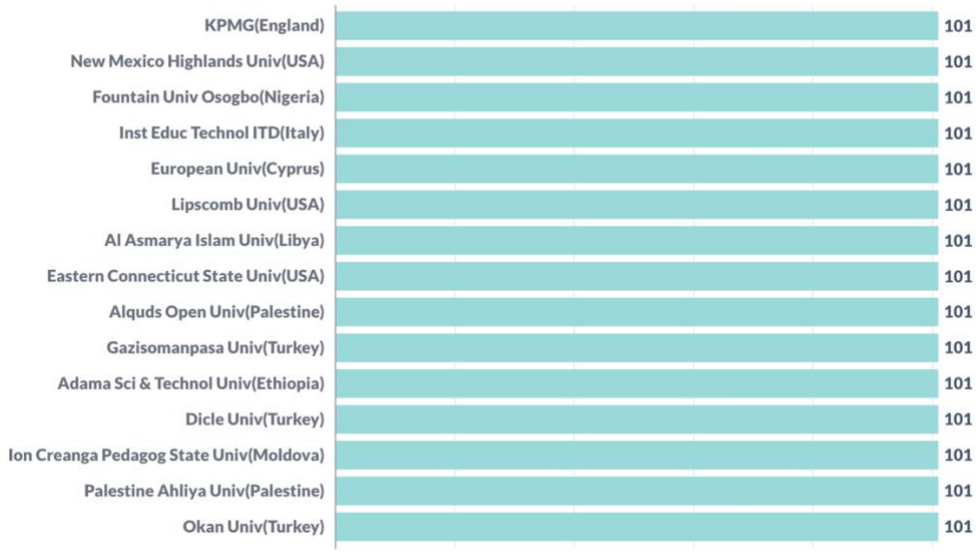


Fonte: dados da pesquisa (2023).

Na sexta coluna Tabela 7, onde temos as Universidades que mais publicaram em revistas e eventos indexados pela *Web of Science* na Área de Ciência da Informação e Biblioteconomia nos últimos 10 anos, o valor da média de autores por artigo nas Universidades apresentou uma variação maior do que quando comparamos com os grupos de países. No *ranking* por universidades, os valores ficaram um pouco maiores, variando entre 3,73 e 6,14 autores por grupo, com maioria das universidades apresentando média acima de 4 autores.

A média de autores em cada um dos artigos foi obtida considerando a coluna “numAuthors”, que armazena o número de autores para cada artigo. Não se trata, portanto, do número total de diferentes autores dividido pelo total de artigos em cada país. O objetivo, mais uma vez, é analisar o tamanho dos grupos envolvidos em cada artigo. Na Figura 51 é possível visualizar as 15 universidades com maior média de autores por artigo. Este gráfico ficou dominado por um único artigo que continha 101 autores. Por ser o único artigo na base de dados de algumas universidades a média foi o resultado deste único registro.

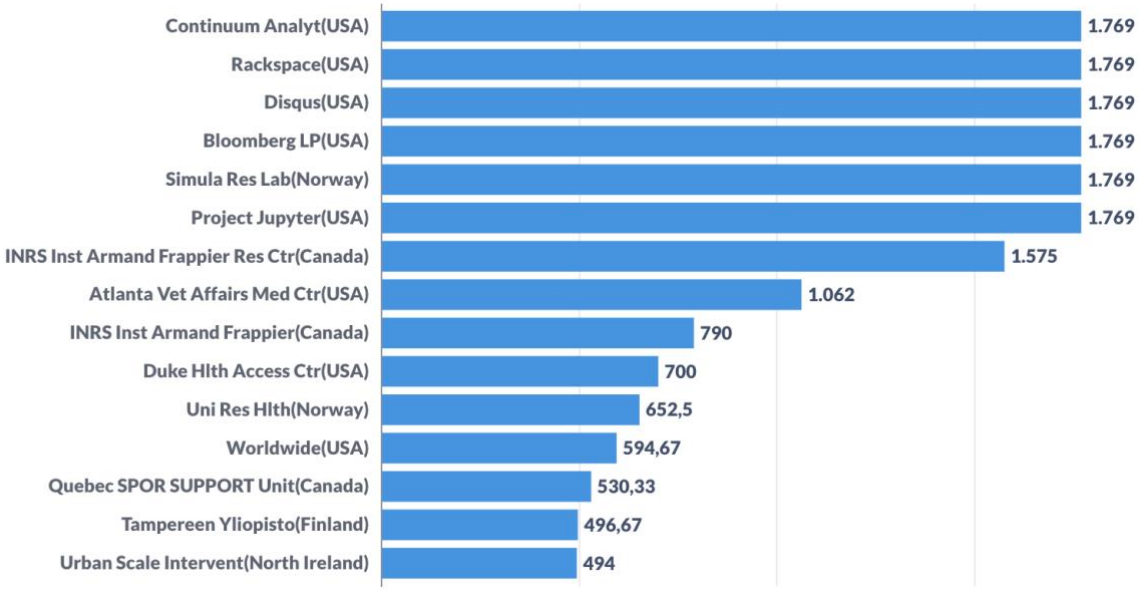
Figura 51 - Média de Autores por Artigo Top 15 Universidades



Fonte: dados da pesquisa (2023).

A média de citações por artigo em cada universidade é demonstrada na oitava coluna da Tabela 7. Ainda na Tabela 7, as médias de citações nos países que mais produzem variam de 11,65 (Nanjing Univ) a 24,45 (Univ Amsterdam). Ao ordenar pela média de citações, obtém-se um conjunto de 6 instituições com o mesmo número de citações, 1769, referentes ao mesmo artigo (Figura 52). As principais Universidades em termos de produção e número de autores não figuram nesta lista. Mais uma vez a média se prova como um indicador não muito interessante na análise da produção de grupos de autores.

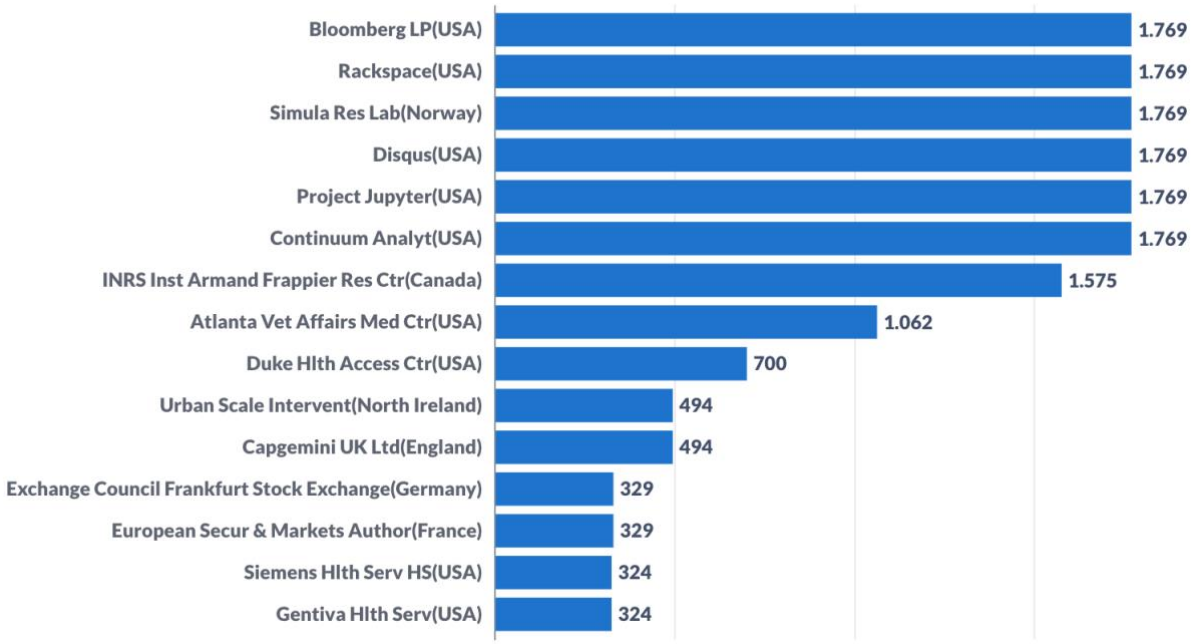
Figura 52 - Média de citações por artigo (Top 15 Universidades)



Fonte: dados da pesquisa (2023).

Considerando que o valor da média não é o que melhor representa um conjunto de dados que não possui uma curva normal, apresentamos na Figura 53 o *ranking* de acordo com a mediana. Pode-se verificar que este gráfico apresenta os mesmos valores da Figura 52, por causa da mesma situação: instituições com pouca produção, mas com um valor de citações mediano elevado.

Figura 53 - Mediana de citações por artigo (Top 15 Universidades)



Fonte: dados da pesquisa (2023).

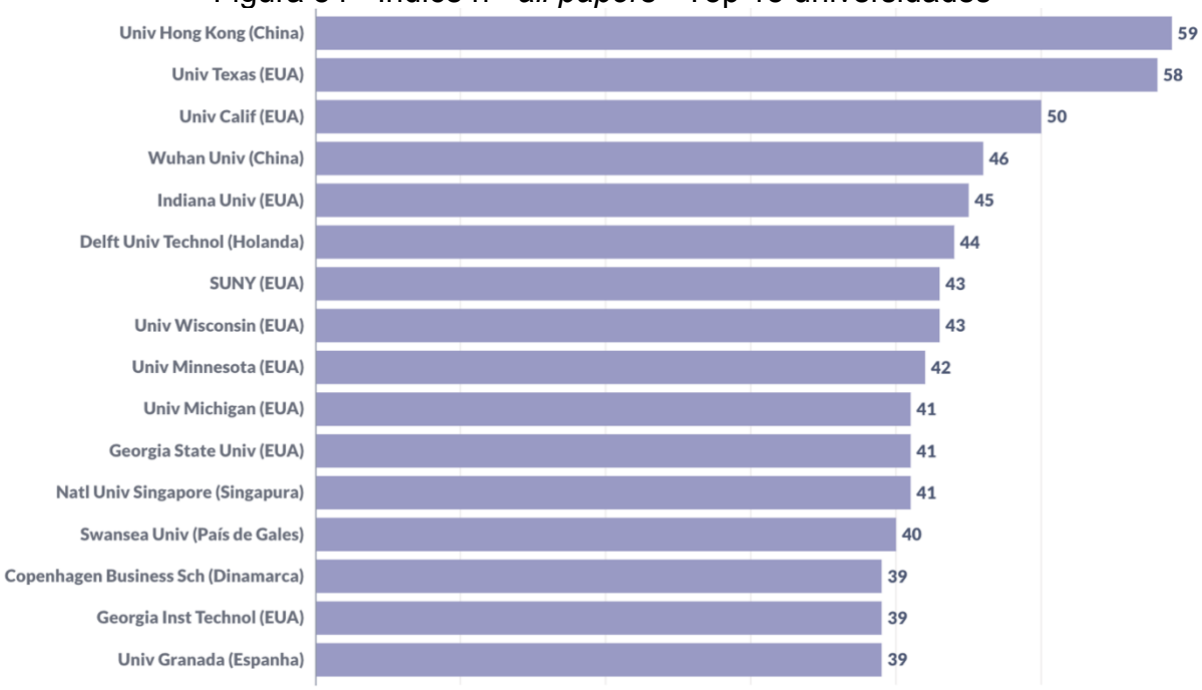
Os gráficos das Figura 52 e Figura 53 apresentam não somente universidades, mas, também, outras instituições. Como a base de dados não possui a informação Universidade normalizada, outras instituições também são passíveis de aparecer nas análises. No entanto, quando observamos a produção científica e número de citações, apenas universidades aparecem entre as primeiras instituições. Isso pode demonstrar uma predominância na publicação de artigos em universidades, nos últimos 10 anos, nas áreas de Ciência da Informação e Biblioteconomia.

6.4.1 Valor de Índice h - *all papers* e *IN-GROUP* por Universidade

Mais uma vez observamos que os valores de média podem não representar muito bem os grupos analisados. As universidades que compõem o *ranking* top 15 de

média de citação (Figura 53) são bem diferentes das que somam o maior número de citações (Figura 50). Como sabemos, o índice h, é utilizado para corrigir as distorções no uso da média. O valor do índice h calculado de maneira "*all papers*" é apresentado na Figura 54. Nesta representação, todos os artigos são considerados para calcular o indicador do grupo, independentemente do número de autores no artigo.

Figura 54 - Índice h - *all papers* - Top 15 universidades



Fonte: dados da pesquisa (2023).

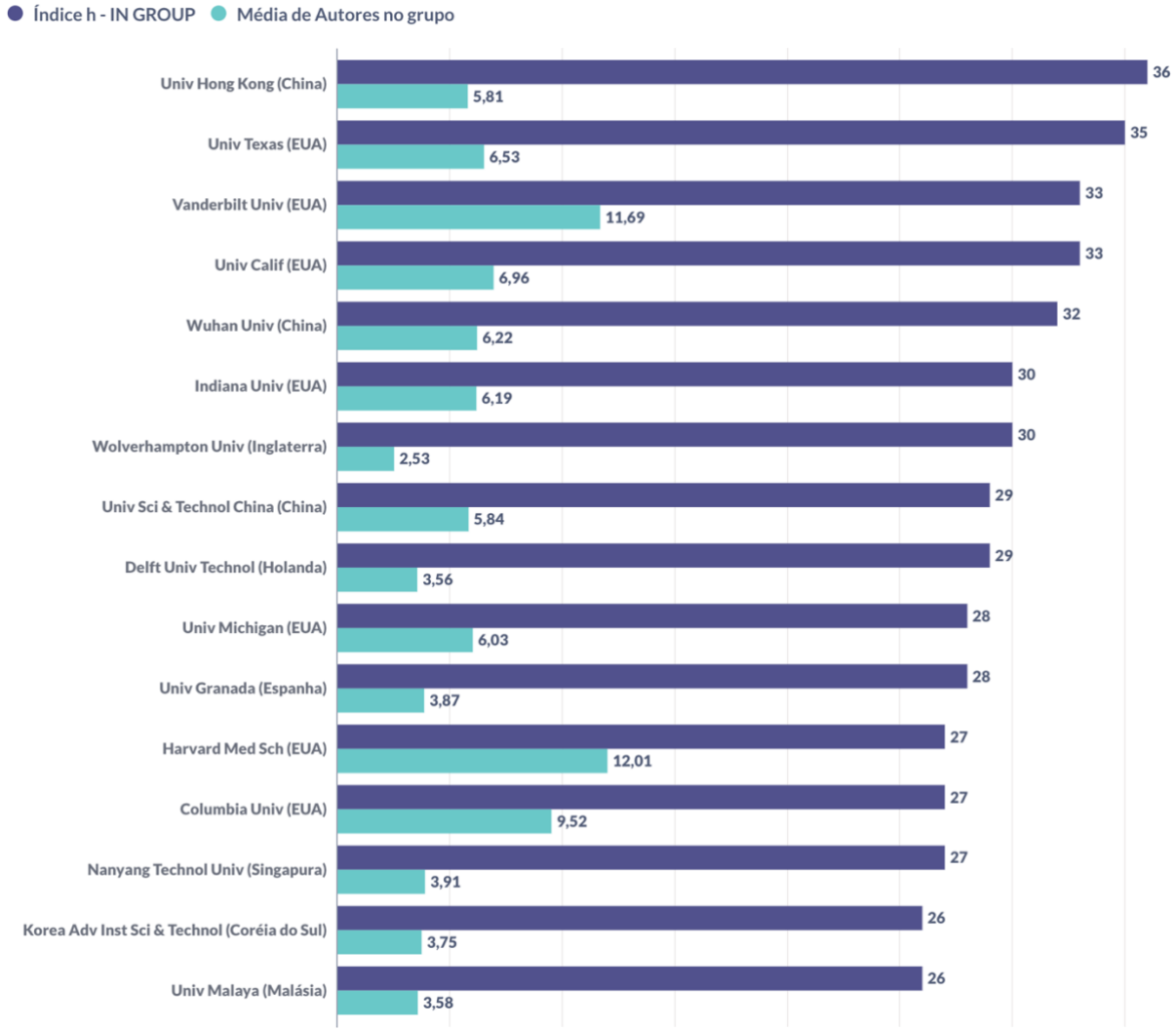
O *ranking* top 15 de índice h em universidades é mais uma vez predominantemente composto por instituições da China e EUA. A lista possui 16 universidades, uma vez que houve um empate da décima quinta posição. As universidades de Hong Kong e do Texas são as que lideram o *ranking* com índice h de 59 e 58, respectivamente. São, ao todo, 9 universidades americanas no topo quando o indicador utilizado é o índice h com agregação *all papers*. Duas universidades chinesas compõem o *ranking*, que ainda possui a universidade de Delft na Holanda (44), Universidade Nacional de Singapura (41), Swansea do país de Gales (40) e Universidade de Granada na Espanha (39).

Diferente do que ocorreu entre países, ao se analisar somente as universidades, não se observa uma discrepância nos valores de índice h tão acentuados. No *ranking* das primeiras 15 universidades, o maior valor foi 59 e o menor 39. Dentre os países, a diferença entre o primeiro país (USA - 172) e o décimo quinto

(Malásia - 60) foi de 112. No entanto, quando se analisa os países nos quais as universidades estão localizadas, mais uma vez é observada uma predominância americana.

O índice h quando calculado por meio da abordagem *IN-GROUP*, varia de acordo com o Índice de Colaboração, que é o número mínimo de autores do grupo que cada artigo deve possuir para ser considerado. A média de autores na mesma universidade apresentou valores bastante baixos, com cerca de 2 autores (Tabela 7). Apresentamos nas Figura 55, Figura 56 e 57 os valores de índice h *IN-GROUP* com diferentes valores de Índice de Colaboração, sendo de 2 a 4 respectivamente.

