Universidade Federal do ABC
Centro de Matemática, Computação e Cognição
Programa de Pós-Graduação em Ciência da Computação

Predição na Ciência da Ciência: Explicativas de Modelos para Predições de Impacto Futuro de Cientistas Júnior

Antonio de Abreu Batista Junior

Santo André - SP, Dezembro de 2021

Antonio de Abreu Batista Junior

Predição na Ciência da Ciência: Explicativas de Modelos para Predições de Impacto Futuro de Cientistas Júnior

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação (área de concentração: Ciência da Computação), como parte dos requisitos necessários para a obtenção do Título de Doutor em Ciência da Computação.

Universidade Federal do ABC – UFABC

Centro de Matemática, Computação e Cognição

Programa de Pós-Graduação em Ciência da Computação

Orientador: Jesús Pascual Mena Chalco

Santo André - SP

Dezembro de 2021

Sistema de Bibliotecas da Universidade Federal do ABC
Elaborada pelo Sistema de Geração de Ficha Catalográfica da UFABC
com os dados fornecidos pelo(a) autor(a).

Batista Junior, Antonio de Abreu
Predição na Ciência da Ciência : Explicativas de Modelos para Predições de Impacto Futuro de Cientistas Júnior / Antonio de Abreu Batista Junior. — 2021.

99 fls.

Orientador: Jesús Pascual Mena Chalco

Tese (Doutorado) — Universidade Federal do ABC, Programa de Pós Graduação em Ciência da Computação, Santo André, 2021.

1. predições científicas. 2. explicativas de modelos. 3. aprendizado de máquina. 4. pesquisadores júnior. 5. modelo Q. I. Mena Chalco, Jesús Pascual. II. Programa de Pós-Graduação em Ciência da Computação, 2021. III. Título.

Este exemplar foi revisado e alterado em relação à versão original, de acordo com as observações levantadas pela banca examinadora no dia da defesa, sob responsabilidade única do(a) autor(a) e com a anuência do(a) (co)orientador(a).

gov.br Documento assinado digitalmente
Antonio de Abreu Batista Junior
Data: 28/01/2022 11:31:48-0300
Verifique em https://verificador.iti.br

de .

Nome completo e Assinatura do(a) autor(a)

gov.br Documento assinado digitalmente
Jesus Pascual Mena Chalco
Data: 28/01/2022 13:00:45-0300
Verifique em https://verificador.iti.br

(co)orientador(a)

MINISTÉRIO DA EDUCAÇÃO
Fundação Universidade Federal do ABC
Avenida dos Estados, 5001 – Bairro Santa Terezinha – Santo André – SP
CEP 09210-580 · Fone: (11) 4996-0017

FOLHA DE ASSINATURAS

Assinaturas dos membros da Banca Examinadora que avaliou e aprovou a Defesa de Tese de Doutorado do candidato, ANTONIO DE ABREU BATISTA JUNIOR realizada em 17 de Dezembro de 2021:

P/ DocuSigned by: Jesús Pascual Mena Chalco 0C1EA80D064E48B...

Prof.(a) ALEXANDRE DONIZETI ALVES
UNIVERSIDADE FEDERAL DO ABC

P/ DocuSigned by: Jesús Pascual Mena Chalco 0C1EA80D064E48B...

Prof.(a) FABIO MASCARENHAS E SILVA
UNIVERSIDADE FEDERAL DE PERNAMBUCO

P/ DocuSigned by: Jesús Pascual Mena Chalco 0C1EA80D064E48B...

Prof.(a) FERNANDO ANTONIO BASILE COLUGNATI
UNIVERSIDADE FEDERAL DE JUIZ DE FORA

P/ DocuSigned by: Jesús Pascual Mena Chalco 0C1EA80D064E48B...

Prof.(a) LUC QUONIAM
AIX-MARSEILLE UNIVERSITE

_____

Prof.(a) DAVID CORREA MARTINS JUNIOR
UNIVERSIDADE FEDERAL DO ABC

_____

Prof.(a) ESTEBAN FERNANDEZ TUESTA
UNIVERSIDADE DE SÃO PAULO

_____

Prof.(a) LEANDRO INNOCENTINI LOPES DE FARIA
UNIVERSIDADE FEDERAL DE SÃO CARLOS

DocuSigned by: Jesús Pascual Mena Chalco 0C1EA80D064E48B...

Prof.(a) JESUS PASCUAL MENA CHALCO
UNIVERSIDADE FEDERAL DO ABC - Presidente

* Por ausência do membro titular, foi substituído pelo membro suplente descrito acima: nome completo, instituição e assinatura

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado do Maranhão - Brasil (FAPEMA) - BD-08792/17 - Edital Nº 046/2017 e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Aos meus pais.
À Priscila e Luíza.

Agradecimentos

É com grande alegria que gostaria de registrar meus mais sinceros agradecimentos a todos que contribuíram para que este trabalho pudesse ser feito.

À Deus, meu senhor.

Ao meu orientador, e amigo, prof. Jesús P. Mena-Chalco, por suas orientações seguras e encorajamento por todos esses anos. Seu suporte e conhecimento científico têm sido um fator determiante para o sucesso desse trabalho. Eu sou muito grato por ter depositado confiança em minhas idéias e ter acreditado em mim.

Aos amigos e amigas do grupo de Cientometria da UFABC, Muhsen, Rozivaldo, Rafael, Victor, Wellington, Luciano Rossi, Diogo, Andréia, com os quais convivi e muito aprendi durante os anos de doutorado.

À FAPEMA pela bolsa de doutorado (BD-08792/17) e à CAPES. Sem o apoio financeiro destas instituições, a realização deste trabalho não teria sido possível.

À UFMA pela suporte e pela oportunidade de cursar o doutorado.

Ao Prof. Fábio Gouveia pela parceria, com ele tive a oportunidade de discutir e refinar boa parte das idéias apresentadas neste trabalho.

Por fim, agradeço a toda minha família pela compreensão e apoio que sempre recebi durante todo o processo. Em especial, a minha mãe, esposa e filha por me encorajar a lutar pelos meus sonhos.

Resumo

Batista-Jr, A. d. A. **Predição na Ciência da Ciência: Explicativas de Modelos para Predições de Impacto Futuro de Cientistas Júnior**. 2021. Tese (Doutorado) - Centro de Matemática, Computação e Cognição, Universidade Federal do ABC, Santo André, 2021.

Indicadores bibliométricos têm sido utilizados amplamente por governos, agências de governos e outros atores (e.g., universidades e pelos próprios cientistas) para mensurar o desempenho de pesquisadores, com o objetivo de orientar decisões de aprovação em estágio probatório, e como critério para progressão de carreira, financiamento de pesquisa, seleção de membros de quadros editoriais entre outras aplicações. A racionalidade por trás do uso desses indicadores como ferramentas de suporte à decisão, nesses contextos, é a força preditiva que supostamente eles carregam. Para essas e outras aplicações, o potencial para impacto futuro de um avaliado é a preocupação central. Indicadores alternativos (i.e., o índice-h futuro estimado via modelos de aprendizado de máquina) capturando o potencial para impacto futuro de cientistas claramente têm uma vantagem sobre indicadores tradicionais. Entretanto, vieses de toda natureza (e.g., favorecimento de certos grupos privilegiados) encontrados nesses modelos e a ausência de explicativas para suas tomadas de decisões têm tornado o uso deles inadequados nesses contextos em que decisões fundamentadas são pré-requisitos. Esta tese foca em como resolver essa questão. Em uma tentativa de aumentar a confiabilidade de modelos, nós propomos novos modelos interpretáveis para predição do Q futuro (impacto futuro) de jovens cientistas, e comparamos as suas acurácias e as suas explicativas para suas decisões contra outros modelos de aprendizado de máquina (Redes neurais profundas) e analíticos (e.g., o modelo Q, o índice h). Nós encontramos que esses modelos preditivos (indicadores alternativos) são confiáveis. Entretanto, por um lado esses testes revelaram que houve propagação involuntária de vieses (e.g., favorecimento de pesquisadores júnior já em posição de destaque) por algoritmos de aprendizado de máquina. Por outro lado, eles mostraram que as explicativas de modelos para suas decisões poderiam ser avaliadas por julgadores humanos para aliviar essa questão. Como esperado, nossos experimentos mostram que mesmo com poucos dados, o desempenho futuro (ou impacto) de pesquisadores júnior pode em grande medida ser predito. Em geral, as explicativas de modelos dadas para suas decisões são razoáveis. No entanto, a decisão final deve sempre ser do julgador humano porque existe sempre um risco embutido em uma predição. Adicionalmente, nós propomos o Q para periódicos, uma nova medida de impacto complementar às medidas de impacto de periódico. A sua principal vantagem sobre as demais medidas é ser uma medida não cumulativa, produzindo um ranking permanente.

Palavras-chave: predições científicas, explicativas de modelos para predições, aprendizado de máquina, cientometria, pesquisadores júnior, modelo Q.

Abstract

Batista-Jr, A. d. A. **Prediction in Science of Science: Explanations of Machines for Predictions of Future Impact of Young Scientists**. 2021. Tese (Doutorado) - Centro de Matemática, Computação e Cognição, Universidade Federal do ABC, Santo André, 2021.
Bibliometric indicators have been broadly used by governments, government agencies, and other actors to measure the performance of researchers, aiming to guide tenure decisions and as a criterion for career progression, research funding, selection of editorial board members, among other applications. The rationale behind the applicability of performance-based indicators as decision support tools in these contexts is the predictive capability that they supposedly carry. For this range of applications, the potential for the future impact of an appraisee is the principal concern. Alternative indicators (e.g., future index-h) seem to have a clear advantage over traditional indicators. However, diverse preferences found in these models for predicting future indicators and the need for explanations for their decisions have negatively impacted their use in real applications, mainly in contexts where reasoned assessments are a need. This thesis focuses on how to solve this barrier. In an attempt to increase the reliability of models, we propose novel interpretable models and compare their accuracy and explanations for their decisions against other machine learning and analytic models. We found that these models are reliable in estimating a researcher's future impact. Furthermore, these tests revealed that machine learning algorithms unintentionally discriminated against people. On the other hand, they also showed that explanations for model decisions alleviated this obstacle. As expected, our experiments show that the future performance of junior researchers can, to a large extent, be predicted, even with bounded data. In general, the reasons given by models for their decisions are reasonable. However, the final decision must always be of human beings because there is always a risk embedded into a prediction. Additionally, we propose the Q for journals, a novel measure complementary to the traditional journal impact measures. Its main advantage over others it is a non-cumulative measure producing an immutable ranking.
Keywords: scientific predictions, explanations for predictions, machine learning, scientometrics, researchers junior, Q model.