Figura 55 - índice h *IN-GROUP* e média de Autores Top 16 universidades - Índice de Colaboração = 2

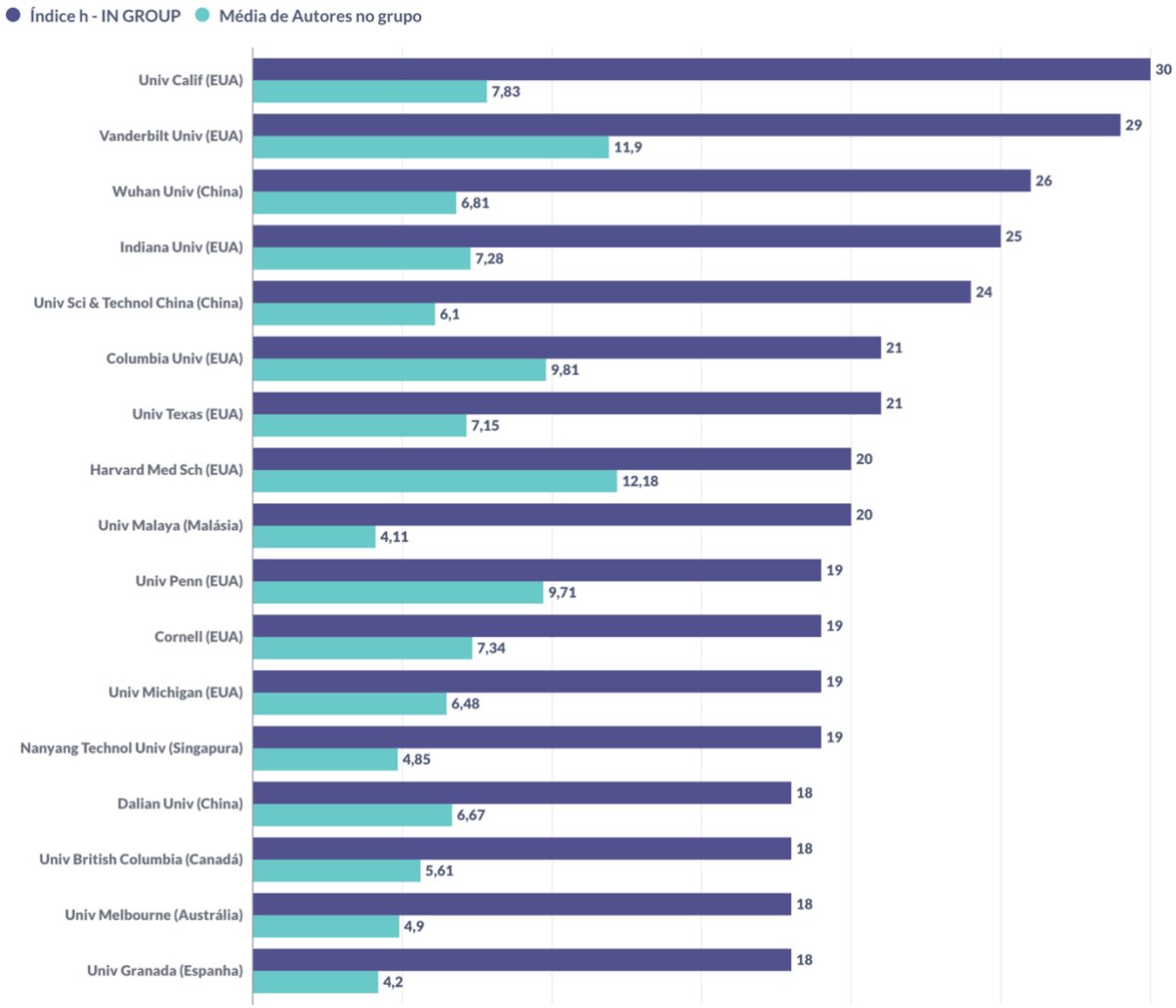


Fonte: dados da pesquisa (2023).

Este conjunto de gráficos apresenta em suas barras na cor violeta o valor do índice h calculado através da abordagem *IN-GROUP*. As barras de cor azul

representam a média do número de autores nos artigos analisados, independente da instituição. A média de autores pode variar à medida que o Índice de Colaboração aumenta, uma vez que quanto menor o valor do índice h, menos artigos são utilizados para compor o índice. Neste conjunto de dados, por haver empates nas últimas posições, os gráficos são trazidos com mais de 15 universidades cada.

Figura 56 - índice h *IN-GROUP* e média de Autores Top 15 universidades - Índice de Colaboração = 3

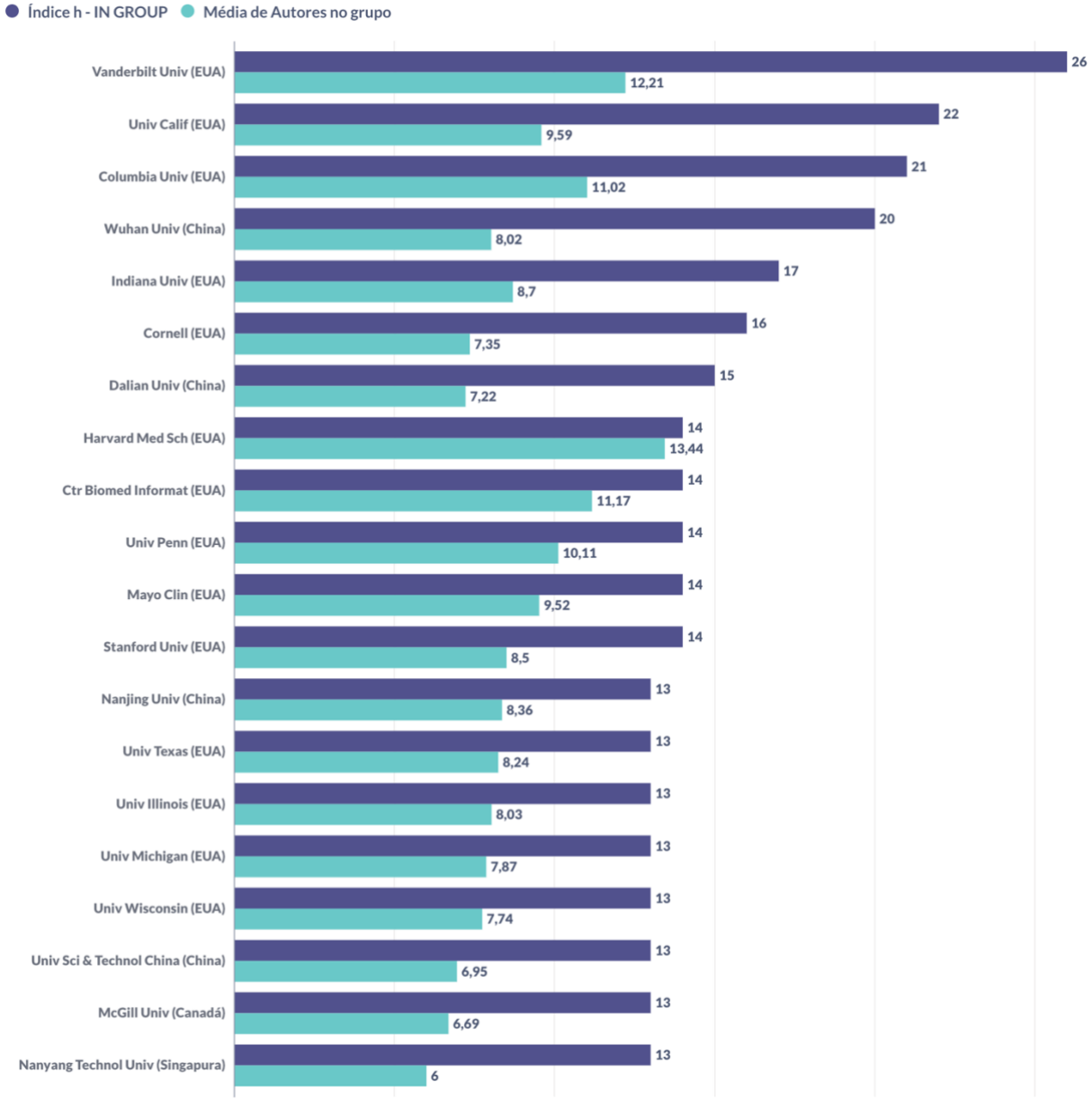


Fonte: dados da pesquisa (2023).

Ao aumentar o grau de colaboração (Figura 55, 56 e 57) é possível observar que o valor do índice h tende a diminuir em todas as universidades. A universidade de Hong Kong, primeira no *ranking* quando o Índice de Colaboração é 2 (Figura 55), não aparece nos demais *rankings*. Isso ocorre, pois, a maioria dos artigos com maior citação possuem somente 2 autores desta universidade. Quando o Grau de

colaboração passa a ser 3, o índice h *IN-GROUP* da Universidade de Hong Kong passa a ser 17. Comportamento semelhante ocorre com as universidades de Wolverhampton da Inglaterra, Delft da Holanda, Granada na Espanha e no Instituto Avançado Coreano de Ciência e Tecnologia, com destaque para a baixa média de coautores dessas universidades, variando entre 2,53 e 3,91 (Figura 55).

Figura 57 - índice h *IN-GROUP* e média de Autores Top 15 universidades - Índice de Colaboração = 4



Fonte: dados da pesquisa (2023).

Quando se analisa as Universidade do Texas e de Michigan ao longo dos diferentes valores de Índice de Colaboração, observamos que o valor do índice h diminui, bem com as posições no *ranking*. Essas universidades permanecem com

produção expressiva, estando presente nos *rankings* quando usamos o Índice de Colaboração entre 2 e 4. O número de autores médio por artigo não sofre grande variações nessas universidades, se mantendo entre 6 e 8 autores de acordo com os diferentes valores de IC.

A universidade de Vanderbilt permanece no top 3 para todos os valores de Índice de Colaboração. Sua posição no *ranking* sobe à medida que o valor do Índice de Colaboração aumenta; concomitantemente, o valor de seu índice h diminui para 33, 29 e 26, respectivamente. Esta universidade era a última do *ranking* quando o valor do índice h não considerava a produção de ao menos dois autores do mesmo grupo, com índice h de 38. O número médio de autores por artigo nessa universidade é expressivo, variando entre 11,69 e 12,21. Ou seja, os artigos mais citados da universidade de Vanderbilt tendem a ter um número grande de autores, e isso inclui autores da própria universidade.

A universidade da California apresenta comportamento semelhante, subindo no *ranking* quando o IC = 3 e caindo quando o IC = 4. O índice h varia de 33, 30 e 22 conforme o valor do IC varia de 2 a 4. Além disso, o tamanho total do grupo dos artigos onde a Universidade da California possuiu colaboração varia entre 6,96 e 9,59, sendo o maior número de autores total com a colaboração mínima de 4 autores da universidade. O índice h desta universidade com a forma de agregação *all papers* é de 50. Podemos inferir que a Universidade da California possui uma colaboração interna expressiva e consistente, pois mesmo com a variação do Índice de Colaboração interno, manteve valores expressivos de índice h no grupo. Um comportamento bastante similar ocorre nas universidades de Wuhan (China) e Indiana (EUA).

Quando o Índice de Colaboração muda de 2 para 3, a Universidade de Ciência e Tecnologia da China sobe da nona para a quinta posição no *ranking* e seu índice h cai de 29 para 24. No entanto, ao subir a colaboração para 4 pesquisadores da mesma universidade, seu índice h despenca para 13, mesmo número de sua posição no *ranking*, o que configura um “empate” com mais 7 universidades. O número médio de autores por artigo pouco varia, com um valor próximo de 6 autores por artigo nos artigos mais citados da universidade de Ciência e Tecnologia da China. É provável que os principais artigos dessa universidade mantenham um número de autores próximo a 6, mas que, no geral, o número de autores da Universidade de Ciência e Tecnologia da China seja mais baixo, entre 1 e 3 autores.

Nanyang, universidade em Singapura, se manteve na décima segunda posição com IC 2 e 3, caindo para a décima terceira posição no *ranking* com IC = 4. Foi a universidade com posição mais estável no *ranking*. Seus valores de índice h variaram entre 27 e 13 (Figura 55, 56 e 57). O número médio de autores por artigo subiu de 3,91 (para um IC = 2) para 4,85 (para um IC = 3) e 6 (quando IC = 4). Isso demonstra que entre os autores dos principais artigos da Universidade Tecnológica de Nanyang costuma haver entre 50% e 60% de integrantes desta universidade.

Dentre as universidades que são beneficiadas por um número maior de Índice de Colaboração interno, temos a americana Columbia. Esta Universidade possui uma média de autores por artigo elevada (Figura 55, 56 e 57), variando entre 9,52 e 11,02. Esse número de autores, combinado com uma participação de autores da própria universidade, contribui para que o valor do índice h não diminua à medida que o IC aumenta, neste caso o valor do índice h variou entre 27 e 21. A Universidade Harvard apresenta comportamento semelhante, com um número bastante elevado de coautores nos seus principais artigos, variando entre 12,01 e 13,44: a maior média entre as top 15 universidades.

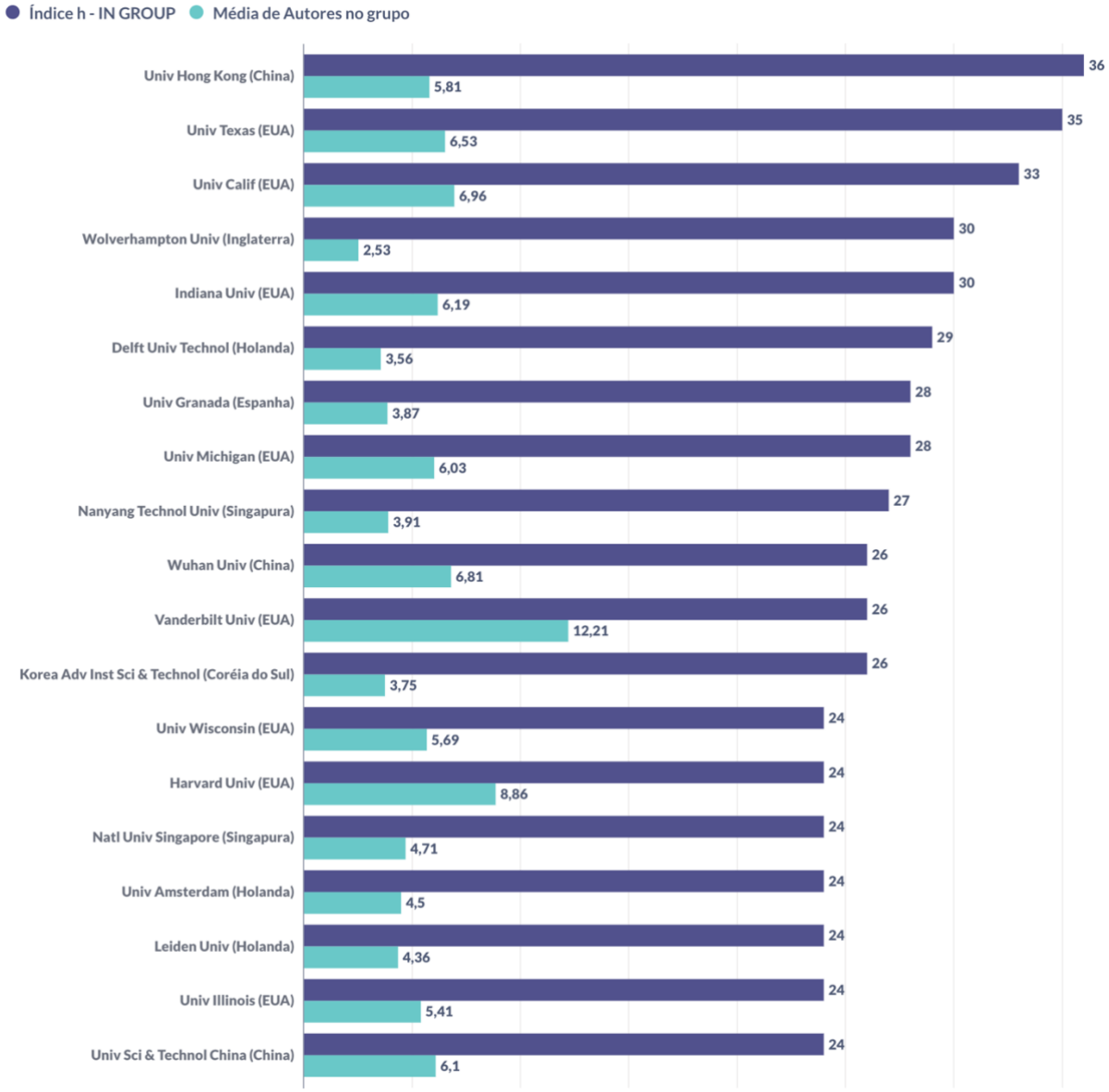
A universidade da Malásia, última do *ranking* quando o IC = 2 (Figura 55), chegou a subir de posições quando se aumenta o Índice de Colaboração para 3; entretanto, não conseguiu se manter no *ranking* quando o IC passou para 4. Isso ocorreu porque essa universidade possui como característica um baixo número de autores médio por artigo.

As universidades do Minesota (EUA), SUNY (EUA), Wisconsin (EUA), Georgia (EUA), Illinois (EUA), Swansea (País de Gales) e Copenhagen (Dinamarca) estavam presentes no *ranking* quando o índice h foi calculado sem considerar o número de autores de cada universidade (Figura 54). É provável que as principais contribuições dessas universidades sejam compostas por artigos onde poucos autores dessas instituições estavam presentes. Considerando que o objetivo dessa avaliação é identificar os grupos que melhor atuam entre seus pares, essas universidades apresentaram valores insuficientes nesse quesito.

Considerando que o Grau de Colaboração pode influenciar em um *ranking* baseado no índice h com *IN-GROUP*, uma tarefa não trivial passa a ser definir qual o Índice de Colaboração deve ser usado. Em especial, quando os grupos analisados são tão diferentes uns dos outros. O que é considerado um grupo médio em universidades como Columbia e Harvard pode ser diferente quando estamos

comparando com as universidades de Granada ou Hong Kong. Para tentar equalizar a medida, considerando que diferentes grupos podem ter diferentes características, vamos utilizar a média e a mediana de autores em cada universidade como o Índice de Colaboração para calcular o índice h com *IN-GROUP* de cada uma das universidades (Figura 58 e 59).

Figura 58 - índice h com *IN-GROUP* e média de Autores (Top 15 universidades - Índice de Colaboração = Média de autores da Universidade)



Fonte: dados da pesquisa (2023).

O valor utilizado como parâmetro para o gráfico da Figura 58 é o Índice de Colaboração baseado na média de autores nos artigos de cada universidade. Este valor foi utilizado para equalizar o tamanho dos grupos de acordo com as

características próprias de cada instituição. Desta forma, a Universidade de Hong Kong volta ao topo, seguida por Texas e California, assim como já ocorria o *ranking* do índice h que não considera o envolvimento do grupo. A Universidade Wuhan, a que apresenta o maior número de publicações, perde várias posições em relação ao cenário de índice h *all papers,* o que pode indicar que muitos de seus principais artigos possuem poucos autores da mesma universidade.

As universidades que perdiam muitas posições à medida que o valor do Índice de Colaboração aumentava, voltam ao *ranking* quando utilizamos as médias. São exemplos dessa situação as universidades de Wolverhampton, Delft, Granada e o Instituto Avançado Coreano de Ciência e Tecnologia. Essas universidades possuem médias de autores por artigo menores, mas a colaboração interna entre elas é significativa.