Lista de Figuras

1.1 Sistema predizendo o ranking futuro de pesquisadores individuais e o entendimento por trás de uma decisão do sistema. A equação estima o Q futuro do cientista júnior dado o seu Q (Q_i) e índice-h (h_i) correntes, e do seu coautor sênior (Q_{c_i}). 3
1.2 Comparativo de dois cientistas júnior. 4
1.3 Explicativas para predições individuais. O modelo prediz qual será o índice-h do cientista depois de algum tempo (para uma idade acadêmica futura), e destaca os indicadores de desempenho individual a partir do histórico passado do pesquisador que levou a esse número. A partir do quadro comparativo do dois cientistas, a comissão julgadora toma uma decisão. 4
1.4 Organização da tese. A divisão em partes da tese é ilustrada pelos 4 retângulos. . 7

2.1 Validação cruzada $10 - fold$. 11
2.2 Cada ponto representa um artigo da lista de publicações do cientista correspondente. O eixo y mede o impacto de cada artigo. No caso do cientista 1 seu melhor artigo foi no começo da carreira e do cientista 2 no final da carreira. 17
2.3 O índice-h depende da idade acadêmica do pesquisador. Além disso, o índice-h pode aumentar sem que haja a produção de novo conhecimento. Estas características torna o índice inadequado para avaliar cientistas em diferentes estágios da carreira. Adaptado de Penner et al. (2013). 20
2.4 Comparativo da evolução das métricas Q e H para um cientista, desde o ano 10 da sua carreira até o ano 30. Por toda a carreira do pesquisador o Q é praticamente o mesmo, enquanto que o seu índice-h cresce com a idade acadêmica da pessoa. . 21
2.5 Fluxo de trabalho de um sistema de aprendizado de máquina interpretável. 26

3.1 Crescimento da literatura de Ciência da Computação. 31
3.2 Distribuição de citação de 1.005.924 artigos com pelo menos uma citação na base de dados da ACM entre 1935 e 2017, em uma escala logarítmica dupla. 32
3.3 Distribuição do número de autores por publicação. 33
3.4 Comparativo das funções mais citadas na literatura para descrever N(X). N(X) é a quantidade de publicações com X citações. CDF é a função de distribuição acumulada. 34
3.5 A estatística *log-likelihood ratio* não tende ao infinito. 35

3.6 Tendência de coautoria múltlipa. N(Y) é o número de artigos com Y autores. . . 36
3.7 Número médio de citações por artigo por ano. Como indica a função exponencial $y(x)$ em que x é o ano, número tem crescido exponencialmente. 37
3.8 Crescimento do número de publicações no periódico PRL. 38
3.9 Crescimento da quantidade de citações recebidas por artigo por ano no periódico PRL. Para o cálculo da média considerou-se as citações recebidas por um artigo nos três primeiros anos de vida (depois da publicação). 39
3.10 Distribuição do número de artigos pelo número de autores. 40

4.1 Linha do tempo mostrando publicações sobrepostas do histórico de publicação do pesquisador júnior i e o do coautor principal c_i. Pode existir mais de um coautor do pesquisador júnior elegível, mas o principal é aquele com um maior índice-h no ano da predição. 45
4.2 Acurácias para os modelos testados. 48
4.3 Resultados das regressões de valores de Q de um conjunto de cientistas medidos em dois momentos, no ano 10 e 20 da carreira (eixo y). Claramente, o Q desses pesquisadores medido no ano 10 é inferior ao Q deles medido no ano 20. 49
4.4 Plot no mesmo gráfico da regressão de valores preditos e observados do modelo linear (cor azul, Observado = $0.009 + 1.016$Predito) e do modelo Q (cor vermelha) - O valor do Q corrente do cientista júnior como suposto ser seu Q definitivo. . . 49
4.5 O tamanho da carreira do pesquisador júnior tem efeito significante sobre o desempenho do modelo, e também sobre suas explicativas (o quanto cada fator contribui para uma predição) para uma predição. 50
4.6 O padrão visual percebido nos *plots* de caixas de observados também é visto naqueles de preditos, mostrando que nossos modelos são modelos robustos para lidar com o aumento observado da média de citação ao longo do tempo. 51

5.1 Gráfico de sobreposição de termos mais frequentes encontrados nos títulos e resumos de artigos citando o trabalho de Acuna, Allesina e Kording (2012) gerado pelo software *VOSviewer*. 54
5.2 Indicadores tradicionais versus futuros. 58
5.3 Distribuição de coeficientes Kendall' τ e p-valores para 100 comparações de rankings- o randômico e o observado. 61
5.4 Distribuição de totais de pesquisadores entre os top-30% em cada ranking - o randômico e o observado. 62
5.5 Precisão Top-30% por grupo. Na cor lilás escuro é a precisão Top-30% (Pre) e na cor lilás claro é o erro (No). O erro significa que o candidato apareceu na lista observada, mas não foi predito. 62

6.1 O procedimento para o cálculo do Q para um periódico é como para um cientista. Basta ver o periódico como um cientista e o tempo de vida do periódico como o tempo de carreira do cientista. 66

6.2 Classificação dos seis periódicos induzido pelo Q. Nós consideramos todos os artigos publicados em cada periódico até o final de 2018. Q(t) é o Q do periódico, e t o tempo para cada artigo acumular citações. A quantidade de citações recebida no período t mede o impacto do artigo. 70
6.3 Evolução do Q dos periódicos. 71
6.4 Correlação entre os rankings induzidos pelo SJR e pelo Q. Diferentemente do ranking induzido pelo SJR que sofreu mudanças nas posições entre 2004 e 2019, o ranking pelo Q se manteve sem alterações, e apresenta boa correlação com o ranking induzido pelo SJR recente (ranking 2018). 72
6.5 Boxplots comparativos de valores de Q de autores por periódico. Existe uma correlação positiva fraca entre o Q do periódico e o Q do autor. 73
6.6 Os correlogramas das séries de valores de Q de um periódico gerados pelos métodos de autocorrelação (r_u) e autocovariância (g_u) divergem, indicando que a séries são geradas por um processo não estacionário. Cada série contém os valores de Q do periódico após t=19, 22, 23, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41, 44 anos depois da sua criação. 76
6.7 Ao misturar os valores de Q das séries, os correlogramas gerados pelo método de autocorrelação (r_u) e autocovariância (g_u) praticamente caminham juntos, sugerindo que as séries são geradas por um processo estacionário. 77

A.1 Estrutura de dados representando a lista de publicações de cada autor identificado no conjunto de dados. a2p: nome do hash; Chave: o identificador do autor; Valor: a lista de identificadores das suas publicações. 91
A.2 Estrutura de dados mantendo a informação dos autores de cada publicação. p2a: nome do hash; Chave: o identificador da publicação ; Valor: a lista de identificadores dos autores da publicação. 92
A.3 Estrutura de dados *Hash* representando a lista de publicações citando uma publicação. p2p: nome do hash; Chave: o identificador da publicação ; Valor: a lista de identificadores das publicações citanto a publicação. 92
A.4 Estrutura de dados mantendo a informação do ano de publicação de cada artigo no conjunto de dados. year: nome do hash; Chave: o identificador da publicação; Valor: o ano da publicação. 92
A.5 Estrutura de dados mantendo a informação do veículo de publicação de cada artigo. p2v: nome do hash; Chave: o identificador da publicação; Valor: nome do veículo de publicação. 93

Lista de Tabelas

2.1 Resumo dos resultados do teste de hipóteses. 13
2.2 Conjunto de teste. 15

4.1 Conjunto de dados D criado. n é o número de cientistas. 43
4.2 Variáveis preditoras. 43
4.3 Resultados das regressões para os tamanhos de carreira 2,3,4, e 5. Configuração 1. 47
4.4 Resultados da regressão (do modelo RL) para o tamanho 5 e a configuração 2. . . 47
4.5 Resultado da regressão de valores Observados versus Preditos. 50

5.1 Indicadores usados para induzir os rankings futuros. 58
5.2 Coeficientes de correlação τ de Kendall entre a classificação observada e as que foram preditas pelos modelos de seleção. 61

6.2 Periódicos indexados pelo conjunto de dados APS. 68
6.3 Tabelas de valores críticos com constante. 69
6.4 Valor da estatística KPSS para as séries randomizada e observada. A série observada contem o Q do periódico medido para 19, 22, 23, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41, 44 anos após a sua criação. A série randomizada é composta desses mesmos valores, porém misturados 39, 25, 19, 35, 28, 29, 26, 37, 34, 30, 33, 27, 32, 41, 38, 31, 22, 36, 44, 23. 73

Lista de Algoritmos

1 Estima a diferença de desempenho entre dois métodos de aprendizado L_A e L_B usando Validação Cruzada. 12
2 Computa o Q do cientista. 17
3 Desambiguador de nomes de autores. 34
4 Identifica na base de dados de publicações aqueles pesquisadores com colaboradores elegíveis 44
5 Busca o provável colaborador principal do pesquisador júnior 45
6 Cálcula o Q do periódico. 67

Lista de Acrônimos

ACM	*Association for Computing Machinery*
APS	*American Physical Society*
CDF	*Cumulative Distribution Function* (Função de Distribuição Acumulada)
DBLP	*Digital Bibliography & Library Project*
Deep	Redes Neuaris Profundas
KPSS	*Kwiatkowski-Phillips-Schmidt-Shin*
MSE	*Mean Squared Error* (Erro Quadrático Médio)
PRL	*Physical Review Letters*
RL	Regressão Linear
SJR	*SCImago Journal Rank*
VC	Validação Cruzada
WoS	*Web of science* & *WoS*

Sumário

1 Introdução **1**

1.1 Objetivos . . . 2
1.2 Justificativas . . . 3
1.3 Contribuições . . . 5
1.4 Estrutura da tese . . . 5

2 Fundamentação **8**

2.1 Aprendizado de Máquina . . . 8
2.1.1 Aprendizado Supervisionado . . . 9
2.1.2 Avaliação de Modelos . . . 10
2.2 Cientometria . . . 15
2.2.1 Fator Q . . . 16
2.2.2 Definição de Pesquisador Júnior . . . 18
2.2.3 Definição de Sucesso Científico . . . 18
2.2.4 Classificação de Medidas . . . 19
2.3 Predição na Ciência da Ciência . . . 22
2.3.1 Definição do Problema Predição do Impacto Futuro do Pesquisador . . . 22
2.3.2 Barreias à Aceitação de Modelos Preditivos . . . 22
2.4 Considerações Finais . . . 27

3 Conjuntos de Dados **29**

3.1 Introdução . . . 29
3.2 ACM . . . 30
3.3 APS . . . 33
3.4 Considerações Finais . . . 36

4 Equações para Prever o Sucesso Futuro de Pesquisadores Júnior **41**

4.1 Introdução . . . 41
4.2 Materiais e Método . . . 42
4.2.1 Identificação do Colaborador Chave do Pesquisador Júnior . . . 44
4.3 Resultados . . . 46
4.3.1 Configuração dos Experimentos . . . 46

4.3.2 Acurácias para as abordagens *Deep* e de Regressão Linear (RL) 48
4.3.3 Efeito da curta presença na academia do pesquisador júnior no seu Q . . . 48
4.3.4 Resultados da regressão de valores Observado x Predito 48
4.4 Discussão e Conclusões 51

5 Arcabouço Computacional para Selecionar Pesquisadores Importantes ainda em suas Fases de Pesquisadores Júnior **53**
5.1 Introdução 53
5.2 Revisão de Literatura 55
5.3 Abordagem 57
5.3.1 Cálculo dos Indicadores Futuros 57
5.4 Avaliação 59
5.4.1 Configuração experimental 59
5.4.2 Resultados 61
5.4.3 Discussão 63
5.5 Conclusão 64

6 Q para Períodicos **65**
6.1 Introdução 65
6.2 Método 67
6.3 Resultados 69
6.4 Discussão e Conclusão 74

7 Conclusões **78**
7.1 Resumo de Contribuições 79
7.2 Lista de Publicações 80
7.3 Limitações e Trabalhos Futuros 81

Bibliografia **83**

A Estruturas de Dados usadas na Tese **91**

B Codigo R para Leitura de Dados **94**

C Codigo R para Cálculo de Métricas **97**

Capítulo 1

Introdução

Predições precisas do sucesso futuro do pesquisador têm recebido muita atenção (BAI et al., 2020; ACUNA; ALLESINA; KORDING, 2012). Isso se deve à existência de muitas aplicações do mundo real (e.g., seleção de novos membros de quadros editoriais, aprovação em estágio probatório, promoção docente) que podem se beneficiar delas.

Para um ampla faixa dessas aplicações, o potencial de impacto futuro de um candidato (cientista) é crítico, mais do que seu impacto passado (SCHWEITZER, 2014). E, embora nenhuma evidência clara ligue o desempenho individual passado do cientista ao seu sucesso futuro, comissões avaliadoras têm usado isso (o desempenho passado do cientista) como uma aproximação para o seu sucesso futuro.

No entanto, mais do que uma predição, uma decisão fundamentada é necessária. Quando cientistas estão sendo considerados para financiamento, promoção, direito à estabilidade na carreira, decisões justificadas são um pré-requisito. E, isso tem criado vários desafios únicos para aplicação de aprendizado de máquina nesses contextos, porque muito do desempenho espetacular de aprendizado de máquina em outras tarefas, como de reconhecimento de imagens, é atualmente alcançado ao custo de aumentar as complexidades das estruturas internas das máquinas, o que compromete o entendimento de suas decisões. Portanto, existe muito espaço para melhorias nesse ponto.

Nesta tese, uma abordagem guiada por dados usando aprendizado de máquina é adotada. Guidado por uma abordagem supervisionada, e por dados reais de carreiras científicas, algoritmos de aprendizado de máquina podem encontrar o modelo mais adequado para uma tarefa de predição, com a de predizer o provável sucesso futuro de novos cientistas entrando no sistema.

Particularmente, nesta tese, os esforços são colocados no desenvolvimento de um modelo entendível do fator Q de um pesquisador júnior em termos de suas características bibliométricas do começo da carreira. O fator Q é um elemento do modelo Q (SINATRA et al., 2016), capturando o talento científico de uma pessoa. Neste trabalho, um pesquisador é classificado como pesquisador júnior pelos cinco primeiros anos desde a sua primeira publicação, e devido a isso, os dados sobre pesquisadores júnior usados para predizer seus impactos futuros são limitados.

Um estudo anterior, por Penner et al. (2013), apontou falhas no modelo de Acuna (ACUNA;

ALLESINA; KORDING, 2012), um modelo do índice h futuro de um cientista, e concluiu que o modelo é inclinado para pesquisadores mais velhos. O modelo de Acuna e seus colegas depende da idade acadêmica do pesquisador (PENNER et al., 2013), quanto maior é a idade maior é o desempenho do modelo.

Ao focar em grupos por faixa etária, e por campo científico, Gogoglou (2017) encontrou modelos mais retos (menos enviesados). Também, trabalhando especificamente com pesquisadores júnior, Lee (2019) e Lindahl (2020) concluíram que seus impactos (ou desempenhos) de curto prazo (nos próximos anos) podem ser preditos. Lee (2019) demonstrou que o número de publicações em periódicos ou eventos científicos, durante a fase de pesquisador júnior, foi o fator que contribuiu mais para o desempenho individual do cientista nos anos seguintes.

Em linha com Lee (2019), Lindahl (2020) concluiu que cientistas júnior que publicam em maior quantidade e em mais periódicos prestigiados, são mais prováveis de desenvolverem pesquisas de ponta no anos seguintes, e que usar isso como critério para embasar decisões de financiamento de pesquisa ou orientar decisões de aprovação em estágio probatório aumentará as desigualdades de gênero já existentes.

No entanto, apesar do crescente interesse pela importância futura de jovens cientistas, os fatores relacionados aos primeiros anos de carreira contribuindo mais para seus sucessos futuros (impactos de longo prazo) não são bem entendidos.

1.1 Objetivos

O objetivo principal desta tese é propor um novo arcabouço computacional com adequada compreensibilidade, para selecionar pesquisadores importantes ainda em suas fases de pesquisadores júnior.

Os objetivos específicos do trabalho são os seguintes:

1. Propor modelos com excelente compreensibilidade para predizer o Q estável do cientista júnior, em termos de seu desempenho individual e do seu influenciador principal (coautor chave).

2. Demonstrar a superioridade desses modelos, estimando o potencial impacto futuro do cientista, e chamados aqui de indicadores alternativos, sobre indicadores tradicionais quanto à produzir um ranking futuro deles.

3. Desenvolver um arcabouço computacional para selecionar pesquisadores importantes ainda em suas fases de pesquisadores júnior, usando esses modelos.

4. Adicionalmente, propor o fator Q para periódico.

Observa-se ainda que não há na literatura trabalho similar a este.

1.2 Justificativas

Predições corretas do impacto futuro do pesquisador, acompanhadas de uma justificativa, podem ajudar avaliadores a justificar uma decisão em que um candidato é preferido a um outro, principalmente, quando humanamente é muito difícil justificar a diferença na avaliação. O entendimento do raciocínio das máquinas, por avaliadores humanos, resultará em justificativas claras da razão de uma dada classificação final, mesmo quando uma decisão dos avaliadores divergir da máquina.

O sistema na Figura 1.1 apresenta o porquê de uma decisão para uma comissão julgadora. Nele, a mesma tarefa de produzir um ranking futuro dos pesquisadores é atribuída tanto aos avaliadores quanto ao modelo preditivo, e ambos farão suas predições do ranking futuro. O grau de concordância entre eles poderá ser medido, e através da interação com a interface de explicação, os avaliadores, poderão ter um entendimento do motivo de um discrepância entre uma predição sua e uma predição da máquina, em relação a um caso particular. Uma explicativa poderia ser a da Figura 1.2. A racionalidade é dada pela equação compreensiva do arcabouço.

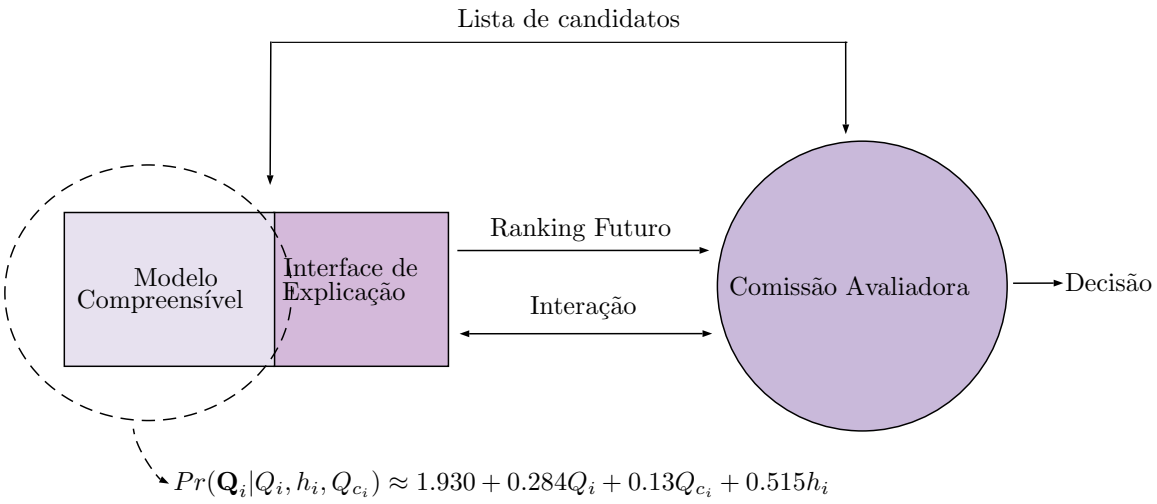
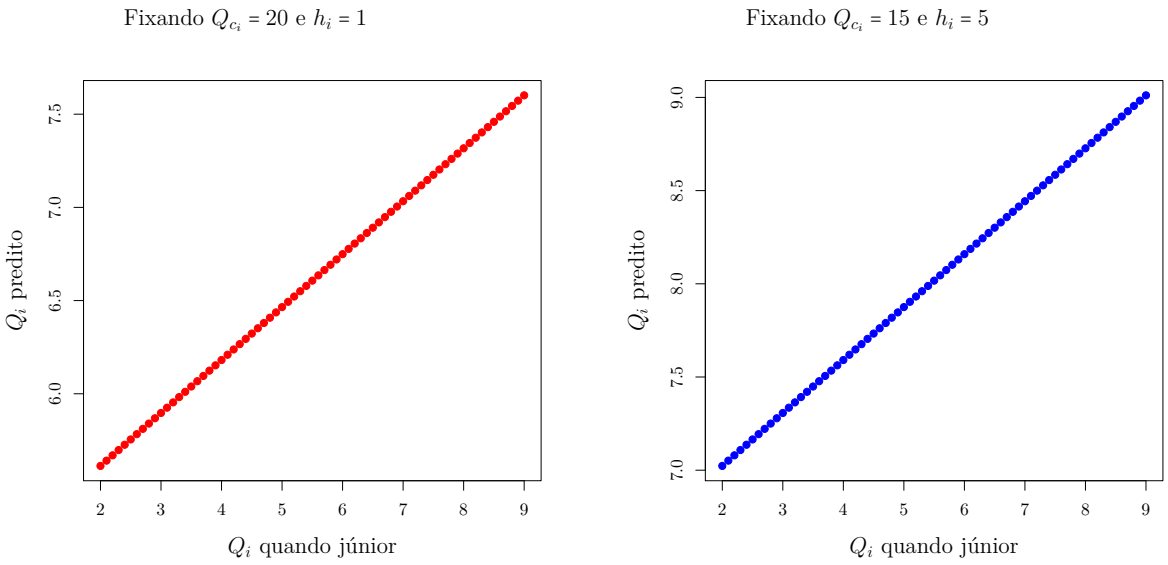


Figura 1.1: Sistema predizendo o ranking futuro de pesquisadores individuais e o entendimento por trás de uma decisão do sistema. A equação estima o Q futuro do cientista júnior dado o seu Q (Q_i) e índice-h (h_i) correntes, e do seu coautor sênior (Q_{c_i}).

O quadro comparativo na Figura 1.3 ilustra uma explicativa alternativa. O sistema prediria a evolução do índice-h futuro do cientista (ou outro indicador ou indicadores), e apresentaria um motivo para esse número, destacando a contribuição de cada indicador. Por serem, o índice-h e os outros indicadores do cientista, métricas conhecidas dos avaliadores, eles saberiam interpretar a razão do modelo ter decidido por esse número. Os avaliadores poderiam avaliar os cientistas a partir de uma visão mais ampla: não só o passado e o presente, mas também o futuro.



(a) Cientista 1.

(b) Cientista 2.

Figura 1.2: Comparativo de dois cientistas júnior.