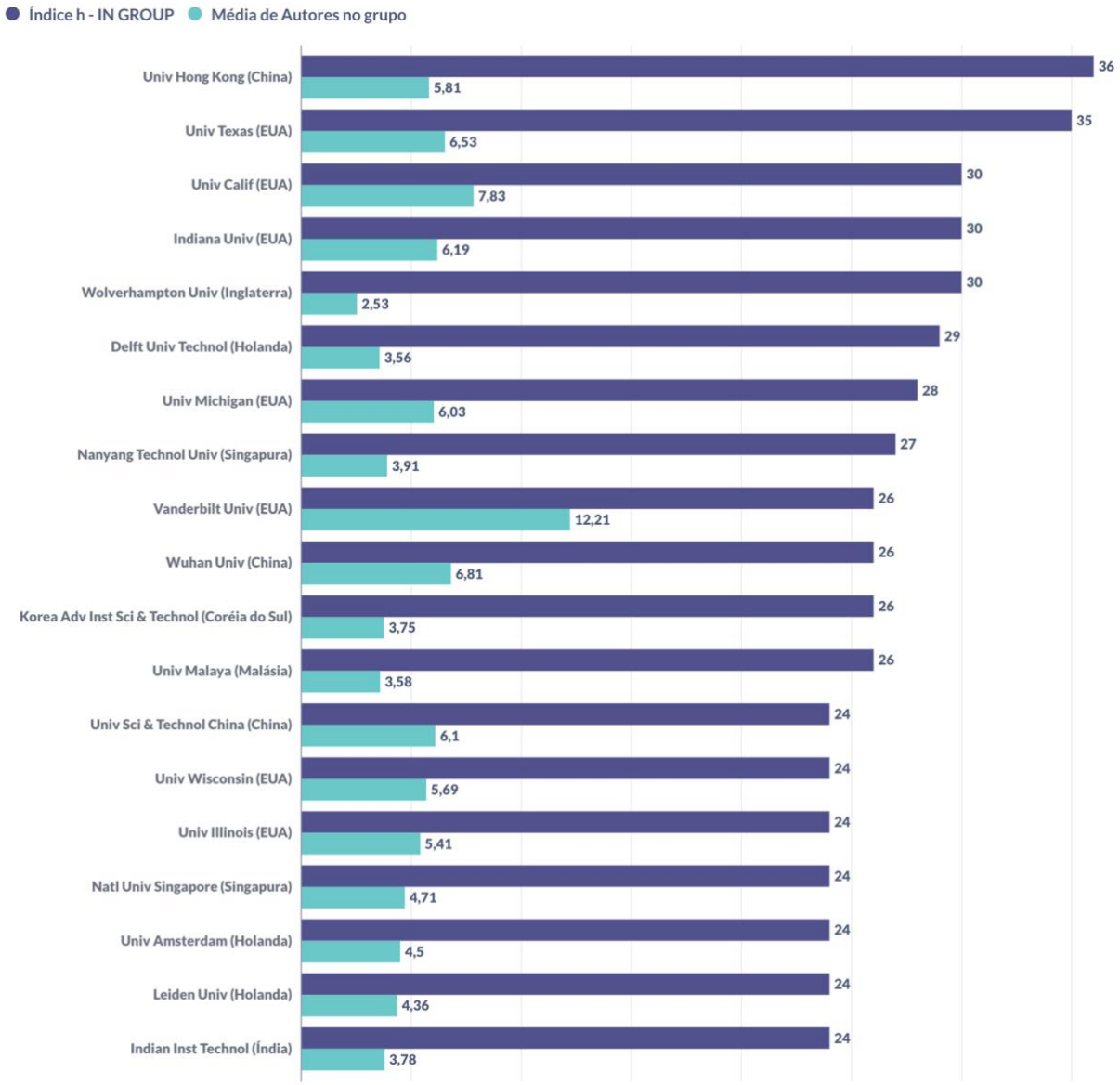
Universidades que possuem as suas principais produções com um número médio de autores elevado tiveram seus valores de índice h atenuados. Isso ocorreu com a Universidade de Columbia, que não esteve presente no *ranking* que considera a média de número de autores por artigo como Índice de Colaboração. No entanto, um número elevado de média de autores por artigo não significa que não há uma colaboração interna. As Universidades de Vanderbilt e Harvard se mantiveram no top 15 mesmo com uma média de autores de 12,21 e 8,86, respectivamente (Figura 58).

Mais duas universidades holandesas voltam ao *ranking* quando consideramos a média do número de autores por artigo: a Universidade de Amsterdam e a Universidade de Leiden. Estas duas instituições não estavam presentes nos demais *rankings* analisados aqui com índice h. No entanto, por possuírem muitos artigos com número de citações elevado e onde vários membros pertencem a elas, agora figuram entre as universidades que melhor colaboram internamente.

A Universidade de Ciência e Tecnologia da China e a Universidade do Illinois completam o *ranking* quando utilizamos as médias, por também possuírem expressiva produção com colaboração interna acima da média.

É possível observar que nem sempre a média é o melhor número para demonstrar de forma resumida o valor de um grupo. Em especial, em variáveis de cauda longa como número de citações e número de autores. Apresentamos na Figura 59 as primeiras 15 universidades quando o valor do índice h é calculado considerado a mediana do grupo, o que traz para análise somente os artigos onde o número de autores daquela universidade é maior do que a mediana de autores.

Figura 59 - índice h *IN-GROUP* e média de Autores (Top 15 universidades - Índice de Colaboração = Mediana de autores da Universidade)



Fonte: dados da pesquisa (2023).

Em um primeiro momento, o *ranking* que utiliza as medianas do número de autores se assemelha muito à que utiliza a média de autores. Isso ocorre pois os valores de média e mediana para essas universidades é muito próximo, conforme demonstrado na Tabela 8. As 6 primeiras universidades se mantêm em ambas as listagens (Hong Kong, Texas, Califórnia, Indiana, Wolverhampton e Delft). Quando se considera este conjunto, apenas a universidade da California muda o valor de seu índice, de 33 para 30 (Figura 58 e 59).

Tabela 8 - Média e Mediana do número de autores em Universidades

Universidade	País	Média Autores Artigo	Mediana Autores Artigo	Média Autores Universidade	Mediana Autores Universidade
Wuhan Univ	China	6,14	6	2,29	3
Univ Texas	EUA	5,68	4	1,6	2
Univ Hong Kong	China	5,16	5	1,53	2
Univ Calif	EUA	5,31	4	1,87	3
Indiana Univ	EUA	5,16	4	1,7	2
Univ Illinois	EUA	4,24	3	1,63	2
Nanjing Univ	China	5,92	5	2,05	3
Univ Wisconsin	EUA	5,24	4	1,69	2
SUNY	EUA	4,56	4	1,45	2
Univ Amsterdam	Netherlands	3,99	3	1,57	2
Univ London	England	4,36	3	1,11	2
Univ Washington	EUA	6	4	1,74	3
Univ Michigan	EUA	5,56	4	1,75	2
Univ N Carolina	EUA	5,01	4	1,68	3
Univ Granada	Spain	3,73	3	1,92	3

Fonte: dados da pesquisa (2023).

Como principal diferença entre os gráficos presentes nas Figura 58 e 59, perecebe-se a ausência das universidades de Granada e Harvard. As universidades de Michigan, Nanyang e Wuhan mantêm os valores de índice h e sobem uma posição (devido à ausência da universidade de Granada). Vanderbilt e o Instituto Coreano

mantêm seus valores de índice h e suas posições no *ranking*. A universidade da Malásia passa a compor o *ranking*, assumindo a nona posição, empatada com Wuhan, Vandebilt e Instituto Coreano.

Ne décima terceira posição temos 7 universidades empatadas: Amsterdam, Ciência e Teconologia da China, Instituto de Tecnologia de Indiana, Wisconsin, Leiden, Illionois e Universidade Nacional de Singapura. A Universidade de Harvard não consta mais no *ranking*, tendo sido substituída pela Universidade de Indiana (Figura 59).

Nesta seção são apresentados os experimentos realizados para a análise das teorias propostas na tese. Foram feitas as descrições das variáveis número de autores e número de citações, apresentando histogramas e diagramas de caixa. Verificou-se que a maioria dos artigos possui entre 2 e 4 autores. Ficou evidenciado que tanto as citações quanto o número de autores apresentam curvas assimétricas de distribuição. Observou-se ainda que a média, mediana e concentração das citações em artigos feitos em grupos se apresentou maior do que nos artigos feitos por um único autor. No entanto, não foi possível encontrar uma correlação entre número de citações e número de autores.

O *framework F-GROUP* foi aplicado para avaliar os grupos formados por países. Além da avaliação do índice h *IN-GROUP*, foram realizadas análises com outros indicadores. Essas análises foram realizadas para compreender melhor o conjunto de dados e analisar como cada um dos indicadores se comportava ao ser aplicado no grupo de países. As métricas relacionadas à contagem, como número de autores, citações e publicações foram bastante homogêneas, trazendo um conjunto reduzido de países. Ao analisar as médias dessas mesmas variáveis os grupos de países acabaram variando bastante.

Ao aferir os valores do índice h *IN-GROUP* para universidades, apareceram instituições dos principais países trazidos na primeira análise de grupos, em especial EUA e China. Ao utilizar a média como Índice de Colaboração, a discrepância no grupo dos 15 primeiros países foi atenuada. Embora o *ranking* tenha se mantido, EUA e China acabaram ficando com valores mais próximos do que quando o valor do Índice de Colaboração era fixo. Essa diferença mais atenuada pode ocorrer por diferentes fatores, um deles é que a China é um país com média de autores maior (5,48) dentre os principais países analisados. No entanto, a diferença principal ao utilizar a média foi uma baixa no valor do índice h *IN-GROUP* para os EUA. Como a média de autores

por artigo dos EUA é 3,9, todos os artigos com 4 ou mais autores foram desconsiderados, e este deve ter sido o principal motivo do impacto do índice h *IN-GROUP*.

A avaliação de universidades foi realizada com dados inferidos e tratados. O relacionamento entre os artigos e as instituições foram realizados com aproximações utilizando o nome mais curto possível disponível na base de dados para representar uma mesma instituição. Ainda assim, considerando a análise por país, com dados normalizados e definidos de forma padronizada na base de origem, foi possível verificar que as análises em universidades e outras instituições foi consistente com os números apresentados na avaliação realizada com países. Isto é, houve uma predominância de instituições americanas, seguida por instituições chinesas.

Identificou-se aspectos interessantes na avaliação da produção científica em Ciência da Informação e Biblioteconomia em universidades. A média e mediana do número de autores por artigo ficaram bastante próximas, sendo seguro afirmar que ambas podem ser utilizadas como Índices de Colaboração. A comparação de *rankings* que possuem diferentes Índice de Colaboração também se mostrou uma ferramenta interessante para a análise da composição de grupos de autores em diferentes instituições. Cabe ressaltar que universidades bastante renomadas, como Harvard e Columbia, apresentaram elevado número de autores para seus artigos, chegando a 12 em média.

6.5 COMPARAÇÃO DAS PROPOSTAS F-GROUP E IN-GROUP COM FORMAS DE AVALIAÇÃO CONSOLIDADAS NA LITERATURA

De forma a resumir os principais diferenciais da forma de agregação *IN-GROUP* combinada ao framework *F-GROUP,* apresentamos o Quadro 19. Neste quadro é possível visualizarmos as principais características identificadas como lacunas no decorrer do trabalho, bem como aspectos identificados como relevantes durante a análise dos dados. Além da combinação de *IN-GROUP* com *F-GROUP* apresentamos outras abordagens para cálculo do índice h para grupos. Considerando que os indicadores que apresentam contadores, como total de autores, artigos, colaborações, são atualmente a forma de avaliação mais utilizada, incluímos este grupo de indicadores na coluna Totais, de forma a representar este conjunto.

Quadro 19 - Comparação *IN-GROUP* e *F-GROUP* com abordagens consolidadas na literatura

Características	***IN-GROUP* e *F-GROUP***	**Outras abordagens para índice h em grupos**	**Totais**
Permite combinar qualidade e quantidade ao avaliar a produção científica	Sim	Sim	-
Permite a comparação entre grupos de diferentes tamanhos	Sim	Sim	-
Critérios claros para seleção dos artigos que representam o grupo	Sim	-	-
Forma de agregação dos artigos é clara	Sim	-	-
Facilita o planejamento da análise da produção científica de um grupo de autores	Sim	-	-
Facilita a comparação entre diferentes trabalhos que utilizam indicadores numéricos para avaliar a produção científica de grupos	Sim	-	-
Permite inferir o índice h de um grupo considerando características inerentes ao próprio grupo	Sim	-	-

Fonte: Elaborado pela Autora

A combinação da abordagem de seleção e agregação *IN-GROUP* e do *Framework F-GROUP* apresentam diferenciais em relação aos demais grupos analisados, conforme demonstrado no Quadro 19. Esta evidência permite indicar que existem vantagens na utilização do *F-GROUP* quando comparado com as principais formas de avaliação da produção científica para grupos.

Atualmente, os Totais são a forma mais utilizada para medir a produtividade em grupos. No entanto, ao utilizar os Totais não é possível combinar em um único indicador a qualidade e quantidade da produção científica de um grupo. Além disso, os valores que representam totais não permitem uma comparação justa entre grupos com diferentes números de autores. Mesmo que os valores sejam utilizados com médias, observamos em nossos exemplos para Países e Universidades que os valores medianos também não representam a produção científica da melhor forma.

A definição de critérios claros para seleção e agregação dos artigos também influencia na composição de medidas em Totais. Mesmo para contar os artigos é necessário estabelecer de forma clara quais serão os artigos utilizados, e de que forma eles representarão todo o grupo que está sendo analisado.

A utilização do *framework F-GROUP* se demonstrou útil também para facilitar o planejamento de análises da produção científica em grupos de autores. Adicionalmente, a utilização de médias e medianas como Índice de Colaboração permite que os grupos sejam comparados de acordo com suas características próprias, tornando assim mais justa a comparação entre diferentes grupos. A contínua utilização do novo *framework* para futuros trabalhos de análise da produtividade em grupos de autores podem ser relevantes para a padronização e permitir a comparação entre diferentes trabalhos.

A seguir são apresentadas as conclusões do presente trabalho.

7 CONCLUSÃO

A avaliação da produção científica de grupos de autores ocorre de forma pouco padronizada, o que dificulta a comparação entre diferentes conjuntos de pesquisadores. Estas dificuldades começam já na definição do grupo de autores, que pode ser definida somente como coautores ou, ainda, com a avaliação de grupos maiores, como países inteiros e áreas de pesquisa. A busca por trabalhos que abordam a mensuração da produção científica é também dificultada pela falta da padronização de termos. Neste trabalho, houve a necessidade de se realizar uma busca inicial para tentar identificar os termos utilizados, dada a falta de padrão na pesquisa sobre esta temática.

Dentre os termos identificados para procurar trabalhos que abordam a mensuração da produção científica em grupos de autores, destacamos alguns pontos principais. O primeiro é o termo indicador, acompanhado dos termos equivalentes: métrica e índice. As métricas para ciência podem ter diversas vertentes. Para delimitar as métricas relacionadas à produção científica, os termos produtividade, performance e qualidade também foram utilizados, sendo também necessário incluir termos relacionados a citação e bibliometria. O conceito de grupo em pesquisa é mais comumente associado a colaboração, e termos como coautoria e rede também são relevantes.

Ao identificar os termos mais relevantes e prosseguir com a busca na literatura, outra falta de padronização emergiu: a forma de agregação e seleção dos artigos utilizados para avaliar os grupos. Mesmo quando se utiliza métricas simples, como total de artigos, a definição das formas de seleção dos artigos pode tornar os resultados diferentes. Neste trabalho, aplicamos três formas de seleção e agregação para atribuir um valor de índice h para um grupo de autores:

i. Empregar a linha dos trabalhos que utilizam a média dos valores dos índices h de cada membro do grupo.
ii. Seguir a dos trabalhos que utilizam os artigos em que algum membro do grupo participou para calcular um índice h único.
iii. Utilizar a forma de agregação *IN-GROUP*, onde um número definido de autores do grupo avaliado precisa ser coautor nos trabalhos para que estes trabalhos sejam contabilizados na composição do índice h do grupo.

A revisão bibliográfica deste trabalho foi realizada a partir de três questões de pesquisa: *RQ1: Como tem sido realizada a avaliação da produtividade dos grupos de pesquisadores nos últimos anos? RQ2: Quais indicadores de produtividade científica têm sido utilizados para avaliar esses grupos de pesquisadores? RQ3: Existe uma metodologia consolidada para a aplicação de indicadores de produtividade a grupos de pesquisadores?* A Revisão Sistemática da Literatura propões alcançar o objetivo específico: Identificar as formas de avaliação existentes para grupos de pesquisadores através de indicadores.

Com base nos resultados recuperados na revisão sistemática, foi possível observar que a avaliação da produtividade em grupos de pesquisadores ocorre de formas variadas. Existem trabalhos que sugerem novos indicadores especificamente para avaliar novos grupos de autores e, ainda, muitos trabalhos com análises em grupos específicos, em países, províncias, instituições e campos de pesquisa.

Quanto à segunda questão de pesquisa *RQ2: Quais indicadores de produtividade científica têm sido utilizados para avaliar esses grupos de pesquisadores?* Os resultados evidenciam que totais de artigos, citações e colaborações são os indicadores mais comumente utilizados para avaliar grupos de pesquisadores. O impacto, relacionado ao número de citações, também é utilizado com frequência. Os artigos que medem a produtividade em ciência para grupos de pesquisadores são, em sua grande maioria, compostos por casos de uso. As métricas são compostas por totais e somatórios de artigos e citações. A combinação de índices que agregam contagem de artigos e citações ainda é bastante incipiente. Enquanto a produtividade de um único autor tem sido avaliada com a proposição de diversos novos indicadores, a produtividade em grupos ainda possui sua avaliação centrada em métricas simples.

Sobre a terceira questão *RQ3: Existe uma metodologia consolidada para a aplicação de indicadores de produtividade em grupos de pesquisadores?* Com base nos artigos analisados, é possível afirmar que há uma certa consolidação acerca da utilização de indicadores totalizadores: artigos, colaborações e citações. Na revisão, encontramos um *framework* (WANG *et al.*, 2021) que define em linhas gerais os passos para identificar os indicadores que melhor representem a produtividade de um grupo de pesquisa, mas não especifica como deve acontecer a avaliação nem os critérios necessários para avaliar o grupo.

A partir da terceira questão de pesquisa, que avalia se há uma metodologia consolidada para medir a produção de grupos de pesquisadores, foi possível identificar a necessidade de padronização na avaliação de grupos de pesquisadores. Um *framework* para facilitar a medição da produção científica para um grupo de pesquisadores é proposto neste trabalho, o *F-GROUP*. O *framework F-GROUP* é centrado na seleção e agregação de artigos de forma a facilitar tanto o planejamento de uma avaliação quanto a comparação entre diferentes estudos. A proposição do *framework F-GROUP* e da forma de agregação *IN-GROUP* está alinhada ao objetivo *de* propor novas formas de avaliar grupos de pesquisadores, considerando as lacunas presentes nas formas atuais.

O *framework F-GROUP* possui três fases:

I. Identificação de Parâmetros iniciais, que engloba as etapas de descrição do(s) grupo(s), objetivos da avaliação e fonte dos dados.
II. Seleção de Artigos, com intuito de que os artigos selecionados melhor representem o grupo que está sendo avaliado, observando e melhor forma de agregação e contagem. Seguida da especificação dos aspectos representativos dos grupos.
III. Definição dos Indicadores que serão utilizados para avaliar os grupos. São ainda destacados no *framework* o Índice de Colaboração e a abordagem *IN-GROUP*.