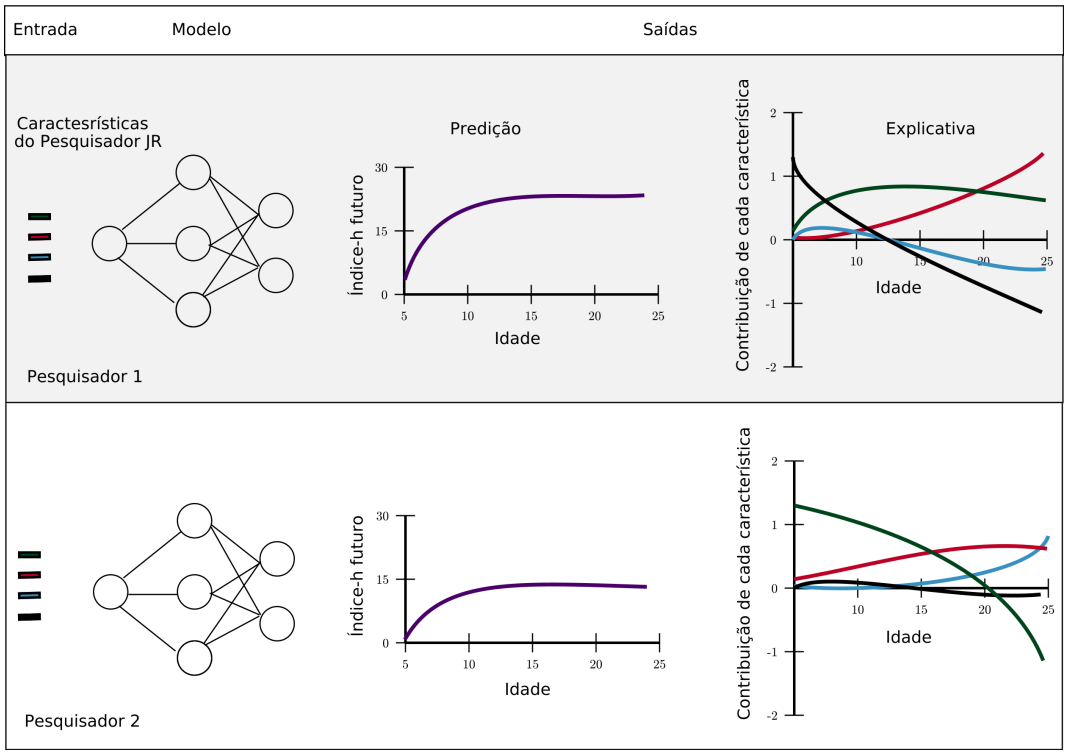


Figura 1.3: Explicativas para predições individuais. O modelo prediz qual será o índice-h do cientista depois de algum tempo (para uma idade acadêmica futura), e destaca os indicadores de desempenho individual a partir do histórico passado do pesquisador que levou a esse número. A partir do quadro comparativo do dois cientistas, a comissão julgadora toma uma decisão.

1.3 Contribuições

Este trabalho contribui para estender o conhecimento atual do uso de explicativas de modelos para o provável impacto futuro de cientistas, em contextos de promoção, montagem de quadros editoriais, aprovação de candidatos em estágio probatório, etc. Uma explicativa é uma fundamentação (um raciocínio) servindo para explicar um decisão.

Como principais contribuições destacam-se:

1. A proposição de novas equações aproximadas para o Q futuro de pesquisadores júnior. As equações usam características ligadas aos seus inícios de carreira, como o sucesso de seus colaboradores chaves e seus próprios desempenhos individuais.

2. O desenvolvimento de uma solução inovadora (um arcabouço computacional) para auxiliar comissões avaliadoras em suas tarefas. O arcabouço usa indicadores alternativos (e.g., o Q futuro, a quantidade de citações futuras de um artigo, o número de novas ocorrências futuras de publicações para um cientista), que tentam estimar o potencial para impacto futuro de um pesquisador, para avaliar pesquisadores individuais. Este trabalho argumenta, e também demonstra, que esses indicadores superam indicadores tradicionais (e.g, o índice-h) que são enviesados contra (prejudicam injustamente) pesquisadores júnior, algo que resulta em problemas para várias aplicações (e.g, selecionar membros de quadros editoriais ou comitê de programas).

3. A introdução do Q para periódicos, uma medida complementar as medidas de impacto de periódicos usadas atualmente. O diferencial dessa medida é que ela é uma medida não cumulativa, e portanto produz um ranking de periódicos permanente.

1.4 Estrutura da tese

A tese está organizada em sete capítulos.

O Capítulo 1 dá uma visão geral do tópico de pesquisa. A Seção 1.3 resume as principais contribuições dessa tese, enquanto que a seção 1.4 provê um visão geral da estrutura do texto.

O Capítulo 2 apresenta conceitos e definições fundamentais dessa tese. Descrições das metodologias de avaliação usadas nos capítulos seguintes e dos paradigmas de aprendizado de máquina são providas.

O Capítulo 3 dá uma visão geral dos dois conjuntos de dados usados nesse trabalho, o conjunto de dados da ACM é usado nos dois capítulos seguintes a esse, 4 e 5, para reconstruir as carreiras dos cientistas e computar seus impactos de curto e longo prazo, enquanto que o conjunto de dados da APS é usado no Capítulo 6 para avaliar as diferenças de impacto dos principais periódicos de revisão de física, entre eles *Physical Review Letters*, *Physical Review D*, e *Reviews of Modern Physics*.

Os três próximos capítulos são os capítulos mais importantes dessa tese, cada um diz respeito a uma publicação, em produção ou publicada recentemente.

No primeiro deles, o Capítulo 4, são propostas novas equações para estimar o Q de pesquisadores júnior usando dados de publicações de seus cinco anos depois do início de suas carreiras de pesquisa (e.i, de sua primeira publicação observada no conjunto de dados). O trabalho descrito nesse capítulo é baseado na:

- BATISTA-JR, A. d. A.; GOUVEIA, F. C.; MENA-CHALCO, J. P. Predicting the Q of-junior researchers using data from the first years of publication. *Journal of Informetrics*,v. 15, n. 2, p. 101130, 2021.

No Capítulo 5, métricas alternativas (e.g., Q futuro ao invés do Q corrente) de impacto de nível de autor são testadas e suas acurácias comparadas contra a de métricas tradicionais (e.g., índice-h) quanto à tarefa de produzir um ranking aproximado do ranking de fato observado no futuro. Demonstra-se que as métricas alternativas, para essa tarefa em particular, são mais adequadas do que as tradicionais. O capítulo é baseado na:

- BATISTA-JR, A. d. A.; GOUVEIA, F. C.; MENA-CHALCO, J. P. Identification of promising researchers through fast-and-frugal heuristics. In: MANOLOPOULOS Y., VERGOULIS T. (Org.). Predicting the Dynamics of Research Impact. Cham: Springer International Publishing, 2021. p. 195-207.

No terceiro, o Capítulo 6, o Q para periódico é introduzido. Este último é baseado em um artigo em produção.

Por fim, algumas conclusões tiradas são apresentadas no Capítulo 7, e também são discutidas direções para pesquisa futura nesse capítulo. O diagrama da Figura 1.4 apresenta graficamente a estrutura da tese.

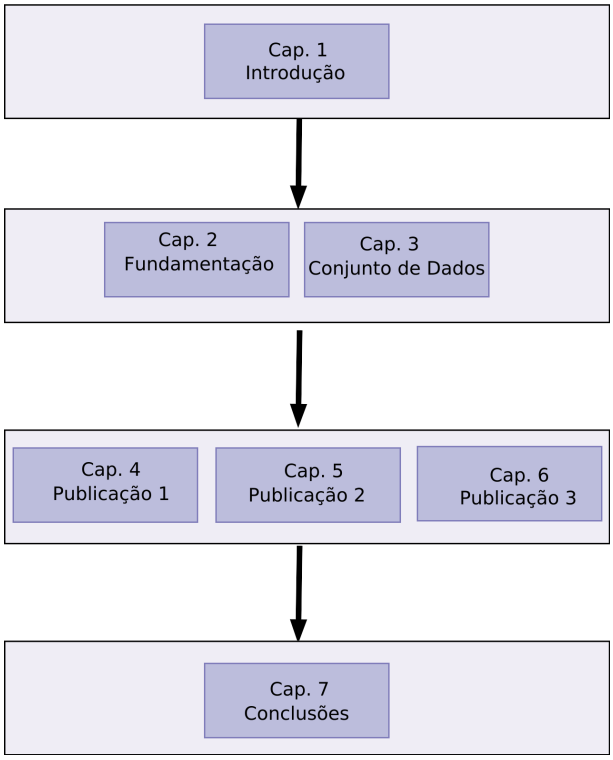


Figura 1.4: Organização da tese. A divisão em partes da tese é ilustrada pelos 4 retângulos.

Capítulo 2

Fundamentação

Para desenvolver um arcabouço computacional para identificar pesquisadores importantes de amanhã ainda em suas fases iniciais da carreira, é preciso antes inferir modelos lineares para prever o potencial impacto futuro do pesquisador júnior. Para isso uma abordagem guiada por dados usando aprendizado de máquina é adotada. Neste Capítulo, detalhes da abordagem e do paradigma de aprendizado são providos. Também, são providas uma introdução do campo de estudo em que esta pesquisa está inserida e alguns conceitos chaves para essa tese.

Especificamente, na seção 2.1 é dada uma breve introdução de aprendizado de máquina, dando especial atenção a abordagem supervisionada e a avaliação de modelos. Também são descritos os modelos lineares e de redes neurais. Na seção 2.2 é dada uma breve visão geral do campo da Cientometria, um área de estudo situada na intersecção entre as ciências sociais, a ciência da informação, e a ciência da computação. Finalmente, na seção 2.3, o problema de predição de impacto futuro do pesquisador júnior é descrito, e alguns entraves à aceitação de modelos preditivos pela comunidade científica são discutidos.

2.1 Aprendizado de Máquina

Aprendizado de Máquina, *Machine Learning* (ML) em inglês, é um subcampo da ciência da computação interessado na criação de máquinas (programas de computadores) que podem melhorar automaticamente através da experiência (MITCHELL, 1997). Para o autor, um programa de computador é dito aprender a partir da experiência E e medida de desempenho P, para alguma classe de tarefas T, se seu desempenho nas tarefas em T, como medida por P, melhora com a experiência E.

Em ML, experiência existe na forma de dados (ZHOU, 2021), e a tarefa principal da ML é desenvolver algoritmos que criem novos modelos, a partir destes dados, que podem fazer predições sobre novos exemplos. Portanto, aprendizado de máquina é a técnica que melhora o desempenho de máquinas aprendendo a partir da vivência via métodos computacionais. Nesta tese, modelos são funções inferidas por algoritmos de aprendizado de máquina com bons desempenhos na tarefa de aprendizado. Tais modelos são usados para mapear novas observações.

Atualmente, o sucesso empírico nesse campo move-se mais rápido do que o nosso entendi-

mento matemático sobre o assunto (ARORA, 2018). E, isso tem mudado o mundo, sobretudo na ciência, em que ML tem assumido um papel cada vez mais central em apoiar a pesquisa científica (MJOLSNESS; DECOSTE, 2001). As aplicações do ML vão da predição de novas ocorrências de publicações para um cientista até a sugestão do parceiro certo para ele quanto à um novo projeto de pesquisa. Essas mudanças terão profundo impacto nas próximas décadas e são uma oportunidade ímpar para cientistas de dados e cientométricos.

2.1.1 Aprendizado Supervisionado

É um paradigma de sucesso de aprendizado de máquina. O objetivo de aprendizado supervisionado é construir um modelo conciso da distribuição de uma variável definida como alvo, em termos de características preditoras (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2006). O modelo inferido é então usado para predizer novos exemplos em que os valores das características preditoras são conhecidos, mas o valor da variável alvo é desconhecido.