A abordagem *IN-GROUP* compreende uma forma de selecionar os artigos que serão agregados ao se avaliar grupos de autores. Alguns trabalhos não especificam de que forma os artigos que representam o grupo foram selecionados. Nesta abordagem, proposta no escopo desta tese, apenas os trabalhos onde parte da equipe trabalhou em conjunto são considerados na avaliação.

O objetivo de avaliar a aplicabilidade das novas abordagens de avaliação por meio de casos de uso foi atendido com a utilização de três casos. Os casos de uso incluíram avaliações através de editais de pesquisa, avaliação entre diferentes países e entre departamentos de universidades. A apresentação dos casos de uso trouxe exemplos de situações cotidianas em que é necessário avaliar a produção de autores em conjunto.

Foram realizados dois experimentos com o novo *framework F-GROUP* e com a nova forma de agregação *IN-GROUP*. Os experimentos foram realizados a partir de

uma consulta realizada na base de dados *Web Of Science* com pesquisa entre 10 de abril de 2013 a 10 de abril de 2023. O objetivo de comparar os dispositivos propostos no trabalho com as formas de avaliação consolidadas na literatura foi, portanto, contemplado.

No primeiro experimento, considerou-se a produção científica em Ciência da Informação e Biblioteconomia nos últimos 10 anos. Utilizando o *framework F-GROUP*, definiu-se os grupos como sendo países e universidades. As formas de agregação utilizadas foram *IN-GROUP* e *all papers*. Os indicadores obtidos foram bastante variados com contadores de citações, autores e artigos, bem como médias e índice h com diferentes formas de agregação. As avaliações foram realizadas a partir de *rankings* com análise dos 15 primeiros colocados.

Uma análise descritiva dos dados indicou que as variáveis número de autores e citação possuem característica assimétrica positiva, com pico à esquerda e cauda longa à direita no histograma. Ao discretizar a variável número de autores para classificação entre trabalhos individuais ou trabalhos em grupo, observou-se que a média, mediana e *box plot* apresentam valores maiores de citações para trabalhos realizados em grupo. Não há correlação entre as variáveis número de autores e número de citações no grupo analisado, uma vez que um maior número de autores não está diretamente relacionado a um maior número de citações.

Na avaliação da produção dos países, foi possível identificar que a produção dos EUA é bastante superior, seja em número de autores, artigos ou citações. Ao utilizar as médias ao invés dos números totais, outros países com quantidade de produção inferior acabavam apresentando valores elevados. Por este motivo, não é segura a utilização de médias ao avaliar a produção científica de grandes grupos de autores. Ao se utilizar o índice h em diferentes formas de agregação, os EUA mantiveram seu destaque, sendo seguidos pela China, segunda colocada, em diferentes métricas. Cabe mais uma vez destacar que a *Web Of Science* é uma base de dados de artigos científicos americana, onde o idioma predominante é o inglês. Mesmo assim, há uma aproximação consistente, ao longo do período analisado, entre a produção desses países.

A análise da distribuição do número de autores por país revelou que os EUA possuem uma distribuição do número de autores com cauda mais alongada que os demais países, chegando próximo a 60 autores por artigo. A China possui no máximo

15 autores por artigo, estando mais próxima dos demais países que compõem o *ranking*. Ao se aferir os valores do índice h *IN-GROUP*, os mesmos países que compunham os grupos líderes nos indicadores de totais foram apresentados. Ao utilizar a média de autores do mesmo grupo como Índice de Colaboração, a discrepância no grupo dos 15 primeiros países foi atenuada. Embora o *ranking* tenha se mantido, EUA e China acabaram ficando com valores mais próximos do que quando o valor do Índice de Colaboração era fixo. Essa diferença mais atenuada pode ocorrer por diferentes fatores, um deles é que a China é um país com média de autores maior (5,48) dentre os principais países analisados. No entanto, a diferença principal ao utilizar a média foi uma queda no valor do índice h para os EUA.

O segundo experimento foi realizado com 20.567 diferentes instituições. Ao considerar os países vinculados a cada instituição, mais uma vez são observadas as predominâncias americana e chinesa. A universidade com maior número de artigos foi Wuhan, na China. O maior número de autores foi identificado na California, EUA. Quanto ao número de citações, a Universidade de Hong Kong ocupa a primeira posição. As primeiras universidades, considerando o número de produção, possuem média de autores por artigos com valor maior do que o observado dentre os países com maior produção.

O índice h *IN-GROUP* aplicado às universidades apresentou valores significativamente menores do que quando utilizamos índice h *all papers*. Mesmo com uma variação menor do Índice de Colaboração, o *ranking* variou bastante à medida que o valor do IC mudou de 2 para 3 e 4. Destaque para Hong Kong, primeira na produção *all papers* e que não apareceu mais no *ranking* com quando o valor de IC era 3 ou 4. As universidades americanas ficaram mais bem posicionadas com o aumento do IC. Ao utilizar IC como média ou mediana do número de autores, Hong Kong volta à primeira posição. Hong Kong possui média de autores por artigo de 5, mas média de autores da própria universidade de 1,53.

Os resultados evidenciam que a pesquisa sobre indicadores para produtividade em grupos é pouco explorada. Os trabalhos são compostos, em grande maioria, por estudos de caso e utilizando métricas simples. Os experimentos evidenciaram que os grupos formados pelos autores nos trabalhos de Ciência da Informação e Biblioteconomia nos últimos dez anos é bastante heterogêneo. Adicionalmente, observamos que os países e instituições de destaque, possuem um número de coautores elevado.

Uma das perguntas centrais levantadas nesta pesquisa é "Como avaliar quantitativamente e qualitativamente a produção científica de grupos de pesquisadores?", focado nesta pergunta apresentamos o *F-GROUP* como uma nova possível forma de avaliação. São considerados aspectos que facilitam o planejamento e, em especial, a seleção dos artigos que serão utilizados para avaliar o grupo. No entanto, a implementação das pesquisas é bastante diversa e com muitos cenários inexplorados. É através da sua utilização que poderemos aprimorar o Framework e, em trabalhos futuros, implementar extensões de acordo com as diferentes necessidades de avaliação.

Em futuros trabalhos, é possível explorar diferentes valores para o Índice de Colaboração. Apesar de apresentarmos valores fixos e característicos de cada grupo, como média e mediana, novos trabalhos podem abordar qual é o melhor Índice de Colaboração de acordo com as características dos grupos analisados. Novos experimentos comparando outras formas de agregação como média e sucessiva, podem ser realizadas, utilizando bases de dados que possuam a listagem de artigos e, também o índice h dos autores. De forma geral, tanto a abordagem *IN-GROUP* quanto o *framework F-GROUP* sugerem direcionamentos para futuras pesquisas que proponham a avaliação da produtividade de grupos de pesquisadores. Novas pesquisas que avaliem grupos de autores através de indicadores, podem deixar mais claro como contabilizam os artigos individualmente para compor os índices.

REFERÊNCIAS

ABRAMO, G.; D'ANGELO, C. A. **A methodology to compute the territorial productivity of scientists: The case of Italy**. **JOURNAL OF INFORMETRICS**PO BOX 211, 1000 AE AMSTERDAM, NETHERLANDSELSEVIER SCIENCE BV, , out. 2015.

ACERO, L.; KLEIN, H. E. Coautorias nas publicações brasileiras sobre medicina regenerativa: assimetrias na colaboração científica internacional. **Revista Eletrônica de Comunicação, Informação e Inovação em Saúde**, v. 15, 2021.

AFTAB, U.; SIDDIQUI, G. F. **Big data augmentation with data warehouse: A survey**. 2018 IEEE International Conference on Big Data (Big Data). **Anais**...IEEE, 2018.

AGARWAL, A. et al. Bibliometrics: tracking research impact by selecting the appropriate metrics. **Asian Journal of Andrology**, v. 18, n. 2, p. 296, 2016.

ALEIXANDRE, J. L. et al. Scientific productivity and collaboration in viticulture and enology in Latin American countries [Productividad y colaboración científica en viticultura y enología en los países latinoamericanos]. **Ciencia e Investigacion Agraria**, v. 40, n. 2, p. 429–443, 2013a.

ALEIXANDRE, J. L. et al. Scientific productivity and collaboration in viticulture and enology in Latin American countries. **Ciencia e investigación agraria**, v. 40, n. 2, p. 429–443, maio 2013b.

ALMEIDA, A. J. Contributos da Sociologia para a compreensão dos processos de profissionalização. v. 1, p. 13, 2010.

ALTMANN, J.; ABBASI, A.; HWANG, J. Evaluating the productivity of researchers and their communities: The RP-Index and The CP-Index. **International Journal of Computer Science and Applications**, p. 15, 2009.

ALVES, E. F. T., B. H. ;PAVANELLI, M. A. ;OLIVEIRA. Rede de coautoria institucional em Ciência da Informação: uma comparação entre indicadores de rede e os conceitos CAPES. **Em Questão**, v. 20, 2014.

ANGELIN, P. E. PROFISSIONALISMO E PROFISSÃO: TEORIAS SOCIOLÓGICAS E O PROCESSO DE PROFISSIONALIZAÇÃO NO BRASIL. **Revista Espaço de Diálogo e Desconexão**, v. 3, 2010.

ANTUNES, R.; ALVES, G. As mutações no mundo do trabalho na era da mundialização do capital. **Educação & Sociedade**, v. 25, n. 87, p. 335–351, ago. 2004.

ARAÚJO, C. A. Bibliometria: evolução histórica e questões atuais. **Em Questão**, v. 12, n. 1, 2006.

AUSLOOS, M. A scientometrics law about co-authors and their ranking. The co-author core. **arXiv:1207.1614 [physics]**, 14 jan. 2013.

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. **Estatística: para cursos de engenharia e informática.** São Paulo: Atlas, 2004. v. 3

BORNMANN, L. et al. Excellence networks in science: A Web-based application based on Bayesian multilevel logistic regression (BMLR) for the identification of institutions collaborating successfully. **Journal of Informetrics**, v. 10, n. 1, p. 312–327, fev. 2016a.

BORNMANN, L. et al. **Excellence networks in science: A Web-based application based on Bayesian multilevel logistic regression (BMLR) for the identification of institutions collaborating successfully**. **JOURNAL OF INFORMETRICS**RADARWEG 29, 1043 NX AMSTERDAM, NETHERLANDSELSEVIER, , fev. 2016b.

BOSSY, M. J. The Last of the Litter : “Netometrics”. 1995.

ÇAKIR, M. P. et al. Multi-authoring and its impact on university rankings: a case study of CERN effect on Turkish universities. **Studies in Higher Education**, v. 44, n. 6, p. 1052–1068, 3 jun. 2019.

CAKIR, M. P. et al. **Multi-authoring and its impact on university rankings: a case study of CERN effect on Turkish universities**. **STUDIES IN HIGHER EDUCATION**2-4 PARK SQUARE, MILTON PARK, ABINGDON OX14 4RN, OXON, ENGLANDROUTLEDGE JOURNALS, TAYLOR & FRANCIS LTD, , 3 jun. 2019.

CICERO, T.; MALGARINI, M. Research Collaboration and Bibliometric Performance. **Handbook Bibliometrics**, p. 319, 2020.

CODINA-CANET, A., M. A. ;PERIANES-RODRíGUEZ. Análisis de la colaboración científica de la Universidad Politécnica de Valencia (Scopus, 2003-2008). **Métodos de información (Espanha)**, v. 3, 2012.

CRESWELL, J. W. **Projeto de pesquisa: métodos qualitativo, quantitativo e misto; tradução Magda Lopes**. Porto Alegre: Artmed, 2010.

CSILLAG, J. M. **Análise do valor: metodologia do valor**. São Paulo: Atlas, 1995.

CUPANI, A. **Filosofia da tecnologia: um convite**. 3. ed. Florianópolis: Editora da UFSC, 2016.

CURTY, R. G.; DELBIANCO, N. R. As diferentes metrias dos estudos métricos da informação:: evolução epistemológica, inter-relações e representações. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 25, p. 01–21, 20 out. 2020.

DE MOYA-ANEGON, F. et al. Statistical relationships between corresponding authorship, international co-authorship and citation impact of national research systems. **Journal of Informetrics**, v. 12, n. 4, p. 1251–1262, nov. 2018.

DOS SANTOS, A. A.; DUTRA, M. L. Is the h-index Applicable and Sufficient to Assess Research Groups? **Mobile Networks and Applications**, 13 abr. 2022.

DUBAR, C. **A socialização: construção das identidades sociais e profissionais**. Brasil: São Paulo, 2005.

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

FASSIN, Y. The emergence of China in international academic management research: A nuanced analysis following the new f^2-methodology. p. 21, 2021a.

FASSIN, Y. **The emergence of China in international academic management research: A nuanced analysis following the new f(2)-methodology**. **MALAYSIAN JOURNAL OF LIBRARY & INFORMATION SCIENCE**UNIV MALAYA, FAC COMPUTER SCIENCE & INFORMATION TECH, KUALA LUMPUR, 50603, MALAYSIAUNIV MALAYA, FAC COMPUTER SCIENCE & INFORMATION TECH, , 2021b.

FELIPE DIAZ-CARDENAS, A. et al. **Characteristics of the scientific production of Puebla institutions indexed in the Web of Science**. **INVESTIGACION BIBLIOTECOLOGICA**CIUDAD UNIV, CENTRO UNIV BIBLIOTECOLOGICAS, TORRE II HUMANIDADES, PISO 11, 12 & 13, MEXICO CITY, CP 04510, MEXICOUNIV NACIONAL AUTONOMA MEXICO, , 2017.

FLATT, J.; BLASIMME, A.; VAYENA, E. Improving the Measurement of Scientific Success by Reporting a Self-Citation Index. **Publications**, v. 5, n. 3, p. 20, 1 ago. 2017.

FREIDSON, E. Para uma análise comparada das profissoes: a institucionalização do discurso e do conhecimento formais. p. 9, 1996.

FRUNZA, O. et al. Exploiting the systematic review protocol for classification of medical abstracts. **Artificial intelligence in medicine**, v. 51, n. 1, p. 17–25, 2011.

GASPARYAN, A. Y. et al. Researcher and Author Impact Metrics: Variety, Value, and Context. **Journal of Korean Medical Science**, v. 33, n. 18, p. e139, 2018.

GAUFFRIAU, M. A categorization of arguments for counting methods for publication and citation indicators. **Journal of Informetrics**, v. 11, n. 3, p. 672–684, ago. 2017.

GAUFFRIAU, M. Counting methods introduced into the bibliometric research literature 1970–2018: A review. **Quantitative Science Studies**, p. 1–44, 12 out. 2021.

GIL, A. C. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 1991.

GOLICHENKO, O. G.; MALKOVA, A. A. The Analysis of Processes of New Knowledge Production in Key World Regions and Russia. **Journal of the Knowledge Economy**, v. 8, n. 4, p. 1133–1145, dez. 2017.

GONZALEZ-ALCAIDE, G. et al. **Dominance and leadership in research activities: Collaboration between countries of differing human development is reflected through authorship order and designation as corresponding authors in scientific publications**. **PLOS ONE**1160 BATTERY STREET, STE 100, SAN FRANCISCO, CA 94111 USAPUBLIC LIBRARY SCIENCE, , 8 ago. 2017.

GONZÁLEZ-VALIENTE, C. L. et al. Mapping the Evolution of Intellectual Structure in Information Management Using Author Co-citation Analysis. **Mobile Networks and Applications**, 22 fev. 2019.

GRÁCIO, M. C. C. Colaboração científica: indicadores relacionais de coautoria. **Brazilian Journal of Information Science: research trends**, v. 12, n. 2, 1 ago. 2018.

GRACIO, M. C. C. et al. **Does corresponding authorship influence scientific impact in collaboration: Brazilian institutions as a case of study**. **SCIENTOMETRICS**VAN GODEWIJCKSTRAAT 30, 3311 GZ DORDRECHT, NETHERLANDSSPRINGER, , nov. 2020.

GREGORIO-CHAVIANO, O. et al. Dialnet Metrics as a bibliometric evaluation tool: contributions to the analysis of the scientific activity in Social Sciences and Humanities. **PROFESIONAL DE LA INFORMACION**, v. 30, n. 3, jun. 2021.

GYULA NAGY. **Text Mining-based Scientometric Analysis in Educational Research**. Official Conference Proceedings. **Anais**... Em: THE EUROPEAN CONFERENCE ON EDUCATION 2018. 2018.

HARARI, Y. N.; GEIGER, P.; COMPANHIA DAS LETRAS (FIRM). **21 lições para o Século XXI**. [s.l: s.n.].

HIRSCH, J. E. An index to quantify an individual's scientific research output. **Proceedings of the National Academy of Sciences**, v. 102, n. 46, p. 16569–16572, 15 nov. 2005.

HIRSCH, J. E. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. **Scientometrics**, v. 85, n. 3, p. 741–754, dez. 2010.

HIRSCH, J. E. hα: An index to quantify an individual's scientific leadership. **Scientometrics**, v. 118, n. 2, p. 673–686, fev. 2019.

JACSÓ, P. The h-index for countries in Web of Science and Scopus. **Online Information Review**, v. 33, n. 4, p. 831–837, 7 ago. 2009.

JAMIL, G. L.; NEVES, J. T. DE R. A era da informação: considerações sobre o desenvolvimento das tecnologias da Informação. **Perspect. cienc. inf.**, v. 5, n. 1, p. 13, 2000.

JIN, B. et al. The R- and AR-indices: Complementing the h-index. **Chinese Science Bulletin**, v. 52, n. 6, p. 855–863, mar. 2007.

JOSHI, M. A. Bibliometric Indicators for Evaluating the Quality of Scientific Publications. **The Journal of Contemporary Dental Practice**, v. 15, n. 2, p. 258–262, abr. 2014.

KHAN, N. et al. Part I: The Application of the h-Index to Groups of Individuals and Departments in Academic Neurosurgery. **World Neurosurgery**, v. 80, n. 6, p. 759-765.e3, dez. 2013.

KITCHENHAM, B. et al. Systematic literature reviews in software engineering – A systematic literature review. **Information and Software Technology**, v. 51, n. 1, p. 7–15, jan. 2009.

KOSMULSKI, M. A new Hirsch-type index saves time and works equally well as the original h-index. **International Society for Scientometrics and Informetrics (ISSI)**, p. 3, 2006 2005.

KUMAR, S. Scientometric analysis of research publications in Astronomy and Astrophysics research in India: a study based on WoS. **Library Philosophy and Practice**, v. 2020, p. 1–20, 2020.

KUMAR, S. et al. Social Indicators Research: A Retrospective Using Bibliometric Analysis. **Social Indicators Research**, 15 nov. 2021.

LANCHO-BARRANTES, B. S.; CANTU-ORTIZ, F. J. Measuring the incidence of social factors on scientific research: A socio-scientometrics analysis of strategic countries. **Investigación Bibliotecológica: archivonomía, bibliotecología e información**, v. 34, n. 85, p. 61, 6 out. 2020.