Problemas de aprendizagem supervisionados são classificados em problemas de "regressão"e "classificação". Se a variável alvo for categórica (e.g., [macho ou fêmea],[verdadeiro ou falso],[ligado, desligado]) podemos definir esta tarefa como uma tarefa de classificação. Por outro lado, se a variável alvo for uma variável contínua, por exemplo, preço, salário, idade, esta é uma tarefa de regressão.

Em aprendizado supervisionado, funções de perda são componentes chaves, o papel delas é mapear os desvios (os erros) entre os valores reais e preditos. Para obter um modelo de melhor desempenho, algoritmos de aprendizado tentam continuamente ir atualizando os parâmetros dessas funções de maneira a minimizar esses desvios.

Abordagens de Aprendizado Supervisionado são classificadas baseadas em sua metodologia de modelagem, em modelos lineares, árvores de decisão, modelos baseados em exemplos e redes neurais (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2006). Nesta tese, o foco é em modelos lineares e de redes neurais.

Modelos Lineares e Regressão linear

Modelos lineares tentam identificar e estabelecer a relação entre variáveis explicativas e de resposta, esta última também conhecida como variável dependente. Enquanto que Regressão Linear é um método de aprendizado supervisionado que visa aprender os parâmetros de um determinado modelo linear. Formalmente, dado um conjunto de dados $D = \{(\mathbf{x}_1, y_1); (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$, em que y_i é a saída para a entrada $\mathbf{x}_i = \{x_{i1}; x_{i2}, \cdots, x_{im}\}$, $y_i \in \mathbb{R}$. O método de regressão linear tenta aprender a função:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b, \text{ tal que } f(\mathbf{x}_i) \simeq y_i, \ i = 1, 2, \cdots, m$$

Para determinar θ^*, isto é, $\mathbf{w}^T$ e b, os parâmetros ótimos da função, o segredo é medir a diferença entre os valores preditos e observados. Existem muitas métricas para esse propósito, uma das mais comumente usada é a métrica MSE. Portanto, encontrar θ^* corresponde a minimizar MSE, isto é:

$$\theta^* = \arg\min_{\theta} \underbrace{\frac{1}{m}\sum_{1}^{m}(y_i - \mathrm{f}_\theta(\mathbf{x}_i))^2}_{MSE} \tag{2.1}$$

MSE corresponde à distância euclidiana e tem uma interpretação geométrica simples. Um método geral para minimizar MSE é o método dos Mínimos Quadrados.

Os pesos **w** aprendidos precisamente expressam a importância de cada variável de entrada. Isso reveste o modelo linear com excelente compreensibilidade.

Modelos de Redes Neurais profundas

Uma Rede Neural Profunda com d camadas escondidas consiste de d matrizes $A_1, A_2, \cdots, A_d$ e uma função específica $\sigma : \mathbb{R} \mapsto \mathbb{R}$ chamada não-linearidade. A não-linearidade mais utilizada nos dias atuais é a função *rectilinear linear*, $\mathrm{RELU}_b = \max\{0, x - b\}$. Nesta função b é chamado *bias* e é também um parâmetro da rede juntamente com as matrizes $A_1, A_2, \cdots, A_d$. Definindo $y^0 = x^0$, essa rede computa $y^1, y^2, \cdots, y^d$ em que $y^i + 1 = \sigma(A_i y^i)$. A função $\sigma(z)$ denota o vetor obtido aplicando σ a cada coordenada de z. Cada coordenada de um vetor computado y^i representa um nó da rede e cada entrada das matrizes $A_1, A_2, \cdots, A_d$ relaciona-se a uma aresta. A saída da rede é y^d. O tamanho da rede é o número de nós nela. O número de parâmetros é o número de arestas mais o número de nós.

Uma Rede Neural Profunda, portanto, é uma função que mapeia o vetor x^0 para o vetor saída $y^d = f_{A_1, A_2, \cdots, A_d, \vec{b}}(x^0)$.

Portanto, dado um conjunto de dados $D = \{(\mathbf{x}_1, y_1); (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$, em que y_i é a saída para a entrada $\mathbf{x}_i = \{x_{i1}; x_{i2}, \cdots, x_{im}\}$, $y_i \in \mathbb{R}$. O método de aprendizado supervisionado tenta aprender a função:

$$f_{A_1, A_2, \cdots, A_d, \vec{b}}(\mathbf{x}), \text{ tal que } f(\mathbf{x}_i) \simeq y_i, \ i = 1, 2, \cdots, m$$

Contrariamente aos modelos lineares, os modelos de redes neurais são conhecidos como modelos caixas pretas. Os parâmetros aprendidos, as matrizes e o vetor de *bias*, não dizem nada sobre a contribuição de cada variável de entrada para uma saída da rede.

2.1.2 Avaliação de Modelos

Nesta seção, discute-se o erro de generalização do modelo, isto é, o erro calculado sobre as novas amostras. Um modelo pode ser avaliado quanto a sua capacidade de generalização, seu custo computacional, ou sua compreensibilidade. A discussão seguinte concentra-se em sua capacidade de generalização.

Em geral, usa-se o erro do modelo no conjunto de teste como uma aproximação para o erro de generalização. Quando avaliando um modelo, o objetivo é obter o modelo com o menor erro de generalização. O erro calculado no conjunto de treino é chamado erro de treinamento ou erro empírico, e é esse erro que se busca minimizar na prática, porém o modelo é avaliado no conjunto de tese.

Comumente, assume-se que as amostras dos conjuntos de teste e de treino são independentes e identicamente amostradas. Também deve-se garantir que exemplos de testes não apareçam no conjunto de treino.

A seguir, discute-se as duas abordagens comumente usadas para dividir os dados do conjunto D em um conjunto de treino S e um conjunto de teste T.

Hold-Out

O método *hold-out* divide o conjunto de dado D em dois subconjuntos disjuntos: um como o conjunto de treino S e o outro como o conjunto de teste T, em que $D = S \cup T$ e $S \cap T = \varnothing$. Treina-se um modelo no conjunto de treino S e então calcula-se o erro de teste no conjunto de teste T como uma estimativa do erro de generalização.

É importante notar que a divisão deve manter a distribuição de dado original para evitar introduzir viés adicional.

Na prática, aplicar uma única vez a rotina leva a uma estimativa do erro de teste não confiável, algo comum é repetir várias vezes a rotina e calcular a média do erro de teste como a estimativa do erro de generalização. Uma outra alternativa é o método de validação cruzada.

Validação Cruzada

Validação Cruzada (VC) divide o conjunto de dado D em k subconjuntos disjuntos com tamanhos similares, isto é, $D = D_1 \cup D_2 \cup \cdots \cup D_k$, em que $D_i \cap D_j = \varnothing$, para $i \neq j$. Tipicamente, cada subconjunto D_i tenta manter a distribuição de dados original via amostragem estratificada. Em cada tentativa de VC, usa-se a união de $k-1$ subconjuntos como o conjunto de treino para treinar o modelo, e o subconjunto não incluso como o conjunto de teste para avaliar o modelo. Na validação cruzada $k-fold$, o procedimento é repetido k vezes, e cada subconjunto D_i é usado como o conjunto de teste precisamente uma vez. Valores comuns de k são 5, 10 e 20. A Figura 2.1 ilustra a ideia da validação cruzada $10-fold$.

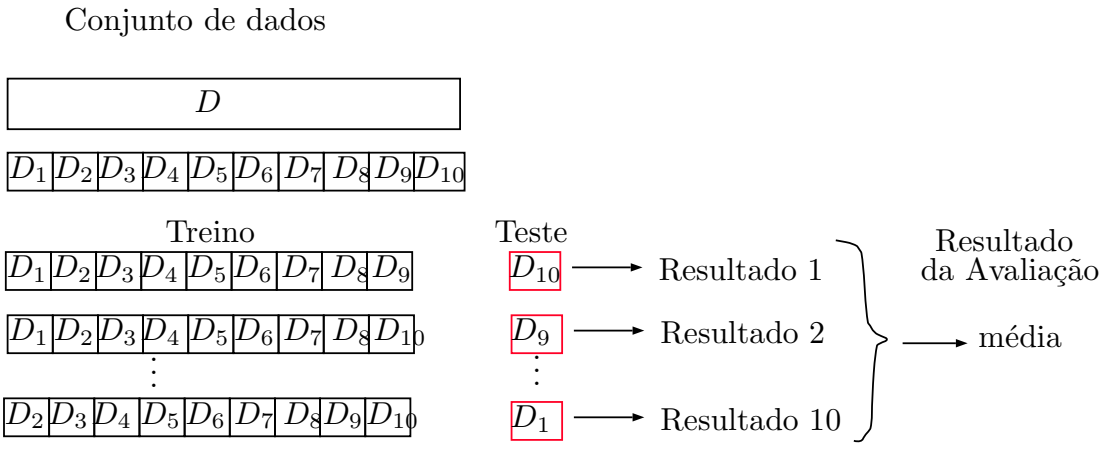


Figura 2.1: Validação cruzada $10-fold$.

Medidas de Desempenho

Em problemas de predição, é dado um conjunto de dados $D = \{(\mathbf{x}_1, y_1); (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$, em que y_i é a saída para a entrada $\mathbf{x}_i = \{x_{i1}; x_{i2}, \cdots, x_{im}\}$. Para avaliar o desempenho de um preditor f, compara-se sua predição $f(\mathbf{x}_i)$ contra y_i. Para problemas de regressão, a medida de desempenho mais comumente usada é o MSE (Mean Squared Error):

$$MSE = M(f, D) = \frac{1}{m}\sum_{1}^{m}(y_i - \mathrm{f}(\mathbf{x}_i))^2$$

Mas outras meditas também são muito usadas: o $RMSE = \sqrt{MSE}$ e o

$$R^2 = M(f, D) = 1 - \frac{\sum_1^m (y_i - \mathrm{f}(\mathbf{x}_i))^2}{\sum_1^m (y_i - \overline{y})^2}$$

Comparação de Algoritmos

A seguir, discute-se uma abordagem baseada em VC para comparar o desempenho de dois algoritmos de aprendizado de máquina L_A e L_B em uma base de dados D. O algoritmo 5 resume o passo a passo dessa abordagem. A abordagem consiste em comparar L_A e L_B em D, usando uma abordagem de validação cruzada *k-fold* e uma medida de desempenho M. Os resultados das avaliações individuais são comparados, e a diferença é retornada pelo algoritmo. E, uma interpretação simples pode então ser obtida, em que uma diferença pequena indica que L_A e L_B são equivalentes.

Algoritmo 1 Estima a diferença de desempenho entre dois métodos de aprendizado L_A e L_B usando Validação Cruzada.

Entrada: Um conjunto de dados D, um inteiro K e uma métrica M
Saída: $\bar{\delta}$

Particione o conjunto de dados disponível D em k subconjuntos disjuntos $\mathrm{T}_1, \mathrm{T}_2, \cdots, \mathrm{T}_\mathrm{k}$ de tamanho igual
para i **de** 1 **até** k **faça**
// Use T_i para o conjunto de teste, e o dado restante para o conjunto de treino S_i
    $\mathrm{S}_i \leftarrow \{\mathrm{D} - \mathrm{T}_i\}$
    $\mathrm{modelo}_A \leftarrow \mathrm{L}_A(\mathrm{S}_i)$
    $\mathrm{modelo}_B \leftarrow \mathrm{L}_B(\mathrm{S}_i)$
    $\delta_i \leftarrow \mathrm{M}(\mathrm{modelo}_A, \mathrm{T}_i) - \mathrm{M}(\mathrm{modelo}_B, \mathrm{T}_i)$
retorne o valor de $\bar{\delta}$, em que $\bar{\delta} = \frac{1}{k}\sum_{i=1}^{k}\delta_i$=média

Teste de Comparação de Algoritmos

Teste de hipótese é uma das técnicas para compara o desempenho de algoritmos. Um teste de hipótese estatística é um tipo de inferência estatística ou aprendizado como é chamado na computação, um processo de usar dado para inferir a distribuição verdadeira que gerou esse dado.