LEYDESDORFF, L. **The challenge of scientometrics: The development, measurement, and self-organization of scientific communications**. [s.l.] Universal-Publishers, 2001.

MACEDO, M.; SOUZA, M. R. D. **TEORIA, MODELOS E FRAMEWORKS: CONCEITOS E DIFERENÇAS**. . Em: CONGRESSO INTERNACIONAL DE CONHECIMENTO E INOVAÇÃO (CIKI). 15 fev. 2023. Disponível em: <https://proceeding.ciki.ufsc.br/index.php/ciki/article/view/1249>. Acesso em: 2 out. 2023

MESCHINI, F. O.; ALVES, B. H.; OLIVEIRA, E. F. T. DE. Coautorias internacionais do brasil em estudos métricos da informação e seus canais de comunicação. **Revista Conhecimento em Ação**, v. 3, n. 2, p. 54–69, 31 dez. 2018.

MITRA, P. Hirsch-type indices for ranking institutions scientific research output. **Current Science**, v. 91, n. 11, p. 1439, 2006.

MORENO-DELGADO, A.; GORRAIZ, J.; REPISO, R. Assessing the publication output on country level in the research field communication using Garfield's Impact Factor. **Scientometrics**, v. 126, n. 7, p. 5983–6000, jul. 2021.

MUELLER, S. P. M. Uma profissão em evolução: profissionais do informação no Brasil sob a ótica de Abbott - proposta de estudo. p. 32, 2004.

MUGNAINI, R.; PACKER, A. L.; MENEGHINI, R. Comparison of scientists of the Brazilian Academy of Sciences and of the National Academy of Sciences of the USA on the basis of the h-index. **Brazilian Journal of Medical and Biological Research**, v. 41, n. 4, p. 258–262, abr. 2008.

OECD. **Gross domestic spending on R&D**. OECD, , 2023. Disponível em: <https://data.oecd.org/rd/gross-domestic-spending-on-r-d.htm>

PAKKAN, S. et al. Quest for Ranking Excellence Impact Study of Research Metrics. **DESIDOC Journal of Library & Information Technology**, v. 41, n. 1, p. 61–69, 11 fev. 2021.

PAN, R. K.; FORTUNATO, S. Author Impact Factor: tracking the dynamics of individual scientific impact. **Scientific Reports**, v. 4, n. 1, p. 4880, maio 2015.

PARK, H. W.; YOON, J.; LEYDESDORFF, L. The normalization of co-authorship networks in the bibliometric evaluation: the government stimulation programs of China and Korea. **Scientometrics**, v. 109, n. 2, p. 1017–1036, nov. 2016.

PARRA, M. R.; COUTINHO, R. X.; PESSANO, E. F. C. UM BREVE OLHAR SOBRE A CIENCIOMETRIA: ORIGEM, EVOLUÇÃO, TENDÊNCIAS E SUA CONTRIBUIÇÃO PARA O ENSINO DE CIÊNCIAS. **Revista Contexto & Educação**, v. 34, n. 107, p. 126–141, 28 mar. 2019.

PFRIEGER, F. W. **TeamTree analysis: A new approach to evaluate scientific production**. **PLOS ONE**1160 BATTERY STREET, STE 100, SAN FRANCISCO, CA 94111 USAPUBLIC LIBRARY SCIENCE, , 21 jul. 2021.

PILCEVIC, I.; JEREMIC, V.; VUJOSEVIC, D. Evaluating the scientific performance of institutions within the university: An example from the University of Belgrade leading institutions. **Journal of the Serbian Chemical Society**, v. 83, n. 11, p. 1285–1295, 2018.

PINTO, A. L. Arquivometria. **ÁGORA: Arquivologia em debate**, v. 21, p. 59–69, 2011.

PINTO, A. L.; MATIAS, M. Indicadores Científicos e as Universidades Brasileiras; Indicadores Científicos y las Universidades Brasileñas. **Informação & Informação**, v. 16, n. 3, p. 1–18, 8 maio 2012.

POPPER, K. R. **A logica da pesquisa cientifica**. São Paulo (SP): Cultrix, 1975.

RAD, A. E. et al. The H-Index in Academic Radiology. **Academic Radiology**, v. 17, n. 7, p. 817–821, jul. 2010.

RADOS, G. J. V.; VALERIM, P.; BLATTMANN, U. **VALOR AGREGADO A SERVIÇOS E PRODUTOS DE INFORMAÇÃO**. mar. 1999.

ROEMER, R. C.; BORCHARDT, R. **Meaningful Metrics: A 21st-Century Librarian's Guide to Bibliometrics, Altmetrics, and Research Impact**. [s.l: s.n.].

ROSAS, M. C. C., F. S. ;GRÁCIO. Colaboração científica como procedimento para a análise de um domínio: uma aplicação na área de Zootecnia. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v. 20, 2015.

ROSSI, P.; STRUMIA, A.; TORRE, R. **Bibliometrics for collaboration works**. (Catalano, G and Daraio, C and Gregori, M and Moed, HF and Ruocco, G, Ed.)**17TH INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS (ISSI2019), VOL I**: Proceedings of the International Conference on Scientometrics and Informetrics.KATHOLIEKE UNIV LEUVEN, FACULTEIT E T E W, DEKENSTRAAT 2, LEUVEN, B-3000, BELGIUMINT SOC SCIENTOMETRICS & INFORMETRICS-ISSI,

, 2019.

ROUSSEAU, R.; YANG, L.; YUE, T. A discussion of Prathap's h2-index for institutional evaluation with an application in the field of HIV infection and therapy. **Journal of Informetrics**, v. 4, n. 2, p. 175–184, abr. 2010.

SAHOO, J. et al. Authorship trend and content analysis: A case study on highly cited articles in library and information science journals. **Performance Measurement and Metrics**, v. 21, n. 1, p. 33–51, 15 nov. 2019.

SAHOO, J. et al. Authorship trend and content analysis: A case study on highly cited articles in library and information science journals. **Performance Measurement and Metrics**, v. 21, n. 1, p. 33–51, 2020.

SANTOS, A. A. DOS et al. **Inferring Relationships from Trajectory Data.** GeoInfo. **Anais**...2015.

SANTOS, A. A. DOS; DUTRA, M. L. Indexes for Evaluating Research Groups: Challenges and Opportunities. Em: BISSET ÁLVAREZ, E. (Ed.). **Data and Information in Online Environments**. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Cham: Springer International Publishing, 2021. v. 378p. 41–53.

SANTOS, A. A. DOS; DUTRA, M. L. Is the h-index Applicable and Sufficient to Assess Research Groups? **Mobile Networks and Applications**, 13 abr. 2022.

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

SENA, P. M. B.; CARVALHO SEGUNDO, W. L. R. D.; MELO, B. A. D. Ciência aberta na parceria para governo aberto: compromisso por um novo modelo de avaliação. **Informação & Informação**, v. 27, n. 3, p. 14–33, 27 abr. 2023.

SILVA, D. A. A., E. D. P. E. ;MARIA, T. C. ;FROGERI, R. F. ;FERREIRA. ANÁLISE SOCIOMÉTRICA DO GRUPO DE TRABALHO 4 DO ENANCIB: UM ESTUDO DAS RELAÇÕES ENTRE OS AUTORES, COAUTORES E INSITUIÇÕES DE ENSINO. **Encontro Nacional de Pesquisa em Ciência da Informação**, 2019.

SILVA, E. L. DA; MENEZES, E. M. **Metodologia da pesquisa e elaboração de dissertação**. 4. ed. Florianópolis: UFSC, 2005.

SILVA, F. C. C. DA; SCHONS, C. H.; RADOS, G. J. V. A gestão de serviços em bibliotecas universitárias: proposta de modelo. **Informação & Informação**, v. 11, n. 2, p. 82, 15 dez. 2006.

SILVA, J. A. DA; BIANCHI, M. DE L. P. Cientometria: a métrica da ciência. **Paidéia (Ribeirão Preto)**, v. 11, n. 21, p. 5–10, 2001.

SILVEIRA, M. A. A. DA; CAREGNATO, S. E. Demarcações epistemológicas dos estudos de citação: o fenômeno da citação. **Informação & Sociedade: Estudos**, v. 27, n. 3, 24 dez. 2017.

SIMON. **The Koli Calling Community**. Proceedings of the 16th Koli Calling International Conference on Computing Education Research. **Anais**...: Koli Calling '16.New York, NY, USA: Association for Computing Machinery, 2016a. Disponível em: <https://doi-org.ez46.periodicos.capes.gov.br/10.1145/2999541.2999562>

SIMON. **The koli calling community**. Proceedings of the 16th Koli Calling International Conference on Computing Education Research. **Anais**... Em: KOLI CALLING 2016: 16TH KOLI CALLING INTERNATIONAL CONFERENCE ON COMPUTING EDUCATION RESEARCH. Koli Finland: ACM, 24 nov. 2016b. Disponível em: <https://dl.acm.org/doi/10.1145/2999541.2999562>. Acesso em: 18 abr. 2022

SOHN, B.-S.; JUNG, J. E. A Novel Ranking Model for a Large-Scale Scientific Publication. 2015.

THOMPSON, D., K. M. ;GARRISON, K. ;SAN TEL ICES-WERCHEZ, C. ;ARELLANO-ROJAS, P. ;REYES-LILLO. Literatura sobre ?Bibliotecología y Ciencias de la Información? en Web of Science: Qué nos dice una década sobre la colaboración académica en el campo (2007-2016). **e-Ciencias de la Información (Costa Rica)**, v. v. 10, 2020.

TORRES-PASCUAL, C.; SÁNCHEZ-PÉREZ, H. J.; ÀVILA-CASTELLS, P. Distribución geográfica y colaboración internacional de las publicaciones científicas latinoamericanas y del Caribe sobre tuberculosis en PubMed. **Revista Peruana de Medicina Experimental y Salud Pública**, v. 38, n. 1, p. 49–57, 26 mar. 2021.

TORRES-SALINAS, D.; ROBINSON-GARCIA, N.; JIMÉNEZ-CONTRERAS, E. Can we use altmetrics at the institutional level? A case study analysing the coverage by research areas of four Spanish universities. p. 8, 2016.

URTEAGA, E. SOCIOLOGÍA DE LAS PROFESIONES: UNA TEORÍA DE LA COMPLEJIDAD. **Revista de Relaciones Laborales**, v. 18, p. 169–198, 2008.

VALLES, M. et al. **An Altmetric Alternative for Measuring the Impact of University Institutional Repositories' Grey Literature**. . Em: INTERNATIONAL CONFERENCE ON DATA AND INFORMATION IN ONLINE. Springer, 2020.

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

VANZ, S. A. S. Redes Colaborativas nos Estudos Métricos de Ciência e Tecnologia ? Collaborative Networks in Metric Studies of Science and Technology. **Liinc em revista**, v. 9, 2013.

VESSURI, H. M. The social study of science in Latin America. **Social Studies of Science**, v. 17, n. 3, p. 519–554, 1987.

WANG, H. et al. **A Data-driven Productivity Assessment Framework for Collaborative Research Teams**. 2021 The 5th International Conference on Compute and Data Analysis. **Anais**... Em: ICCDA 2021: 2021 THE 5TH INTERNATIONAL CONFERENCE ON COMPUTE AND DATA ANALYSIS. Sanya China: ACM, 2 fev.

2021. Disponível em: <https://dl.acm.org/doi/10.1145/3456529.3456530>. Acesso em: 4 abr. 2022

YANG, C. C.; TANG, X. **A content and social network approach of bibliometrics analysis across domains**. Proceedings of the 2012 iConference on - iConference '12. **Anais**... Em: THE 2012 ICONFERENCE. Toronto, Ontario, Canada: ACM Press, 2012. Disponível em: <http://dl.acm.org/citation.cfm?doid=2132176.2132270>. Acesso em: 23 abr. 2022

YUAN, L. et al. Who are the international research collaboration partners for China? A novel data perspective based on NSFC grants. **Scientometrics**, v. 116, n. 1, p. 401–422, jul. 2018.

ZHANG, B. et al. **The Influence of Author Degree Centrality and L-Index on Scientific Performance of Physical Education and Training Papers in China Based on the Perspective of Social Network Analysis**. **COMPLEXITY**ADAM HOUSE, 3RD FL, 1 FITZROY SQ, LONDON, WIT 5HE, ENGLANDWILEY-HINDAWI, , 30 set. 2021a.

ZHANG, C.-T. The e-Index, Complementing the h-Index for Excess Citations. **PLoS ONE**, v. 4, n. 5, p. e5429, 5 maio 2009.

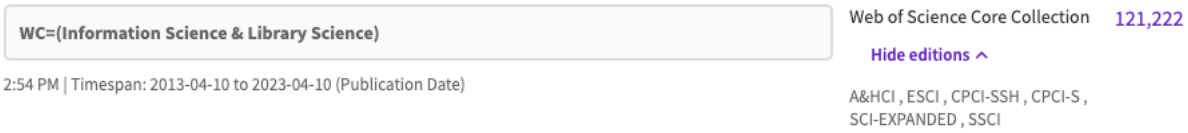
ZHANG, J. et al. Scientific Collaboration Network Analysis for Computing Education Conferences. Em: **Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1**. New York, NY, USA: Association for Computing Machinery, 2021b. p. 582–588.

ZIPF, G.-K. Human behavior and the principle of least effort. **Addison-Wesley: Cambridge Mass**, p. 543, 1949.

APÊNDICE A - Coleta dos dados

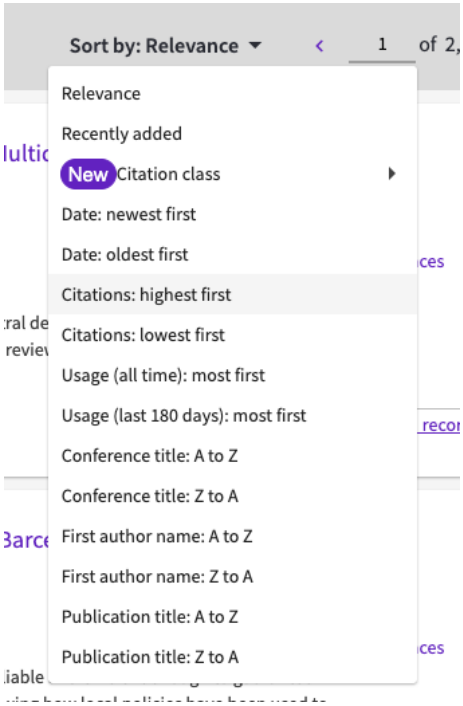
Os dados presentes nesta análise foram coletados em 10 de abril de 2023, coletados da *Web Of Science*. Para extração das informações, foi selecionado o Tópico *Information Science & Library Science* com controle de data dos últimos 10 anos, ou seja, do período entre 10/04/2013 e 10/04/2023, sendo recuperados 121.222 registros no total (Figura 60).

Figura 60 - Pesquisa com número de resultados na *Web of Science*



Fonte: https://www-webofscience.ez46.periodicos.capes.gov.br/wos/woscc/basic-search, acesso em 10 de abril de 2023.

Figura 61 - Ordenação dos resultados na *Web of Science*



Fonte: https://www-webofscience.ez46.periodicos.capes.gov.br/wos/woscc/basic-search, acesso em 10 de abril de 2023.

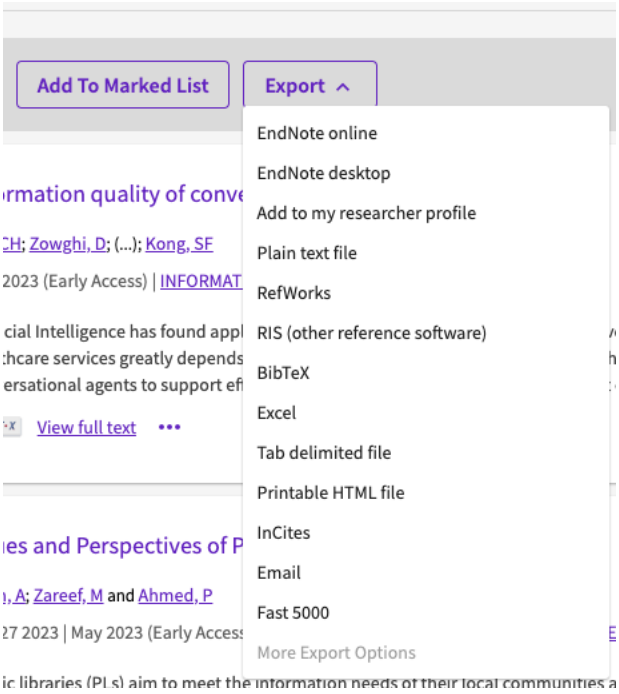
Optamos por trazer para análise somente os artigos com ao menos uma citação. A remoção dos artigos sem citação ocorreu pois neste trabalho o índice h é um dos principais indicadores utilizados, também para facilitação na apresentação e

avaliação dos resultados,. Desta forma, realizamos uma ordenação por total de citações, conforme apresentado na Figura 61.

Foi utilizada a ferramenta própria da plataforma *Web of Science* para exportação dos resultados (Figura 62). A exportação possui limitação de 5 mil registros (Figura 63), por este motivo, foram necessárias diversas exportações até trazer todos os artigos com ao menos uma citação cada.

Apesar da mensagem descrita na Figura 63 informando que os registros conteriam somente Autor, Título e Fonte, os resultados foram trazidos de forma completa. Cada registro possui informações de autores, afiliações, título, citações, endereços, local de publicação, ano de publicação, dentre outras informações relevantes.

Figura 62 - Opções de exportação na *Web of Science*

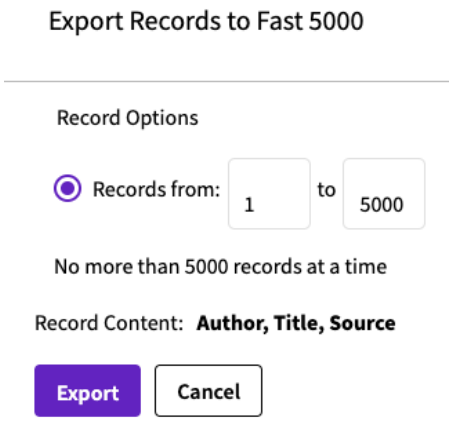


Fonte: https://www-webofscience.ez46.periodicos.capes.gov.br/wos/woscc/basic-search, acesso em 10 de abril de 2023.

As tabelas exportadas do *Web of Science* foram organizadas em uma única planilha do Excel (Figura 64), contendo aproximadamente 55 mil registros. Esta tabela única foi então importada para um banco de dados PostgreSQL, através do comando COPY. Nesse comando, a planilha no formato .csv é importada para uma tabela no

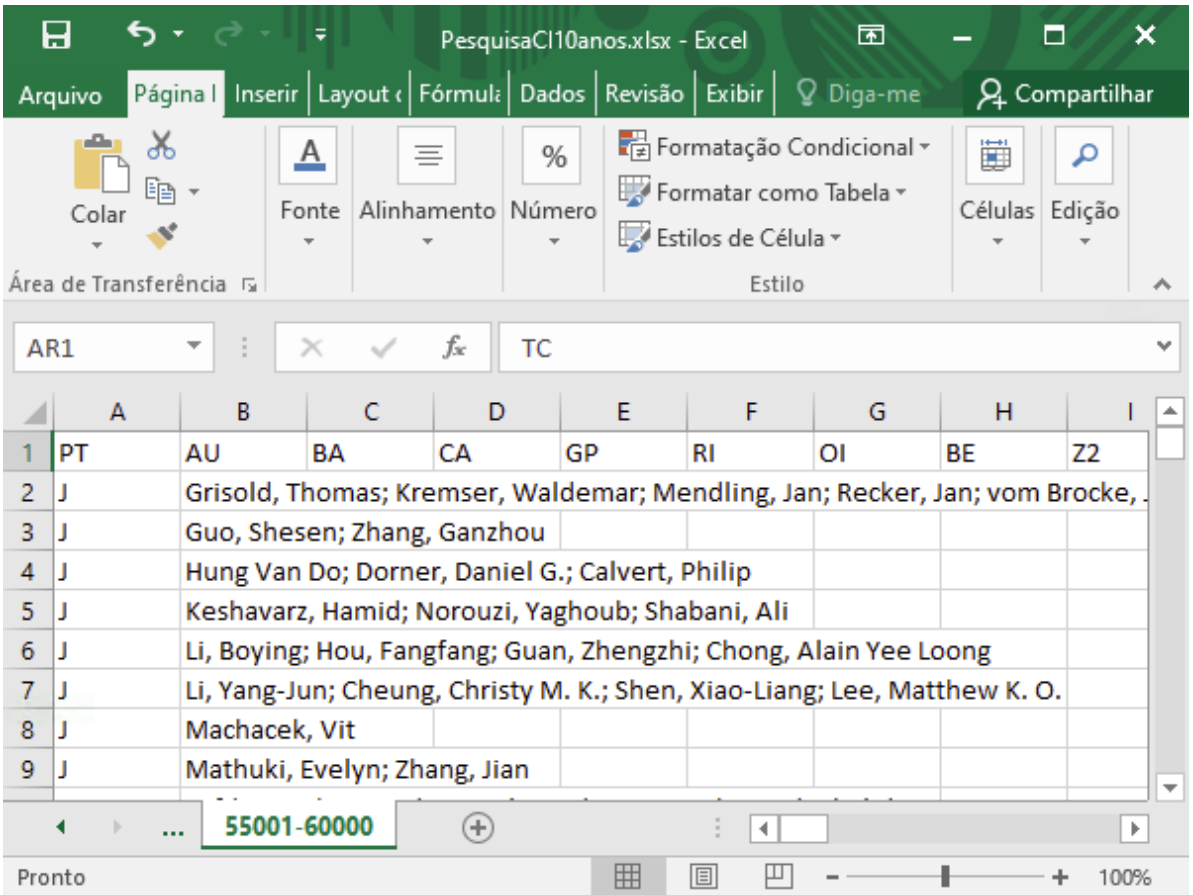
banco que possui colunas com nome e tipos de dados correspondentes com a entrada .csv.

Figura 63 - Exportação Fast 500 na *Web of Science*



Fonte: https://www-webofscience.ez46.periodicos.capes.gov.br/wos/woscc/basic-search, acesso em 10 de abril de 2023.

Figura 64 - Dados importados para o Excel



Fonte: dados da pesquisa (2023).