O propósito de um teste de hipótese é ajudar o pesquisador a alcançar uma conclusão relacionada a uma população, examinando uma amostra dela.

A seguir, define-se um teste de hipótese.

Formalmente um teste de hipótese consiste em particionar o espaço de parâmetros Θ em dois conjuntos disjuntos Θ_0 e Θ_1, e testar:

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

em que, H_0 é chamada a hipótese nula e H_1 a hipótese alternativa.

Suponha que X seja uma variável randômica e χ uma faixa de X. Nós testamos a nula encontrando um subconjunto de resultados $R \subset \chi$ que nós denominamos de região de rejeição. Se $X \subset R$, então nós rejeitamos a hipótese nula, caso contrário, nós não a rejeitamos:

$$X \in R \Rightarrow \text{ rejeite } H_0$$

$$X \notin R \Rightarrow \text{ Não rejeite } H_0$$

Usualmente, a região de rejeição R é da forma:

$$R = \{x : T(x) > c\}$$

em que T é um teste estatístico e c é um valor crítico.

Nós conjecturamos que $X \notin R$, e somente rejeitaremos a nula H_0, caso existam fortes evidências para isso. Nesse contexto, existem dois erros que nós podemos cometer. Rejeitar H_0 quando H_0 é verdadeiro, chamado erro tipo 1 ou não rejeitar H_0 quando H_1 é verdadeiro, chamado erro tipo 2. Os possíveis resultados de um teste de hipótese são resumidos na Tabela 2.1.

Tabela 2.1: Resumo dos resultados do teste de hipóteses.

	Não rejeitar a nula	Rejeitar a nula
A nula H_0 é verdadeira	✓	Erro tipo 1
A alternativa H_1 é verdadeira	Erro tipo 2	✓

A **função força** de um teste, com região de rejeição R, é definida por

$$\beta(\theta) = \mathbb{P}_\theta(X \in R).$$

O **tamanho** de um teste é definido ser:

$$\text{tam} = \sup_{\theta \in \Theta_0} \beta(\theta),$$

e.i., o tamanho do teste é a maior probabilidade de rejeitar a nula H_0 quando ela é verdadeira. Um teste é dito ter nível α se seu tamanho é menor do que ou igual a tam.

Uma hipótese da forma $\theta = \theta_0$ é chamada uma hipótese simples. Uma hipótese da forma $\theta > \theta_0$ ou $\theta < \theta_0$ é chamada uma hipótese composta. Um teste da forma

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

é chamado um teste bilateral (*two-sided test*). Um teste da forma

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0$$

ou

$$H_0 : \theta \geq \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0$$

é um teste unilateral (*one-sided test*).

A seguir, defini-se um **Teste de *Wald***:

Seja θ um parâmetro escalar, seja $\hat{\theta}$ uma estimativa de θ e seja $\hat{se}$ o erro padrão estimado de $\hat{\theta}$.

O teste de *Wald* consiste em testar:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

Assume-se que $\hat{\theta}$ é assintoticamente normal:

$$\frac{\hat{\theta} - \theta_0}{\hat{se}} \rightsquigarrow \mathcal{N}(0, 1.)$$

Nós devemos rejeitar H_0 quando $|\mathrm{W}| > z_{\frac{\alpha}{2}}$, em que:

$$|\mathrm{W}| = \frac{\hat{\theta} - \theta_0}{\hat{se}}$$

Finalmente, defini-se um **Teste de Wald** para comparar dois algoritmos de predição.

Teste ambos os algoritmos em um mesmo conjunto de teste de tamanho n. Faça $\mathrm{X}_i = 1$, se o algoritmo 1 está correto, no caso do teste i, e $\mathrm{X}_i = 0$ caso contrário. Faça o mesmo para o algoritmo 2, faça $\mathrm{Y}_i = 1$, se o algoritmo 2 está correto no caso do teste i, e $\mathrm{Y}_i = 0$ caso contrário. Se o problema for de regressão, então defina um valor limitante Δ tal que a predição de um algoritmo estará correta se a diferença (o erro) entre o valor de fato e o predito for menor do que Δ, e errada caso contrário. Defina $\mathrm{D}_i = \mathrm{X}_i - \mathrm{Y}_i$. Na Tabela 2.2 é mostrado com ficaria um conjunto de teste com as definições anteriores.

Seja $\delta = \mathbb{E}(\mathrm{D}_i) = \mathbb{E}(\mathrm{X}_i) - \mathbb{E}(\mathrm{Y}_i) = \mathbb{P}(\mathrm{X}_i = 1) - \mathbb{P}(\mathrm{Y}_i = 1)$.

Uma estimativa de δ é $\hat{\delta} = \overline{\mathrm{D}} = n^{-1} \sum_{i=1}^{n} \mathrm{D}_i$ e $\hat{\mathrm{se}}(\hat{\delta}) = \frac{\mathrm{S}}{\sqrt{n}}$, em que $\mathrm{S}^2 = n^{-1} \sum_{i=1}^{n} (\mathrm{D}_i - \overline{\mathrm{D}})$. Para testar

$$H_0 : \delta = 0 \quad \text{versus} \quad H_1 : \delta \neq 0,$$

nós usamos $\mathrm{W} = \frac{\hat{\delta}}{\hat{\mathrm{se}}}$ e rejeitamos a nula H_0 se $|\mathrm{W}| > z_{\frac{\alpha}{2}}$.

Tabela 2.2: Conjunto de teste.

Caso de teste i	X_i	Y_i	$D_i = X_i - Y_i$
1	1	0	1
2	1	1	0
3	1	1	0
4	0	1	-1
⋮	⋮	⋮	⋮
n	0	1	-1

2.2 Cientometria

A Cientometria é um campo do conhecimento que tenta estudar a ciência, a tecnologia, e a inovação a partir de uma perspectiva quantitativa (GOGOGLOU, 2017; LEYDESDORFF; MILOJEVIĆ, 2015). Os resultados de pesquisa nesse campo tem muitos usos, havendo interesse de Governos e instituições de pesquisas em utilizar este conhecimento com o objetivo de implementar diferentes formas de apoio ao desenvolvimento científico e tecnológico. Esta seção dá uma breve visão geral do campo.

A disponibilidade crescente de dado digital sobre a literatura científica tem provido cientistas com recursos de dados abundantes para analisar, quantificar e possivelmente predizer novas ocorrências de publicações/citações. Isso tem acelerado as pesquisas nessa área e tornado o campo atrativo (CLAUSET; LARREMORE; SINATRA, 2017).

Devido à necessidade por transparência e justificativas dos investimentos em pesquisa, governos e agências de governos precisam medir o desempenho de pesquisa de diferentes atores(e.g., pesquisadores, grupos de pesquisa, instituições de pesquisa, etc). Nesse contexto, a Cientometria tem provido o ferramental e a metodologia para eles (MINGERS; LEYDESDORFF, 2015; BORNMANN; GUNS et al., 2021a).

A Ciência da Ciência (FORTUNATO et al., 2018; WANG; BARABÁSI, 2021), como ela também é conhecida, é caracterizada pela heterogeneidade (BORNMANN; GUNS et al., 2021a), situando-se na intersecção entre as ciências sociais, a ciência da informação, e a ciência da computação com seus esforços para capturar padrões em "big data" (LEYDESDORFF; MILOJEVIĆ, 2015). O Termo “big data” refere-se ao dado que é complexo, rápido e muito grande, e que, por isso, é difícil de processar e armazenar usando as tecnologias computacionais tradicionais.

A pesquisa cientométrica engloba, mas não está limitada a, avaliação de pesquisa, estudo das dinâmicas de citação na ciência, estudo do desenvolvimento de novos campos da ciência, e desenvolvimento de indicadores para uso em contextos de política científica (MINGERS; LEYDESDORFF, 2015).

Além de aplicações relevantes desses métodos no desenvolvimento de políticas científicas, existem muitas outras aplicações importantes ainda pouco exploradas que podem se beneficiar de estudos cientométricos. A seguir são apresentadas duas delas, e que tem haver com o assunto dessa tese.

Uma aplicação é o provimento de predições fundamentadas na Ciência da Ciência, sobretudo em aplicações em que decisões justificadas são primordiais (e.g., aprovação/desaprovação em

estágios probatórios, promoção docente). Uma segunda aplicação é o uso delas para apoiar revisores em suas tomadas de decisão. Com o desenvolvimento de novos centros de pesquisa por todo o mundo e o acirramento da disputa por recursos de pesquisa, como Gogoglou (2017), nós também consideramos que a avaliação e o julgamento de novas propostas de pesquisa (projetos de pesquisas) não podem continuar dependendo exclusivamente de revisões subjetivas de pares, havendo uma necessidade por suporte de decisão através de ferramentas analíticas.

Por último, periódicos publicando resultados de pesquisas nessa área, e.g., *Journal of informetrics* e *Scientometrics*, têm encorajado a submissão de novos estudos do desenvolvimento da ciência cujas resultados apoiam-se em novos métodos matemáticos, estatísticos, computacionais.

A seguir, apresenta-se o fator Q, uma medida que busca definir e predizer carreiras científicas, quantificando a habilidade de um cientista de transformar uma ideia em uma descoberta com um dado impacto.

2.2.1 Fator Q

Na literatura existem muito poucos exemplos de estudos definindo uma carreira científica. O fator Q busca definir uma. Este fator é suposto capturar o talento científico da pessoa, e é único para cada indivíduo.

O fator Q, ou simplesmente o Q do cientista, é um elemento do modelo Q (SINATRA et al., 2016), um modelo que assume que o impacto dos artigos que nós publicamos depende de dois fatores: sorte e talento científico do pesquisador. Este modelo define o impacto $c_{i\alpha}$ de um artigo α de um cientista i como sendo um produto do potencial impacto da ideia p_α e o Q do cientista.

$$c_{i\alpha} = Q_i p_\alpha$$

No caso de cientistas habilidosos, o fator Q transforma sorte em uma carreira de sucesso.

O modelo Q da Sinatra e seus colegas parece razoável, as predições do modelo estão em excelente conformidade com dados reais de carreiras científicas e parece suficiente para explicar o que diferencia um cientista do outro.

O modelo é suportado pela regra de impacto randômico (SINATRA et al., 2016), uma regra que afirma que o artigo científico de maior impacto de um cientista pode estar, com igual probabilidade, em qualquer lugar na sua lista de artigos (i.e., pode está no começo, no meio ou no fim da lista de publicações) como pode ser visto na Figura 2.2.

Artigos muito citados são, portanto, um resultado de um cientista talentoso (Q muito alto), selecionando casualmente um projeto de pesquisa (ideia) com potencial impacto alto. Por outro lado, qualquer cientista, mesmo com um talento para pesquisa (Q alto), poderia ainda publicar um artigo de baixo impacto, bastando para isso que tenha selecionado um projeto não interessante para sua comunidade. O Q dos cientistas é relativamente o mesmo ao longo das suas carreiras.

O Algoritmo 4 computa o Q do cientista i, a partir da sua lista L de publicações. $C(\alpha, t)$ é uma função que recebe como parâmetro um artigo α e um inteiro t, e retorna a quantidade de citações do artigo α passados t anos da sua publicação. Por último, o m representa quantidade

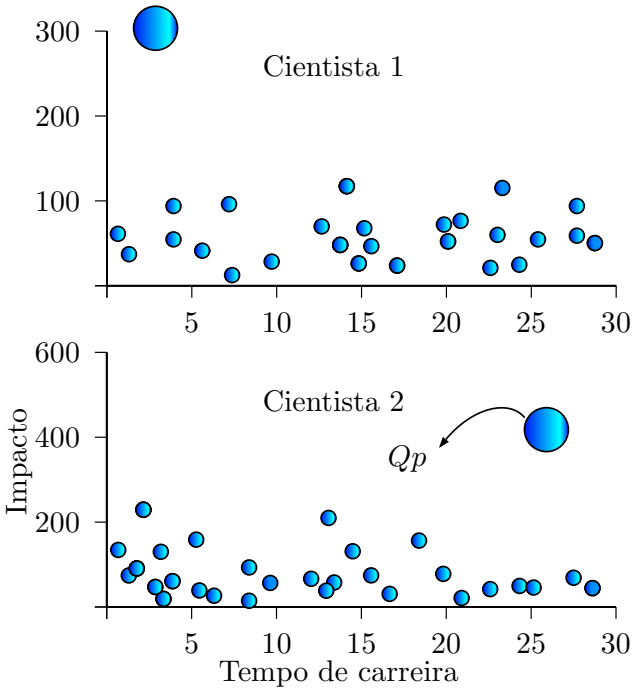


Figura 2.2: Cada ponto representa um artigo da lista de publicações do cientista correspondente. O eixo y mede o impacto de cada artigo. No caso do cientista 1 seu melhor artigo foi no começo da carreira e do cientista 2 no final da carreira.

de artigos com mais de 1 citação após sair do laço **para**.

Algoritmo 2 Computa o Q do cientista.

Entrada: L = $\{1, 2, \cdots, n\}$ e t
Saída: o Q do cientista i, caso i tenha pelo menos 1 artigo com mais de uma citação e zero caso contrário.

```
1    m ← 0
2    soma ← 0
3    para cada artigo α na lista L faça
4        qtd ← C(α, t)
5        se qtd > 1 então
6            soma ← soma + log_e qtd
7            m ← m + 1
8    se m > 0 então
9        retorne exp(soma/m)
10   senão
11       retorne m
```

O fator Q de um cientista pode ser usado para:

1. Calcular como o impacto do artigo de maior sucesso de um cientista é esperado mudar com produtividade, e também,

2. Predizer seu índice-h futuro e sua produtividade esperada.

Recentemente, alguns pesquisadores têm feito comentários críticos sobre o fator Q.

Como por exemplo, a dificuldade de computar o Q do cientista júnior, mencionado por Zeng, Shen et al. (2017), devido ao seu cálculo depender pesadamente de citações. Quanto a esta questão, no Capítulo 4, eu proponho novas equações para estimar o Q do cientista júnior que não depende pesadamente de citações.

Adicionalmente, é importante destacar o comentário feito por Sugimoto (2021), que critica o fato do modelo depender da suposição de que todos os cientistas têm acesso ao mesmos recursos, ignorando as disparidades massivas entre países e instituições.

Na próxima seção, discute-se como um pesquisador na base de dados é classificado como um pesquisador júnior, nesta tese.

2.2.2 Definição de Pesquisador Júnior

A definição de pesquisador júnior é central para essa tese, tentando provar que o desempenho individual de um cientista júnior pode ser predito a partir suas características bibliométricas, mesmo com dado limitado.

Muitos estudos têm tentado propor uma definição que seja amplamente aceita. Entretanto, quando começa ou termina a fase de pesquisador júnior ainda não é claro.Alguns autores, por exemplo Li, Aste et al. (2019), classificam um cientista como um pesquisador júnior pelos três primeiros anos desde a sua primeira publicação. Essa definição simples é a mais comum. Porém o tempo que conta como a fase de pesquisador júnior tem variado entre diferentes atores (instituições de pesquisa, agências de fomento à pesquisa). E, é ás vezes entre três e oito anos após o término do período de doutoramento do pesquisador (BAZELEY, 2003).

Uma definição mais complexa foi dada por Laudel e Gläser (2008), ele posicionou a fase de pesquisador júnior dentro de um arcabouço teórico definindo carreiras científicas, e identificou que, durante a fase de pesquisador júnior, ocorre uma mudança de status do pesquisador, de um pesquisador conduzindo pesquisa sob a orientação de outros para a de um pesquisador que desenvolve pesquisa autonomamente.

O momento exato dessa passagem não é claro, e pode variar bastante de cientista para cientista. Segundo Laudel e Gläser (2008), a fase de pesquisador júnior pode ser prolongada caso o pesquisador não atinga sua autonomia. E, isso pode acontecer por várias razões, como, por exemplo, pouco tempo para se dedicar a pesquisa científica depois do término do doutorado ou alguma interrupção programada pelo cientista.

Nesta tese, nós classificamos um pesquisador como pesquisador júnior pelos cinco primeiros anos desde a sua primeira publicação, similar à classificação de cientistas júnior dada por Lindahl (2020) e Lee (2019). Nós entendemos que esse tempo é um tempo razoável.

2.2.3 Definição de Sucesso Científico

Para entender esse conceito, é importante notar que a validação de um resultado científico é sempre coletiva, no sentido que ele tem sido escrutinado, criticado e validado por um número

de pares.

Nesta tese, sucesso capturado por fama, celebridade, popularidade, impacto ou visibilidade, é uma medida coletiva que captura as reações da comunidade científica para o desempenho individual de uma entidade (e.g., um pesquisador, uma instituição de pesquisa, etc.) (YUCESOY; BARABÁSI, 2016).

Um estudo anterior, por Yucesoy e Barabási (2016), analisou o grau de concordância entre o desempenho de um indivíduo e seu sucesso. Os autores concluíram que na maioria das áreas de alcance humano (e.g., esporte, ciência), visibilidade excepcional pode ser explicada por medidas de desempenho detectáveis.

Diferentemente de uma medida de sucesso, uma medida de desempenho captura a totalidade de alcances mensuráveis objetivamente em um certo domínio de atividade (e.g., o total de publicações de um cientista ou o número de campeonatos vencidos por um time de futebol), capturando as ações de uma entidade individual (e.g., o pesquisador ou o time de futebol) (LEHMANN; JACKSON; LAUTRUP, 2006).

Portanto, citação é uma medida de sucesso porque captura as reações da comunidade do cientista para o seu desempenho de pesquisa. Consequentemente, outras medidas baseadas em citações (e.g., o Q de um cientista, o índice-h) são também medidas de sucesso. Nesse trabalho, assume-se isso.

Assim como, o índice-h esperado para um cientista é uma medida de seu sucesso futuro, porque essa medida indiretamente captura a reação esperada da comunidade para o seu desempenho individual nos anos que virão.

O sucesso de um professor em uma universidade pode ser medido por diversos ângulos diferentes daquele focado em impacto de pesquisa (citação). O total de alunos de doutorado/-mestrado orientados e o total de projetos de pesquisa finalizados são algumas das medidas de desempenho individual alternativas. O reconhecimento da comunidade para o desempenho do docente pode ser medido através de números de convites recebidos para participação em bancas diversas, totais de indicações para coordenar cursos, departamentos, setores administrativos da Universidade, etc.

Por último, é importante destacar que por toda a tese, usa-se o termo sucesso e impacto intercambiavelmente, seguindo a mesma prática adotada por outros autores, por exemplo Kanellos et al. (2019) e Yucesoy e Barabási (2016).

2.2.4 Classificação de Medidas

As medidas podem ser classificas quanto a sua natureza em cumulativas e não cumulativas.

Medidas cumulativas (e.g., o número de publicações, o índice-h e o total de citações para uma publicação) são aquelas que aumentam com o passar do tempo, isto é, dependem da idade acadêmica da pessoa, e por isso não são adequadas para avaliar o desempenho/impacto de pesquisadores júnior ou o impacto recente de cientistas (WILSON; TANG, 2020). Tais medidas favorecem pesquisadores mais velhos, e o seu uso, como critério para progressão de carreira, financiamento de pesquisa e seleção de membros de quadros editoriais, têm sido questionado.

O índice-h de Hirsch (HIRSCH, 2005) é uma dessas medidas, um índice que ajuda a qualificar o impacto dos trabalhos publicados por pesquisador e se o impacto se concentra em poucos ou vários trabalhos. O índice depende do total de publicações de um cientista e do impacto das publicações. Ambas as medidas compondo o índice são também medidas cumulativas. A Figura 2.3 ilustra o comportamento comum do índice-h para um cientista em função de sua idade acadêmica.

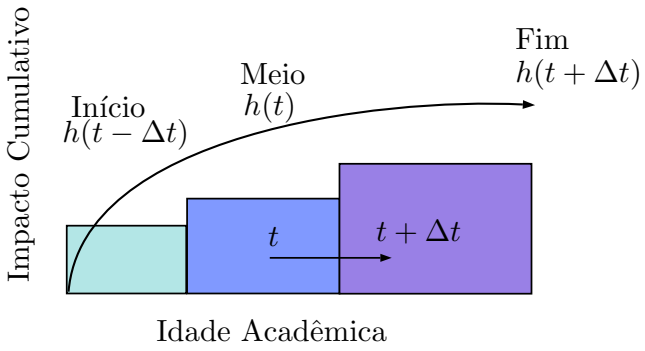


Figura 2.3: O índice-h depende da idade acadêmica do pesquisador. Além disso, o índice-h pode aumentar sem que haja a produção de novo conhecimento. Estas características torna o índice inadequado para avaliar cientistas em diferentes estágios da carreira. Adaptado de Penner et al. (2013).

Por outro lado, o fator Q do cientista tem sido mostrado ser praticamente o mesmo por toda a sua carreira, e portanto, uma suposta medida não cumulativa. Teoricamente, isso significa dizer que para uma série de valores observados dessa métrica em diferentes fases da carreira do pesquisador (e.g., fase de pesquisador júnior, sênior), a média e a variância é praticamente a mesma.

Graficamente, a diferença entre uma medida cumulativa e uma não cumulativa pode ser vista na Figura 2.4.

O modelo teórico para evolução do índice-h do pesquisador (PENNER et al., 2013) é utilizado para descrever a evolução do índice-h, em que seu h em um dado ano é a soma de incrementos randômicos e independentes Δh. Portanto, para um dado pesquisador p, seu índice-h depois de t anos é dado pela equação 2.2.

$$h^p(t) = \sum_{i=1}^{t} \Delta h_i^p \tag{2.2}$$

Usando o modelo anterior, refizemos o impacto das publicações ano a ano do cientista e computamos a evolução do seu Q. A Figura 2.4 mostra a evolução do Q e do h do cientista lado a lado.