APÊNDICE B – Código em SQL para Tratamento dos Dados

```
--Criando a tabela que irá receber os valores do arquivo CSV exportado da
Web of Science
drop table a;
create table a
(PT varchar(40000),
AU varchar(2000),
BA varchar(1250),
CA varchar(1250),
GP varchar(1250),
RI varchar(1250),
OI varchar(1250),
BE varchar(1250),
Z2 varchar(1250),
TI varchar(2000),
X1 varchar(1250),
Y1 varchar(1250),
Z1 varchar(1250),
FT varchar(1250),
PN varchar(1250),
AE varchar(1250),
Z3 varchar(100),
SO varchar(1250),
S1 varchar(1250),
SE varchar(1250),
BS varchar(1250),
VL varchar(1250),
ISI varchar(1250),
SI varchar(1250),
MA varchar(1250),
BP varchar(1250),
EP varchar(1250),
AR varchar(1250),
DI varchar(1250),
D2 varchar(1250),
SU varchar(1250),
PD varchar(1250),
PY varchar(1250),
AB varchar(20000),
X4 varchar(1250),
C1 varchar(10000),
Y4 varchar(1250),
Z4 varchar(1250),
AK varchar(1250),
CT varchar(1250),
CY varchar(1250),
SP varchar(1250),
CL varchar(1250),
TC varchar(1250),
Z8 varchar(1250),
ZB varchar(1250),
ZS varchar(1250),
Z9 varchar(1250),
SN varchar(1250),
BN varchar(1250),
WC varchar(1250),
UT varchar(1250),
PM varchar(1250)
);
```

```
-- Criando a tabela de autores
drop table author;
create table author
(id serial primary key, name varchar(250))

--Inserindo os valores individuais de autores, que são originalmente uma
lista,
--SEPARANDO OS DIFERENTES AUTORES POR ;
insert into author (name)
select SPLIT_PART(au, ';', 1) from a;
insert into author (name)
select SPLIT_PART(au, ';', 2) from a;
insert into author (name)
select SPLIT_PART(au, ';', 3) from a;
insert into author (name)
select SPLIT_PART(au, ';', 4) from a;
insert into author (name)
select SPLIT_PART(au, ';', 5) from a;
insert into author (name)
select SPLIT_PART(au, ';', 6) from a;
insert into author (name)
select SPLIT_PART(au, ';', 7) from a;
insert into author (name)
select SPLIT_PART(au, ';', 8) from a;
insert into author (name)
select SPLIT_PART(au, ';', 9) from a;
insert into author (name)
select SPLIT_PART(au, ';', 10) from a;
insert into author (name)
select SPLIT_PART(au, ';', 11) from a;
insert into author (name)
select SPLIT_PART(au, ';', 12) from a;
insert into author (name)
select SPLIT_PART(au, ';', 13) from a;
insert into author (name)
select SPLIT_PART(au, ';', 14) from a;
insert into author (name)
select SPLIT_PART(au, ';', 15) from a;
insert into author (name)
select SPLIT_PART(au, ';', 16) from a;
insert into author (name)
select SPLIT_PART(au, ';', 17) from a;
insert into author (name)
select SPLIT_PART(au, ';', 18) from a;
insert into author (name)
select SPLIT_PART(au, ';', 19) from a;
insert into author (name)
select SPLIT_PART(au, ';', 20) from a;
insert into author (name)
select SPLIT_PART(au, ';', 21) from a;
insert into author (name)
select SPLIT_PART(au, ';', 22) from a;
insert into author (name)
select SPLIT_PART(au, ';', 23) from a;
insert into author (name)
select SPLIT_PART(au, ';', 24) from a;
insert into author (name)
select SPLIT_PART(au, ';', 25) from a;
insert into author (name)
select SPLIT_PART(au, ';', 26) from a;
```

```
insert into author (name)
select SPLIT_PART(au, ';', 27) from a;
insert into author (name)
select SPLIT_PART(au, ';', 28) from a;
insert into author (name)
select SPLIT_PART(au, ';', 29) from a;
insert into author (name)
select SPLIT_PART(au, ';', 30) from a;
insert into author (name)
select SPLIT_PART(au, ';', 31) from a;
insert into author (name)
select SPLIT_PART(au, ';', 32) from a;
insert into author (name)
select SPLIT_PART(au, ';', 33) from a;
insert into author (name)
select SPLIT_PART(au, ';', 34) from a;
insert into author (name)
select SPLIT_PART(au, ';', 35) from a;
insert into author (name)
select SPLIT_PART(au, ';', 36) from a;
insert into author (name)
select SPLIT_PART(au, ';', 37) from a;
insert into author (name)
select SPLIT_PART(au, ';', 38) from a;
insert into author (name)
select SPLIT_PART(au, ';', 39) from a;
insert into author (name)
select SPLIT_PART(au, ';', 40) from a;
insert into author (name)
select SPLIT_PART(au, ';', 41) from a;
insert into author (name)
select SPLIT_PART(au, ';', 42) from a;
insert into author (name)
select SPLIT_PART(au, ';', 43) from a;
insert into author (name)
select SPLIT_PART(au, ';', 44) from a;
insert into author (name)
select SPLIT_PART(au, ';', 45) from a;
insert into author (name)
select SPLIT_PART(au, ';', 46) from a;
insert into author (name)
select SPLIT_PART(au, ';', 47) from a;
insert into author (name)
select SPLIT_PART(au, ';', 48) from a;
insert into author (name)
select SPLIT_PART(au, ';', 49) from a;
insert into author (name)
select SPLIT_PART(au, ';', 50) from a;
insert into author (name)
select SPLIT_PART(au, ';', 51) from a;
insert into author (name)
select SPLIT_PART(au, ';', 52) from a;
insert into author (name)
select SPLIT_PART(au, ';', 53) from a;
insert into author (name)
select SPLIT_PART(au, ';', 54) from a;
insert into author (name)
select SPLIT_PART(au, ';', 55) from a;
insert into author (name)
select SPLIT_PART(au, ';', 56) from a;
insert into author (name)
```

```
select SPLIT_PART(au, ';', 57) from a;
insert into author (name)
select SPLIT_PART(au, ';', 58) from a;
insert into author (name)
select SPLIT_PART(au, ';', 59) from a;
insert into author (name)
select SPLIT_PART(au, ';', 60) from a;
insert into author (name)
select SPLIT_PART(au, ';', 61) from a;
insert into author (name)
select SPLIT_PART(au, ';', 62) from a;
insert into author (name)
select SPLIT_PART(au, ';', 63) from a;
insert into author (name)
select SPLIT_PART(au, ';', 64) from a;
insert into author (name)
select SPLIT_PART(au, ';', 65) from a;
insert into author (name)
select SPLIT_PART(au, ';', 66) from a;
insert into author (name)
select SPLIT_PART(au, ';', 67) from a;
insert into author (name)
select SPLIT_PART(au, ';', 68) from a;
insert into author (name)
select SPLIT_PART(au, ';', 69) from a;
insert into author (name)
select SPLIT_PART(au, ';', 70) from a;
insert into author (name)
select SPLIT_PART(au, ';', 71) from a;
insert into author (name)
select SPLIT_PART(au, ';', 72) from a;
insert into author (name)
select SPLIT_PART(au, ';', 73) from a;
insert into author (name)
select SPLIT_PART(au, ';', 74) from a;
insert into author (name)
select SPLIT_PART(au, ';', 75) from a;
insert into author (name)
select SPLIT_PART(au, ';', 76) from a;
insert into author (name)
select SPLIT_PART(au, ';', 77) from a;
insert into author (name)
select SPLIT_PART(au, ';', 78) from a;
insert into author (name)
select SPLIT_PART(au, ';', 79) from a;
insert into author (name)
select SPLIT_PART(au, ';', 80) from a;
insert into author (name)
select SPLIT_PART(au, ';', 81) from a;
insert into author (name)
select SPLIT_PART(au, ';', 82) from a;
insert into author (name)
select SPLIT_PART(au, ';', 83) from a;
insert into author (name)
select SPLIT_PART(au, ';', 84) from a;
insert into author (name)
select SPLIT_PART(au, ';', 85) from a;
insert into author (name)
select SPLIT_PART(au, ';', 86) from a;
insert into author (name)
select SPLIT_PART(au, ';', 87) from a;
```

```
insert into author (name)
select SPLIT_PART(au, ';', 88) from a;
insert into author (name)
select SPLIT_PART(au, ';', 89) from a;
insert into author (name)
select SPLIT_PART(au, ';', 90) from a;
insert into author (name)
select SPLIT_PART(au, ';', 91) from a;
insert into author (name)
select SPLIT_PART(au, ';', 92) from a;
insert into author (name)
select SPLIT_PART(au, ';', 93) from a;
insert into author (name)
select SPLIT_PART(au, ';', 94) from a;
insert into author (name)
select SPLIT_PART(au, ';', 95) from a;
insert into author (name)
select SPLIT_PART(au, ';', 96) from a;
insert into author (name)
select SPLIT_PART(au, ';', 97) from a;
insert into author (name)
select SPLIT_PART(au, ';', 98) from a;
insert into author (name)
select SPLIT_PART(au, ';', 99) from a;

select SPLIT_PART(au, ';', 99) from a order by 1 desc; -- " Schrader, P.
G." 99

--CRIANDO UMA TABELA ONDE CADA NOME DE AUTOR TEM SOMENTE UMA ENTRADA
alter table author rename to author1;

create table author
(id serial primary key, name varchar(250));

insert into author (name)
select distinct trim(name) from author1
where name is not null and name <> ''
order by 1;

select count(1)  from author -- TOTAL: 95844 diferentes autores

-- Criando uma nova tabela (pesquisaa) trazendo somente as colunas que
serão utilizadas na análise:
drop table pesquisaa
create table pesquisaa as
select TI AS titulo,
AU as listaAutores,
SO as revista,
C1 as afiliacoes,
TC as citacoes,
PY as ano,
WC as areas
from a;

ALTER TABLE pesquisaa ADD COLUMN id SERIAL PRIMARY KEY;

--ASSOCIANDO OS AUTORES COM O ARTIGOS
select count(*) from pesquisaa

drop table rl_autor_artigo;
--ATENCAO ESSA QUERY LEVA 30 MINUTOS PARA EXECUTAR
```

```
create table rl_autor_artigo as
select p.id as idArtigo, au.id as idAutor
from pesquisaa p
join author au on p.listaautores like '%'||au.name||'%';

CREATE INDEX rl_autor_artigo_idartigo
   ON public.rl_autor_artigo(idartigo);

CREATE INDEX rl_autor_artigo_idauthor
   ON public.rl_autor_artigo(idauthor);

alter table author rename to author2;
create table author as select id, name as original_name, SPLIT_PART(name,
',', 1) as surname, SPLIT_PART(name, ',', 2) as name from author2;

alter table author add column sufix varchar(30) null
update author set sufix = SPLIT_PART(original_name, ',', 3)

--TRATANDO OS NOMES DOS AUTORES
--removendo traços dos nomes

select * from author where name like '%-%'

select au.original_name, r.idArtigo, ar.*
from rl_autor_artigo r
join author au on au.id = r.idAutor
join pesquisaa ar on ar.id = r.idArtigo
where idArtigo = 28685
order by 1

select *
from author a1
join author a2
   on a1.id <> a2.id
   and (a1.name||a1.surname||a1.sufix =
replace((a2.name||a2.surname||a2.sufix), '-', ''))
   --and a2.original_name = replace(a1.original_name, '-', '')
order by a1.original_name

select *
from author a1
join author a2
   on a1.id > a2.id
   and replace((a1.name||a1.surname||a1.sufix), '-', ' ') =
replace((a2.name||a2.surname||a2.sufix), '-', ' ')
   --and a2.original_name = replace(a1.original_name, '-', '')
order by a1.id

update author a1 set id_homonimo = a2.id
from author a2
where a1.id > a2.id
and replace((a1.name||a1.surname||a1.sufix), '-', ' ') =
replace((a2.name||a2.surname||a2.sufix), '-', '')

alter table rl_autor_artigo add column idauthor integer;

update rl_autor_artigo set idauthor = coalesce(a.id_homonimo, a.id)
from author a where a.id = idautor

alter table rl_autor_artigo rename idautor to
idauthor_DESOCONSIDERA_HOMONIMO
```

```
alter table rl_autor_artigo add constraint rl_autor_artigo_PK primary key
(idartigo,idauthor)

select * from author

select a.*
from rl_autor_artigo r
join author a on a.id = r.idauthor
join artigos ar on ar.id = r.idartigo
where r.idartigo = 1

select a.name, b.name
from author a,
 author b
where a.name like '%'||b.name||'%' and a.id <> b.id

select listaAutores from artigos where listaAutores like '%-%-%'

delete from rl_autor_artigo where idautor = 1 --removendo autor em branco

--TRATAMENTO DAS AFILIAÇÕES, para crias as tabelas Institution e
Unibversity
select listaautores, afiliacoes from pesquisaa limit 10

create table rl_artigo_afiliacao as
select p.id as idArtigo, au.id as idAutor
from pesquisaa p
join author au on p.listaautores like '%'||au.name||'%';

drop table affiliation
create table affiliation
(id serial primary key, affiliation varchar(2000))
--SEPARANDO OS DIFERENTES afiliacoes POR ;

insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 1) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 2) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 3) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 4) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 5) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 6) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 7) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 8) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 9) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 10) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 11) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 12) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 13) from pesquisaa;
```

```
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 14) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 15) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 16) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 17) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 18) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 19) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 20) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 21) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 22) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 23) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 24) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 25) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 26) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 27) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 28) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 29) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 30) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 31) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 32) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 33) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 34) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 35) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 36) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 37) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 38) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 39) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 40) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 41) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 42) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 43) from pesquisaa;
insert into affiliation (affiliation)
```

```
select SPLIT_PART(afiliacoes, '; [', 44) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 45) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 46) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 47) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 48) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 49) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 50) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 51) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 52) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 53) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 54) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 55) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 56) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 57) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 58) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 59) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 60) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 61) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 62) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 63) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 64) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 65) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 66) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 67) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 68) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 69) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 70) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 71) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 72) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 73) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 74) from pesquisaa;
```

```
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 75) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 76) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 77) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 78) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 79) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 80) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 81) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 82) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 83) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 84) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 85) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 86) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 87) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 88) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 89) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 90) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 91) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 92) from pesquisaa;
insert into affiliation (affiliation)
select SPLIT_PART(afiliacoes, '; [', 93) from pesquisaa;

select * from affiliation

alter table affiliation rename to affiliation1;

create table affiliation
(id serial primary key, affiliation varchar(2000));

insert into affiliation (affiliation)
select distinct trim(affiliation) from affiliation1
where affiliation is not null and affiliation <> ''
order by 1;

select *, SPLIT_PART(affiliation, ']', 2) from affiliation
alter table affiliation add column institution varchar(2000)
update affiliation set institution = SPLIT_PART(affiliation, ']', 2)

select affiliation  from affiliation where institution = ''
update affiliation Institution = affiliation where institution = ''
select distinct institution from affiliation

create table institution
(id serial primary key, institution varchar(2000))
```

```
insert into institution (institution)
select SPLIT_PART(institution, ';', 1) from affiliation where
SPLIT_PART(institution, ';', 1) <> '';
insert into institution (institution)
select SPLIT_PART(institution, ';', 2) from affiliation where
SPLIT_PART(institution, ';', 2) <> '';
insert into institution (institution)
select SPLIT_PART(institution, ';', 3) from affiliation where
SPLIT_PART(institution, ';', 3) <> '';
insert into institution (institution)
select SPLIT_PART(institution, ';', 4) from affiliation where
SPLIT_PART(institution, ';', 4) <> '';
insert into institution (institution)
select SPLIT_PART(institution, ';', 5) from affiliation where
SPLIT_PART(institution, ';', 5) <> '';
insert into institution (institution)
select SPLIT_PART(institution, ';', 6) from affiliation where
SPLIT_PART(institution, ';', 6) <> '';
insert into institution (institution)
select SPLIT_PART(institution, ';', 7) from affiliation where
SPLIT_PART(institution, ';', 7) <> '';

alter table institution add column country varchar(100) null
update institution set country = SPLIT_PART(institution, ',', 11) where
SPLIT_PART(institution, ',', 11) <> '';
update institution set country = SPLIT_PART(institution, ',', 10) where
SPLIT_PART(institution, ',', 10) <> '' and country is null;
update institution set country = SPLIT_PART(institution, ',', 9) where
SPLIT_PART(institution, ',', 9) <> '' and country is null;
update institution set country = SPLIT_PART(institution, ',', 8) where
SPLIT_PART(institution, ',', 8) <> '' and country is null;
update institution set country = SPLIT_PART(institution, ',', 7) where
SPLIT_PART(institution, ',', 7) <> '' and country is null;
update institution set country = SPLIT_PART(institution, ',', 6) where
SPLIT_PART(institution, ',', 6) <> '' and country is null;
update institution set country = SPLIT_PART(institution, ',', 5) where
SPLIT_PART(institution, ',', 5) <> '' and country is null;
update institution set country = SPLIT_PART(institution, ',', 4) where
SPLIT_PART(institution, ',', 4) <> '' and country is null;
update institution set country = SPLIT_PART(institution, ',', 3) where
SPLIT_PART(institution, ',', 3) <> '' and country is null;
update institution set country = SPLIT_PART(institution, ',', 2) where
SPLIT_PART(institution, ',', 2) <> '' and country is null;
update institution set country = SPLIT_PART(institution, ',', 1) where
SPLIT_PART(institution, ',', 1) <> '' and country is null;

select count(1), country from institution
group by country order  by 1 desc

select * from institution where country like 'USA%'

update institution set country = 'USA' where country like '%USA'

select SPLIT_PART(institution, ',', 10) , institution from institution
where SPLIT_PART(institution, ',', 1) <> '' and country is null

select distinct institution from institution

alter table institution rename to institution1;
```

```
create table institution
(id serial primary key, institution varchar(2000),  country varchar(100))

--CRIANDO UMA TABELA ONDE CADA NOME DE instituicao TEM SOMENTE UMA ENTRADA
insert into institution (institution, country)
select distinct trim(institution), trim(country) from institution1
where institution is not null and institution <> ''
order by 1;

--ATENCAO ESSA QUERY LEVA 1h15MIN PRA EXECUTAR
create table rl_instituicao_artigo as
select p.id as idArtigo, i.id as idInstitution
from pesquisaa p
join institution i on p.afiliacoes like '%'||i.institution||'%'

select
count(distinct p.id) as total_papers,
count(distinct aa.idAuthor) as total_authors,
cast(count(distinct aa.idAuthor) as float)/count(distinct p.id) as
avg_authors_paper, -- aqui é média de autores por artigo, não o número
médio de autores
i.country
from pesquisaa p
join rl_instituicao_artigo r on p.id = r.idArtigo
join institution i on i.id = r.idInstitution
join rl_autor_artigo aa on p.id = aa.idArtigo
group by i.country
order by 1 desc

--Incluindo coluna numAuthors contendo o número de autores em cada artigo:
select distinct
p.id,
(select count(distinct idAuthor) from rl_autor_artigo aa where aa.idArtigo
= p.id) as numAutores
from pesquisaa p;

alter table pesquisaa add column numAuthors integer

update pesquisaa set numAuthors = (select count(distinct idAuthor) from
rl_autor_artigo aa where aa.idArtigo = pesquisaa.id)

--Incluindo colunas de citações com valor inteiro, pois na importação o
valor veio como varchar
select citacoes, * from pesquisaa
alter table pesquisaa add column citacoes1 integer
update pesquisaa set citacoes1 = cast(citacoes as integer)
alter table pesquisaa rename column citacoes to citacoes_varchar
alter table pesquisaa rename column citacoes1 to citacoes

select max(citacoes) from pesquisaa

create table country as
select
ROW_NUMBER() OVER(ORDER BY country) as id,
country
from (
select distinct
country
from institution
order by country) t
```

```
CREATE INDEX country_pk
   ON public.country(id);

--Cálculo do índice H por SQL, query base extraída do stackoverflow e
adaptada
--https://stackoverflow.com/questions/18786156/sql-for-computing-h-score-h-
index
select
   country,
-- max(ranking) as indice_h,
   (select avg(pa.numAuthors)
      from pesquisaa pa
      where pa.id in (
         select distinct p.id
         from pesquisaa p
         join rl_instituicao_artigo r on p.id = r.idArtigo
         join institution i on i.id = r.idInstitution
         where numauthors is not null and country = t.country)
      and pa.id = t.id
   ) as media_autores
from
   (select i.country, p.id, p.numAuthors, sum(p.citacoes) as
citations_count,
   rank() over (partition by i.country order by sum(p.citacoes) desc) as
ranking
   from pesquisaa p
   join rl_instituicao_artigo r on p.id = r.idArtigo
   join institution i on i.id = r.idInstitution
   where numAuthors >= 2
   group by i.country, p.id) as t
--where ranking <= citations_count
--group by country
order by 2 desc
limit 15

--A media de autores nos EUA cai para 3,90 quando o numero de top citacoes
e levado em conta
--Existem 343 artigos com ao menos 343 citacoes, dando um indice h de 343
para EUA, para artigos com ao menos 2 autores
select country,
count(*) as h_index,
avg(numAuthors) as avg_autores,
(select avg(pa.numAuthors)
   from pesquisaa pa
   where pa.id in (
      select distinct p.id
      from pesquisaa p
      join rl_instituicao_artigo r on p.id = r.idArtigo
      join institution i on i.id = r.idInstitution
      where numauthors is not null and country = t.country)
) as media_autores
from (select i.country, p.id, p.numAuthors, sum(p.citacoes) as
citations_count,
rank() over (partition by i.country order by sum(p.citacoes) desc) as
ranking
from pesquisaa p
join rl_instituicao_artigo r on p.id = r.idArtigo
join institution i on i.id = r.idInstitution
where numAuthors >= 2
group by i.country, p.id) t
where ranking <= citations_count
```

```
group by country
order by 2 desc
limit 15


CREATE MATERIALIZED VIEW vi_pais
AS
 SELECT i.country,
    count(DISTINCT p.id) AS total_papers,
    count(DISTINCT aa.idauthor) AS total_authors,
    ( SELECT avg(pa.numauthors) AS avg
           FROM pesquisaa pa
          WHERE (pa.id IN ( SELECT p_1.id
                   FROM ((pesquisaa p_1
                     JOIN rl_pais_artigo r_1 ON ((r_1.idartigo = p_1.id)))
                     JOIN country c ON ((c.id = r_1.idcountry)))
                  WHERE ((c.country)::text = (i.country)::text)))) AS
average_numauthors_paper,
    ( SELECT avg(r_1.numauthors)
          FROM ((pesquisaa p_1
           JOIN rl_pais_artigo r_1 ON ((r_1.idartigo = p_1.id)))
           JOIN country c ON ((c.id = r_1.idcountry)))
        WHERE c.country = i.country) AS average_numauthors_country,
    ( SELECT mode() WITHIN GROUP (ORDER BY pa.numauthors) AS modal_value
           FROM pesquisaa pa
          WHERE (pa.id IN ( SELECT p_1.id
                   FROM ((pesquisaa p_1
                     JOIN rl_pais_artigo r_1 ON ((r_1.idartigo = p_1.id)))
                     JOIN country c ON ((c.id = r_1.idcountry)))
                  WHERE ((c.country)::text = (i.country)::text)))) AS
mode_numauthors_paper,
    ( SELECT mode() WITHIN GROUP (ORDER BY r_1.numauthors) AS modal_value
          FROM ((pesquisaa p_1
           JOIN rl_pais_artigo r_1 ON ((r_1.idartigo = p_1.id)))
           JOIN country c ON ((c.id = r_1.idcountry)))
        WHERE r_1.numauthors > 1 AND c.country = i.country) AS
mode_numauthors_country,
    ( SELECT sum(pa.citacoes) AS sum
           FROM pesquisaa pa
          WHERE pa.numauthors > 1 and (pa.id IN ( SELECT p_1.id
                   FROM ((pesquisaa p_1
                     JOIN rl_pais_artigo r_1 ON ((r_1.idartigo = p_1.id)))
                     JOIN country c ON ((c.id = r_1.idcountry)))
                  WHERE ((c.country)::text = (i.country)::text)))) AS
sum_citacoes,
    ( SELECT avg(pa.citacoes) AS avg
           FROM pesquisaa pa
          WHERE (pa.id IN ( SELECT p_1.id
                   FROM ((pesquisaa p_1
                     JOIN rl_pais_artigo r_1 ON ((r_1.idartigo = p_1.id)))
                     JOIN country c ON ((c.id = r_1.idcountry)))
                  WHERE ((c.country)::text = (i.country)::text)))) AS
avg_citacoes
   FROM (((pesquisaa p
     JOIN rl_pais_artigo r ON ((p.id = r.idartigo)))
     JOIN country i ON ((i.id = r.idcountry)))
     JOIN rl_autor_artigo aa ON ((p.id = aa.idartigo)))
  GROUP BY i.country
  ORDER BY i.country;
```

```
--índice H por país considerando artigos cujo número de autores é maior ou
igual à média de autores em cada país
--Ou seja, artigos que foram realizados em conjunto, por um grupo,
respeitando o número de autores característico de cada país.
select country, count(*) as h_index, avg(numAuthors)
from
(select i.country, p.id, p.numAuthors, sum(p.citacoes) as citations_count,
vp.average_numAuthors,
rank() over (partition by i.country order by sum(p.citacoes) desc) as
ranking
from pesquisaa p
join rl_instituicao_artigo r on p.id = r.idArtigo
join institution i on i.id = r.idInstitution
join vi_pais vp on vp.country = i.country
group by i.country, vp.average_numAuthors, p.id) t
where ranking <= citations_count
and numAuthors >= average_numAuthors
group by country
order by 2 desc

--Seção apresentando a coleta e normalização dos dados

--Criar subseção Análise da amostra, com gráficos com total de artigos e
citações por país, apresentando média e mediana de número de autores geral
e por país

--Criar subseção sobre análise do grau de colaboração, colocando os valores
com diferentes graus de colaboração, plotando em gráfico.
--Mostrando, após, o índice h do grupo de países considerando o grau de
colaboração da média de autores em cada país

--Seção com análise de dados por país, trazendo o grau de colaboração por
país. Nos 5 primeiros países analisar também o grau por instituição

select country, count(*) as h_index, avg(numAuthors)
from (select i.country, p.id, p.numAuthors, sum(p.citacoes) as
citations_count, vp.average_numAuthors,
rank() over (partition by i.country order by sum(p.citacoes) desc) as
ranking
from pesquisaa p
join rl_instituicao_artigo r on p.id = r.idArtigo
join institution i on i.id = r.idInstitution
join vi_pais vp on vp.country = i.country
WHERE numAuthors >= average_numAuthors/2
group by i.country, vp.average_numAuthors, p.id) t
and numAuthors >= average_numAuthors
group by country
order by 2 desc

select * from vi_pais

select *,
   case when numauthors > 1 then 'Group' else 'Individual' end as isGroup,
   case when numauthors > 3 then 'AboveAverage' else 'BelowAverage' end as
gropSizeAverage
from pesquisaa

select avg(numauthors) from pesquisaa
select avg(pesquisaa.citacoes) from pesquisaa

select --numauthors
```

```
count(1)
from pesquisaa where numauthors between 2 and 4

update artigos set citacoes = 0 where citacoes is null

select count(1), citacoes
from pesquisaa
group by citacoes
order by citacoes

select count(1), numauthors
from pesquisaa
group by numauthors
order by numauthors

select * from pesquisaa limit 10

-----------ANÁLISE POR INSTITUICAO----------

select * from institution limit 10

select SPLIT_PART(i.institution, ',', 1),
count(1), i.*
from institution i
join rl_instituicao_artigo r on r.idinstitution = i.id
where i.institution like '%UFSC,%'
   or i.institution like '%Federal%Unive%Santa%Catarina%'
   or i.institution like '%Universidade%Federal%Santa%Catarina%'
group by i.id
order by 1 desc

alter table institution1 add column university varchar(200) null

update institution1 set university = SPLIT_PART(institution, ',', 1) where
SPLIT_PART(institution, ',', 1) <> '';

select count(distinct r.idartigo) diff_affiliations, university, country
from
institution i
join rl_instituicao_artigo r on r.idinstitution = i.id
group by university, country
order by 1 desc

--CRIANDO UMA TABELA ONDE CADA NOME DE universidade TEM SOMENTE UMA ENTRADA
create table university
(id serial primary key, university varchar(2000),  country varchar(100))

insert into university(university, country)
select distinct trim(university), trim(country)
from institution
where university is not null and university <> ''
order by 1;

select * from institution where university = 'Univ Amsterdam' and country =
'Turkey'

--Removendo universidades sem nome completo que possuíam valores que
atrapalhavam as análises.
delete from university
where university
in ('Informat Syst','Sch Informat', 'Univ Sci & Technol', 'Dept
```

```
Management', 'UN', '1', 'Univ', 'Univ Technol',
   'Informat Syst','Sch Informat','Dept Informat','Coll Business','Sch
Management','State Univ','40','Natl Univ','16',
   'Dept Informat Management','200','51','Univ Sci','17','Dept Comp
Sci','Christ','Phi','Lib & Informat Sci','IT','Dept Geog',
   'Tech Univ','Sch Med','Univ Lib','Tech Univ','Sch Med','Univ Lib', 'Dept
Med', 'City Univ')

delete from university where university in ('Sch Commun')

alter table author add column affiliation varchar(2000);
alter table author add column institution varchar(2000);

update author
set affiliation = af.affiliation, institution = af.institution
from affiliation af
where af.affiliation like '%'||original_name||'%'

select * from author
where affiliation is null
--1273 - Casos em que o autor aparece na lista de autores, mas não estão na
lista de afiliações

alter table author add column country varchar(100);
alter table author add column university varchar(200);

update author set institution = ltrim(institution);

update author
set country = i.country, university = i.university
from institution i
where i.institution = author.institution;

alter table author add column idcountry integer;
alter table author add column iduniversity integer;

update author
set idcountry = c.id
from country c
where c.country = author.country;

update author
set iduniversity = u.id
from university u
where u.university = author.university;

--Criando o relacionamento entre país e artigo,
--o relacionamento foi feito a partir das afiliações para evitar falsos
positivos.
drop table rl_pais_artigo
create table rl_pais_artigo as
select distinct p.id as idArtigo, c.id as idCountry
from pesquisaa p
join institution i on p.afiliacoes like '%'||i.institution||'%'
join country c on c.country = i.country

CREATE UNIQUE INDEX rl_pais_artigo_pk
   ON public.rl_pais_artigo USING btree (idartigo, idCountry);

CREATE INDEX rl_rl_pais_artigo_idartigo
   ON public.rl_pais_artigo(idartigo);
```

```
CREATE INDEX rl_pais_artigo_idCountry
   ON public.rl_pais_artigo(idCountry);

alter table rl_pais_artigo add column numAuthors integer;

update rl_pais_artigo
set numAuthors =
(select count(distinct raa.idAuthor)
from rl_autor_artigo raa
join author a on raa.idAuthor = a.id
WHERE raa.idArtigo = rl_pais_artigo.idArtigo and a.idcountry =
rl_pais_artigo.idcountry);

CREATE INDEX rl_pais_artigo_numAuthors
   ON public.rl_pais_artigo(numAuthors);

--5h17m
create table rl_author_institution_country as
select distinct af.listauthors, au.id as idAuthor, p.id as idArtigo,
i.institution, i.country
from affiliation af
join institution i on i.institution = af.institution
join pesquisaa p on p.afiliacoes like '%'||af.affiliation||'%'
join author au
    on af.listauthors = au.original_name --unico na lista
        or af.listauthors like au.original_name||';%'--primeiro da lista
        or af.listauthors like '%; '||au.original_name||';%'--meio da lista
        or af.listauthors like '%; '||au.original_name--último da lista

CREATE INDEX rl_author_institution_country_idartigo
   ON public.rl_author_institution_country(idartigo);

CREATE INDEX rl_author_institution_country_idauthor
   ON public.rl_author_institution_country(idauthor);

CREATE INDEX rl_author_institution_country_idCountry
   ON public.rl_author_institution_country(idCountry);

alter table rl_author_institution_country add column idCountry integer
update rl_author_institution_country set idCountry = c.id
from country c where c.country = rl_author_institution_country.country

update rl_pais_artigo
set numAuthors =
(select count(distinct raic.idAuthor)
from rl_author_institution_country raic
WHERE raic.idArtigo = rl_pais_artigo.idArtigo and raic.idcountry =
rl_pais_artigo.idcountry)

select
   country,
   max(ranking) as h_index,
    (select avg(pa.numAuthors)
    from pesquisaa pa
    where pa.id in (
        select p.id
        from pesquisaa p
        join rl_pais_artigo r on r.idartigo = p.id
        join country c on c.id = r.idcountry
        where c.country = t.country
```

```
        and r.numAuthors >= (select avg(pa.numAuthors)
            from pesquisaa pa
            where pa.id in (
                select p.id
                from pesquisaa p
                join rl_pais_artigo r on r.idartigo = p.id
                join country c on c.id = r.idcountry
                where c.country = t.country
                )
            )
        )
    ) as avg_autores
from
    (
     select c.id as idcountry, c.country, p.id as idartigo, p.citacoes,
     rank() over (partition by c.country order by max(p.citacoes) desc) as
ranking
     from pesquisaa p
     join rl_author_institution_country r on r.idArtigo = p.id
     join country c on c.id = r.idcountry
     group by c.id, c.country, p.id
     having count(distinct r.idAuthor) >= (select avg(pa.numAuthors)
            from pesquisaa pa
            where pa.id in (
                select p.id
                from pesquisaa p
                join rl_pais_artigo r on r.idartigo = p.id
                join country c1 on c.id = r.idcountry
                where c1.country = c.country
                )
            )
    ) as t
where ranking <= citacoes
group by country
order by 2 desc
limit 15

select
    country,
    max(ranking) as indice_h,
    (select avg(pa.numAuthors)
       from pesquisaa pa
       where pa.id in (
          select p.id
             from pesquisaa p
             join rl_pais_artigo r on r.idartigo = p.id
             join country c on c.id = r.idcountry
             where pa.numAuthors >= 2 and c.country = t.country)
    ) as media_autores_pais
from
    (
    select c.id as idcountry, c.country, p.id as idartigo, p.citacoes,
     count(distinct r.idAuthor) as numAutoresPais,
     rank() over (partition by c.country order by max(p.citacoes) desc) as
ranking
     from pesquisaa p
     join rl_author_institution_country r on r.idArtigo = p.id
     join country c on c.id = r.idcountry
     group by c.id, c.country, p.id
     having count(distinct r.idAuthor) >= 2
    ) as t
```

```
where ranking <= citacoes
group by country
order by 2 desc
limit 15

-- Avaliação de universidades:

--Removendo falsos positivos de University
delete from university where university in ('Sch Commun', '64', 'Sch Publ
Hlth', 'Dept Biomed Informat', 'Sch Econ & Management', 'Res Inst',
                                'Dept Hlth', 'Dept Management Informat
Syst', 'Dept Econ', 'Hlth Informat', 'Inst Management', 'Technol Univ',
'168',
                                'Sch Nursing', 'Med Sch', 'Dept Pediat',
'Arizona State', 'Sch Publ Hlth', 'Dept Biomed Informat',
                                'Sch Lib & Informat Sci', 'Sch Informat
Sci', 'Penn State', 'Ciencia Informacao',
                                'City Univ Hong Kong', 'Normal Univ',
'Chinese Acad Sci', 'Childrens Hosp', '75 Francis St');

alter table author add column affiliation varchar(2000)
update affiliation set affiliation = replace(affiliation, '[', '')
update affiliation set institution = ltrim(institution)

alter table rl_author_institution_country add column idUniversity integer

update rl_author_institution_country
set idUniversity = u.id
from university u
join institution i on u.university = i.university
where i.institution = rl_author_institution_country.institution and
i.country = rl_author_institution_country.country;

--ATENCAO ESSA QUERY LEVA 1h51m PRA EXECUTAR
--relacionamento entre universidade e artigo considerando diferentes
grafias de universidade, 145895 registros
drop table rl_university_artigo cascade
create table rl_university_artigo as
select distinct
   p.id as idArtigo,
   u.id as idUniversity
from pesquisaa p
join institution i on p.afiliacoes like '%'||i.institution||'%'
join university u on i.institution like
'%'||u.university||'%'||u.country||'%'

CREATE UNIQUE INDEX rl_university_artigo_pk
   ON public.rl_university_artigo USING btree (idartigo int4_ops,
iduniversity int4_ops);

CREATE INDEX rl_university_artigo_idartigo
   ON public.rl_university_artigo(idartigo);

CREATE INDEX rl_university_artigo_iduniversity
   ON public.rl_university_artigo(iduniversity);

alter table rl_university_artigo add column numAuthors integer;
update rl_university_artigo
set numAuthors =
(select count(distinct raic.idAuthor)
from rl_author_institution_country raic
```

```
WHERE raic.idArtigo = rl_university_artigo.idArtigo
     and raic.iduniversity = rl_university_artigo.iduniversity)

--relacionamento entre universidade e artigo considerando somente a
universidade com o menor nome
create table rl_university_artigo_raic as
select count(distinct raic.idAuthor), raic.idUniversity, raic.idArtigo
from rl_author_institution_country
group by  raic.idUniversity, raic.idArtigo;

CREATE UNIQUE INDEX rl_university_artigo_raic_pk
   ON public.rl_university_artigo_raic USING btree (idartigo int4_ops,
iduniversity int4_ops);

CREATE INDEX rl_university_artigo_raic_idartigo
   ON public.rl_university_artigo_raic(idartigo);

CREATE INDEX rl_university_artigo_raic_iduniversity
   ON public.rl_university_artigo_raic(iduniversity);

select
   university||'('||country||')' as universidade, t.iduniversity,
   max(ranking) as h_index,
    (select avg(pa.numAuthors)
    from pesquisaa pa
    where pa.id in (
        select p.id
        from pesquisaa p
        join rl_university_artigo r on r.idartigo = p.id
        where r.iduniversity = t.iduniversity
        and r.numAuthors >= 2
        )
   ) as avg_autores
from
   (
    select u.university, r.idUniversity, u.country, p.id as idartigo,
p.citacoes,
    rank() over (partition by u.university, u.country order by
max(p.citacoes) desc) as ranking
    from pesquisaa p
    join rl_author_institution_country r on r.idArtigo = p.id
    join university u on r.idUniversity = u.id and u.country = r.country
    group by u.university, r.idUniversity, u.country, p.id
    having count(distinct r.idAuthor) >= 2
   ) as t
where ranking <= citacoes
group by university||'('||country||')', t.iduniversity
having max(ranking) > 24
order by 3 desc;

drop MATERIALIZED VIEW vi_university
CREATE MATERIALIZED VIEW vi_university
AS
 SELECT i.country, i.university, i.id,
    count(DISTINCT p.id) AS total_papers,
    count(DISTINCT aa.idauthor) AS total_authors,
    ( SELECT avg(pa.numauthors) AS avg
           FROM pesquisaa pa
          WHERE (pa.id IN ( SELECT p_1.id
                   FROM ((pesquisaa p_1
                     JOIN rl_university_artigo_raic r_1 ON ((r_1.idartigo =
```

```
p_1.id)))
                    JOIN university c ON ((c.id = r_1.iduniversity)))
                 WHERE c.id = i.id))) AS average_numauthors_paper,
    ( SELECT avg(r_1.numauthors)
         FROM ((pesquisaa p_1
          JOIN rl_university_artigo_raic r_1 ON ((r_1.idartigo = p_1.id)))
          JOIN university c ON ((c.id = r_1.iduniversity)))
        WHERE c.id = i.id) AS average_numauthors_university,

   ( SELECT PERCENTILE_DISC(0.5) WITHIN GROUP (ORDER BY pa.numauthors) AS
median
           FROM pesquisaa pa
          WHERE (pa.id IN ( SELECT p_1.id
                   FROM ((pesquisaa p_1
                     JOIN rl_university_artigo_raic r_1 ON ((r_1.idartigo =
p_1.id)))
                     JOIN university c ON ((c.id = r_1.iduniversity)))
                  WHERE c.id = i.id))) AS median_numauthors_paper,
    ( SELECT PERCENTILE_DISC(0.5) WITHIN GROUP (ORDER BY r_1.numauthors) AS
median
         FROM ((pesquisaa p_1
          JOIN rl_university_artigo_raic r_1 ON ((r_1.idartigo = p_1.id)))
          JOIN university c ON ((c.id = r_1.iduniversity)))
        WHERE r_1.numauthors > 1 AND c.id = i.id) AS
median_numauthors_university,
    ( SELECT mode() WITHIN GROUP (ORDER BY pa.numauthors) AS
mode_numauthors_paper
           FROM pesquisaa pa
          WHERE (pa.id IN ( SELECT p_1.id
                   FROM ((pesquisaa p_1
                     JOIN rl_university_artigo_raic r_1 ON ((r_1.idartigo =
p_1.id)))
                     JOIN university c ON ((c.id = r_1.iduniversity)))
                  WHERE c.id = i.id))) AS mode_numauthors_paper,
    ( SELECT mode() WITHIN GROUP (ORDER BY r_1.numauthors) AS modal_value
         FROM ((pesquisaa p_1
          JOIN rl_university_artigo_raic r_1 ON ((r_1.idartigo = p_1.id)))
          JOIN university c ON ((c.id = r_1.iduniversity)))
        WHERE r_1.numauthors > 1 AND c.id = i.id) AS
mode_numauthors_university,
    ( SELECT sum(pa.citacoes) AS sum
           FROM pesquisaa pa
          WHERE pa.numauthors > 1 and (pa.id IN ( SELECT p_1.id
                   FROM ((pesquisaa p_1
                     JOIN rl_university_artigo r_1 ON ((r_1.idartigo =
p_1.id)))
                     JOIN university c ON ((c.id = r_1.iduniversity)))
                  WHERE c.id = i.id))) AS sum_citacoes,
    ( SELECT avg(pa.citacoes) AS avg
           FROM pesquisaa pa
          WHERE (pa.id IN ( SELECT p_1.id
                   FROM ((pesquisaa p_1
                     JOIN rl_university_artigo r_1 ON ((r_1.idartigo =
p_1.id)))
                     JOIN university c ON ((c.id = r_1.iduniversity)))
                  WHERE c.id = i.id))) AS avg_citacoes,
    ( SELECT PERCENTILE_DISC(0.5) WITHIN GROUP (ORDER BY pa.citacoes) AS
median
           FROM pesquisaa pa
          WHERE (pa.id IN ( SELECT p_1.id
                   FROM ((pesquisaa p_1
```

```
                    JOIN rl_university_artigo r_1 ON ((r_1.idartigo =
p_1.id)))
                    JOIN university c ON ((c.id = r_1.iduniversity)))
                 WHERE c.id = i.id))) AS median_citacoes
   FROM (((pesquisaa p
     JOIN rl_university_artigo r ON ((p.id = r.idartigo)))
     JOIN university i ON ((i.id = r.iduniversity)))
     JOIN rl_autor_artigo aa ON ((p.id = aa.idartigo)))
  GROUP BY i.country, i.university, i.id;


select university,
country,
total_papers,
total_authors,
sum_citacoes,
round(average_numauthors_paper, 2) as average_numauthors_paper,
round(median_numauthors_paper, 2) as median_numauthors_paper,
--round(mode_numauthors_paper, 2) as mode_numauthors_paper,
round(average_numauthors_university, 2) as average_numauthors_university,
round(median_numauthors_university, 2) as median_numauthors_university,
--round(mode_numauthors_university, 2) as mode_numauthors_university,
round(avg_citacoes,2) as avg_citacoes--,
--round(median_citacoes,2) as median_citacoes
from vi_university
order by total_papers desc
limit 31;

select
   university||' ('||country||')' as universidade, t.iduniversity,
   max(ranking) as h_index,
    (select avg(pa.numAuthors)
    from pesquisaa pa
    where pa.id in (
        select p.id
        from pesquisaa p
        join rl_university_artigo r on r.idartigo = p.id
        where r.iduniversity = t.iduniversity
        and r.numAuthors >= 3
        )
   ) as avg_autores
from
   (
    select u.university, r.idUniversity, u.country, p.id as idartigo,
p.citacoes,
    rank() over (partition by u.university, u.country order by
max(p.citacoes) desc) as ranking
    from pesquisaa p
    join rl_author_institution_country r on r.idArtigo = p.id
    join university u on r.idUniversity = u.id and u.country = r.country
    group by u.university, r.idUniversity, u.country, p.id
    having count(distinct r.idAuthor) >= 3
   ) as t
where ranking <= citacoes
group by university||' ('||country||')', t.iduniversity
--having max(ranking) > 24 --top 15 excluia os empates
order by 3 desc;

select
   university||' ('||country||')' as universidade, t.iduniversity,
   max(ranking) as h_index,
```

```
    (select avg(pa.numAuthors)
    from pesquisaa pa
    where pa.id in (
        select p.id
        from pesquisaa p
        join rl_university_artigo r on r.idartigo = p.id
        join university u on r.idUniversity = u.id
        join vi_university vi on vi.university = u.university and
vi.country = u.country
        where r.iduniversity = t.iduniversity
        and r.numAuthors >= vi.average_numauthors_university
        )
   ) as avg_autores
from
   (
    select u.university, r.idUniversity, u.country, p.id as idartigo,
p.citacoes, vi.average_numauthors_university,
    rank() over (partition by u.university, u.country order by
max(p.citacoes) desc) as ranking
    from pesquisaa p
    join rl_author_institution_country r on r.idArtigo = p.id
    join university u on r.idUniversity = u.id and u.country = r.country
    join vi_university vi on vi.university = u.university and vi.country =
u.country
    group by u.university, r.idUniversity, u.country, p.id,
vi.average_numauthors_university
    having count(distinct r.idAuthor) >= vi.average_numauthors_university
   ) as t
where ranking <= citacoes
group by university||' ('||country||')', t.iduniversity
having max(ranking) >= 24
order by 3 desc;
```

APÊNDICE C – Código em Python para Análise dos Dados

```
from google.colab import drive
drive.mount('/content/gdrive')

import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

artigos =  pd.read_csv(r'gdrive/My
Drive/DOUTORADO/ci10anosComCategorias.csv')

citacoes = artigos[['citacoes']]
citacoes.rename(columns={'citacoes':'TotalCitacoes'}, inplace=True)
citacoes.boxplot(figsize=(4,5))

numAuthors = artigos[['numauthors']]
numAuthors.boxplot(figsize=(4,5))

Q1 = artigos['numauthors'].quantile(0.25)
Q3 = artigos['numauthors'].quantile(0.75)
IQR = Q3 - Q1    #IQR is interquartile range.

filter = (artigos['numauthors'] >= Q1 - 1.5 * IQR) & (artigos['numauthors']
<= Q3 + 1.5 *IQR)
arigosSemOutliers = artigos.loc[filter]
numauthorsSemOutliers = arigosSemOutliers[['numauthors']]
numauthorsSemOutliers.boxplot(figsize=(4,5))

Q1 = artigos['citacoes'].quantile(0.25)
Q3 = artigos['citacoes'].quantile(0.75)
IQR = Q3 - Q1    #IQR is interquartile range.

filter = (artigos['citacoes'] >= Q1 - 1.5 * IQR) & (artigos['citacoes'] <=
Q3 + 1.5 *IQR)
arigosSemOutliersCitAut = arigosSemOutliers.loc[filter]
citacoesSemOutliers = arigosSemOutliersCitAut[['citacoes']]
citacoesSemOutliers.boxplot(figsize=(4,5))

emGrupo = artigos.loc[artigos['isgroup'].isin(['Group'])]
emGrupoSemOut =
arigosSemOutliersCitAut.loc[artigos['isgroup'].isin(['Group'])]
mediaCitacoesEmGrupo = (emGrupo["citacoes"].mean())
print('Media Citacoes Em Grupo = ', mediaCitacoesEmGrupo)

individual = artigos.loc[artigos['isgroup'].isin(['Individual'])]
individualSemOut =
arigosSemOutliersCitAut.loc[artigos['isgroup'].isin(['Individual'])]
mediaCitacoesIndividual = (individual["citacoes"].mean())
print('Media Citacoes Individual = ', mediaCitacoesIndividual)

artigos.groupby('isgroup')['citacoes'].mean()

artigos.groupby('isgroup')['citacoes'].std()

pip install pandas-profiling

means_individual = (mediaCitacoesIndividual)
means_grupo = (mediaCitacoesEmGrupo)
```

```
fig, ax = plt.subplots()
index = np.arange(1)
bar_width = 0.35
opacity = 0.8

rects2 = plt.bar(index , means_individual, bar_width,
                 alpha=opacity,
                 color="blue",
                 label='Trabalhos Individuais')

rects1 = plt.bar(index + bar_width, means_grupo, bar_width,
                 alpha=opacity,
                 color="red",
                 label='Trabalhos em Grupo')

plt.ylabel('Média das Citações')
plt.xticks(np.arange(1), ['teste'])
plt.xticks(np.arange(1) + bar_width/2, [''])
#plt.xticks(index + bar_width/2, ("CCA","CCB","CCE","CCJ","CCS",
#                                 "CDS","CED","CFH","CFM","CSE","CTC"))
plt.legend()

plt.ylim(0,17)
plt.tight_layout()
plt.show()

citacoesEmGrupo = emGrupoSemOut[['citacoes']]
citacoesEmGrupo.rename(columns={'citacoes':'citacoesEmGrupo'},
inplace=True)
citacoesEmGrupo.boxplot(figsize=(4,5))

citacoesIndividual = individualSemOut[['citacoes']]
citacoesIndividual.rename(columns={'citacoes':'citacoesindividual'},
inplace=True)
citacoesIndividual.boxplot(figsize=(4,5))

import scipy.stats as stats
stats.probplot(arigosSemOutliersCitAut[arigosSemOutliersCitAut['isgroup']
== 'Individual']['citacoes'], dist="norm", plot=plt)
plt.title("Probability Plot - Citacoes Individual")
plt.show()

stats.probplot(arigosSemOutliersCitAut[arigosSemOutliersCitAut['isgroup']
== 'Group']['citacoes'], dist="norm", plot=plt)
plt.title("Probability Plot - Citacoes Group")
plt.show()

arigosSemOutliersCitAut['citacoesLog2'] =
np.log2(arigosSemOutliersCitAut['citacoes'])
artigos['citacoesLog2'] = np.log2(artigos['citacoes'])
artigos['numauthorsLog2'] = np.log2(artigos['numauthors'])

import scipy.stats as stats
stats.probplot(arigosSemOutliersCitAut[arigosSemOutliersCitAut['isgroup']
== 'Group']['citacoesLog2'], dist="norm", plot=plt)
plt.title("Probability Plot - Citacoes Log2 Group")
plt.show()

stats.probplot(arigosSemOutliersCitAut[arigosSemOutliersCitAut['isgroup']
== 'Individual']['citacoesLog2'], dist="norm", plot=plt)
plt.title("Probability Plot - Citacoes Individual")
```

```
plt.show()

individualLog2 =
arigosSemOutliersCitAut.loc[arigosSemOutliersCitAut['isgroup'].isin(['Indiv
idual'])]
citacoesIndividualLog2 = individualLog2[['citacoesLog2']]

grupoLog2 =
arigosSemOutliersCitAut.loc[arigosSemOutliersCitAut['isgroup'].isin(['Group
'])]
citacoesGrupoLog2 = grupoLog2[['citacoesLog2']]

subdataset = artigos[['citacoes','numauthorsLog2','numauthors', 'revista',
'citacoesLog2']]
subdataset.head()
#subdataset.corr()

import seaborn as sns

pp = sns.pairplot(subdataset,
                  size=1.8,aspect=1.2,
                  plot_kws=dict(edgecolor='k', linewidth=0.5),
                  diag_kws=dict(shade=True),
                  diag_kind='kde')

fig = pp.fig

import seaborn as sb
import pandas as pd

correlation = subdataset.corr()
plot = sb.heatmap(correlation)
plot

X = subdataset.iloc[:,:-1].values
y = subdataset.iloc[:,4].values

from sklearn.preprocessing import LabelEncoder
labelencoder = LabelEncoder()
X[:,3] = labelencoder.fit_transform(X[:,3])

X[:,3]

from sklearn.preprocessing import OneHotEncoder
onehotencoder = OneHotEncoder()
X = onehotencoder.fit_transform(X).toarray()

X = X[:,1:]

citacoesIndividualSorted = np.sort(citacoesIndividual)

citacoesEmGrupoSorted = np.sort(citacoesEmGrupo)

from scipy.stats import mannwhitneyu

# compare samples
stat, p = mannwhitneyu(citacoesEmGrupoSorted, citacoesIndividualSorted)
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
alpha = 0.05
if p > alpha:
```

```
    print('Same distribution (fail to reject H0)')
else:
    print('Different distribution (reject H0)')

# compare samples
stat, p = wilcoxon(citacoesEmGrupoSorted, citacoesIndividualSorted)
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
alpha = 0.05
if p > alpha:
    print('Same distribution (fail to reject H0)')
else:
    print('Different distribution (reject H0)')

stats.kruskal(citacoesEmGrupo, citacoesIndividual)
```