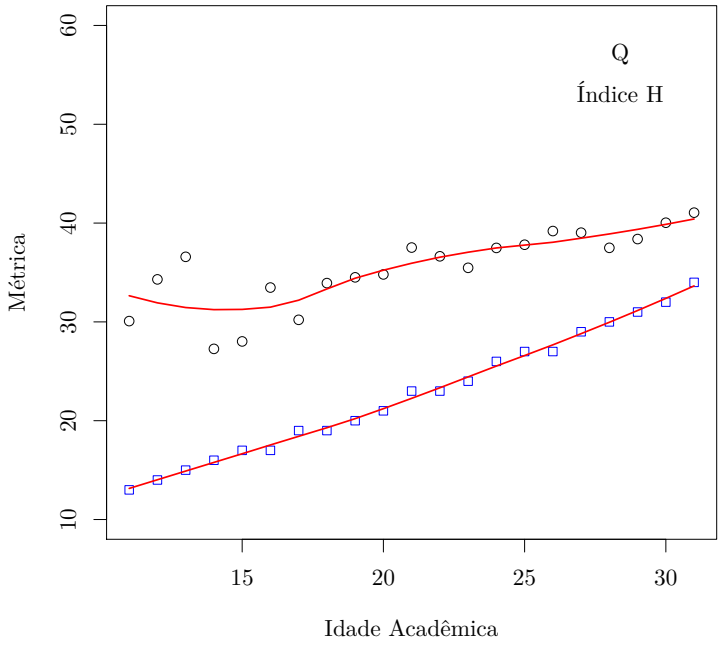


Figura 2.4: Comparativo da evolução das métricas Q e H para um cientista, desde o ano 10 da sua carreira até o ano 30. Por toda a carreira do pesquisador o Q é praticamente o mesmo, enquanto que o seu índice-h cresce com a idade acadêmica da pessoa.

2.3 Predição na Ciência da Ciência

A Predição do impacto futuro do pesquisador (HOU et al., 2019a; CLAUSET; LARREMORE; SINATRA, 2017) tem atraído um interesse considerável nos últimos anos. Isso se deve aos seus muitos usos no mundo real.

Predições fazem parte da vida do cientista, ele está sempre tomando decisões fundamentadas em tendências. Além disso, o desempenho futuro do cientista interessa a muita gente, como, agências de governo, instituições de pesquisa, e os próprios cientistas.

Nesta seção, é dada uma visão geral do tópico.

2.3.1 Definição do Problema Predição do Impacto Futuro do Pesquisador

Para definir o problema, assume-se que uma carreira científica é um empreendimento afetado por muitos fatores, e portanto, o potencial impacto futuro do cientista deve ser definido em termos de suas muitas diferentes características ligadas ao início da carreira. Uma abordagem de aprendizado de máquina supervisionada é usada para definir o problema.

Sejam $\{x_1, x_2, \cdots, x_n\}$ as entradas para um algoritmo de aprendizado de máquina e $\{y_1, y_2, \cdots, y_n\}$ sejam os alvos, em que $x_i \in \mathbb{R}^d$ representa as d características do cientista i relacionadas a sua fase de pesquisador júnior, e $y_i \in \mathbb{R}$ o valor de uma medida de impacto, calculada Δt anos depois da fase de pesquisador júnior.

O algoritmo tem como *tarefa de aprendizado* (objetivo), aprender uma função aproximada f para estimar y_i dado x_i e Δt (Equação 2.3). A função f deve ser avaliada sobre novos exemplos. E, deve ser possível extrair de f a contribuição de cada característica para uma predição.

$$f(y_i|x_i, \Delta t) \approx y_i \tag{2.3}$$

O maior desafio é identificar quais características x_i do cientista usar, mas antes é preciso reuni-las em um único lugar. Dados relativos à ensino não são comumente vistos em bases de dados, a maioria delas reúne ou dados de redes de citação ou dados curriculares. Uma base de dados, mesmo incompleta, integrando todas essas informações em um só lugar ainda é só um desejo da comunidade.

2.3.2 Barreias à Aceitação de Modelos Preditivos

A ampla adoção de sistemas preditivos baseados em aprendizado de máquina depende da aprovação desses sistemas pela comunidade científica. Predições envolvem riscos. Além dos riscos usuais, como, a passagem de uma nova entrada muito específica para o modelo de aprendizado de máquina, gerando uma predição errada, existem outras preocupações também afetando a confiabilidade do modelo.

Esta seção, discute-se barreias à aceitação do uso de modelos preditivos pela comunidade científica. A descriminação não intencional por máquinas é uma delas.

Descriminação não Intencional por Algoritmos

Até agora é sabido que as medidas cumulativas, como o índice-h, prejudicam pesquisadores mais jovens. O que tem sido descoberto recentemente é que elas também causam efeitos colaterais nas predições de modelos (PENNER et al., 2013). Os autores da pesquisa atribuem a natureza cumulativa do índice-h a introdução de vieses no modelo do índice-h futuro do cientista de Acuna (ACUNA; ALLESINA; KORDING, 2012)(e.g. o modelo é mais preciso quando predizendo o impacto futuro de pesquisadores mais velhos).

A discriminação não intencional por modelos, como o caso do modelo do Acuna, é inaceitável, e torna modelos preditivos não confiáveis.

Os dois próximos tópicos discutidos são exemplos de situações difíceis de serem notadas, e que podem causar prejuízos ao cientista, porque isso reduz o seu impacto levando o modelo a predizer valores mais baixos:

- Se um dos artigos da lista de artigos do cientista é *Sleeping Beauty* (RAAN, 2004). Sleeping Beauties é um termo cunhado para descrever um artigo cuja relevância não tem sido reconhecida por décadas, mas então, de repente, torna-se altamente influente e citado.
- Se o desempenho ruim de um cientista é um resultado de uma interrupção na carreira (programada).

Sleeping Beauties

De fato, mesmo a descoberta mais profunda passa despercebida se sua importância não é reconhecida através de discussões, palestras e citações pela comunidade científica. O artigo *paradox* (EINSTEIN; PODOLSKY; ROSEN, 1935), é um caso desse, e um exemplo de *Sleeping Beauty* (KE et al., 2015).

Estudos anteriores (KE et al., 2015; RAAN, 2004; LI; YE, 2016) concluíram que há poucos exemplos desses. Geralmente, as publicações científicas seguem um padrão: referências para elas atingem seu máximo alguns anos após a publicação, e em seguida, diminuem continuamente.

Embora *sleeping beauties* sejam raras, isso pode impactar negativamente a avaliação da confiabilidade do modelo. Devido a sua similaridade com outros artigos no primeiros anos de vida, *sleeping beauties* são difíceis de serem diferenciados dos outros, antes de ganharem popularidade. Certamente, se os algoritmos de aprendizado de máquina, soubesse delas, durante a fase de treino, criariam máquinas mais precisas.

Muito pouco é conhecido sobre *sleeping beauties*. Mais pesquisa é necessária para analisar as estatísticas de *sleeping beauties* para diferentes campos e estudar a possível influência de fatores específicos (e.g., tipos de periódicos , impacto do autor, colaboração).

Interrupções na Carreira e as Barreiras Existentes para Grupos sub Representados

Interrupções na carreira também são muito difíceis de serem preditas. Uma carreira individual pode ser interrompida por várias razões. Períodos de descontinuidade em carreiras

de pesquisa são comuns, e algumas paralisações são forçadas (e.g., causadas por pandemias e guerras, desastres naturais) e outras programadas (e.g., maternidade, paternidade) pelo pesquisador. A produtividade mais baixa durante a interrupção torna avaliações de desempenho (impacto) passado de pesquisadores enganosas.

Tem sido predito que qualquer interrupção, seja ela de qualquer natureza, exerce em algum grau um impacto sobre as atividades de pesquisa de pesquisadores júnior, e com consequências profundas (e.g., interrupção de pesquisa, ansiedade sobre incertezas da carreira futuro).

A pandemia tem paralisado ou reduzido as pesquisas em laboratórios de todo o mundo, cortado orçamentos e ameaçado a disponibilidade de subsídios, bolsas e outras fontes de financiamento de pós-doutorado. Um estudo anterior, por Herman et al. (2021) e Woolston (2020), concluiu que pesquisadores júnior são desproporcionalmente afetados pela pandemia em curso e carregam o fardo das dificuldades decorrentes dela.

McElrath (1992) mediu o grau de concordância entre o total de pareceres desfavoráveis ás mulheres, relativos ás decisões de aprovação em estágios probatórios, e o número de interrupções em sua carreira de pesquisadora, e concluiu que quando a mulher interrompe sua carreira ou muda de emprego acadêmico os efeitos negativos sobre a sua aprovação no estágio probatório (i.e., o direito de permanecer empregada até a aposentadoria) são substanciais. Há poucas evidências em seus dados de que os homens estejam em desvantagem semelhante.

Caso as disparidades entre homens e mulheres na ciência continuem a ser tratadas como naturais, certamente o *gap* de gênero tende a aumentar e a continuar prejudicando as mulheres na ciência. Caso isso não seja corrigido, o potencial para impacto futuro de jovens pesquisadoras seria enganoso.

Medidas Infladas

A confiabilidade do modelo também é afetada por medidas infladas. Medidas baseadas em citações sofrem desse viés. Isso é causado por um fenômeno conhecido de inflação de citação (LARIVIÈRE; ARCHAMBAULT; GINGRAS, 2008). Esse viés temporal, surge do fato que a literatura científica está constantemente crescendo, em aproximadamente 4% ao ano, e a produção total de citações dobra a cada 12 anos (PETERSEN et al., 2019).

Por um lado, esse crescimento mostra que mais pessoas estão desenvolvendo pesquisas científicas, e como um resultado, mais pessoas tem se beneficiado disso. Por outro lado, tem sido mostrado que isso tem reduzido o valor real de um citação, i.e., o valor de fato de uma citação depende de quando ela foi produzida. Pesquisadores contemporâneos tem citado mais do que seus antecessores, e isso, tem efeito sobre as medidas de impacto baseadas nelas (e.g., índice-h).

Portanto, inflação de citação é um aumento geral na produção total de citações e a depreciação no valor de uma citação (LARIVIÈRE; ARCHAMBAULT; GINGRAS, 2008). Tem sido mostrado que isso afeta as comparações de pesquisadores vivendo em momentos diferentes, algo ainda muito comum de ser visto no dia a dia de avaliadores.

Petersen et al. (2019) chama atenção para a necessidade de uma revisão dos métodos de contagem de referências usados para avaliar o impacto de citação de entidades individuais,

principalmente, daqueles cujos os registros de publicações se espalham por várias décadas, e isso parece ser uma necessidade urgente.

Dependendo da abordagem adotada, longitudinal ou mais estreita, tomada com relação a base de dados, o efeito da inflação de citação impacta mais ou menos. A maioria dos trabalhos em predição na ciência foca na segunda abordagem. Essa última restringe a análise ao estudo de carreiras individuais extraídas de um período curto (e.g., que iniciaram no mesmo ano.).

Nesta tese, optou-se por uma abordagem diferente da anterior, por que decidiu-se analisar as carreiras de pesquisadores júnior que durante os primeiros cinco anos da carreira trabalhou com um pesquisador com uma presença duradoura na academia, isso gerou amostras muito pequenas quando encurtando o período de análise.

Recentemente, o uso de predição na Ciência da Ciência tem sido revisada, por Hou et al. (2019a), e nada tem sido mencionado sobre o valor de uma citação para uma predição quando ela origina-se a partir de diferentes épocas. Nesta tese, isso tem sido uma preocupação constante e temos feito experimentos específicos para medir isso, e concluímos que os modelos testados implicitamente classificam os pesquisadores por época. Quanto mais recente for a época, maior é o impacto futuro estimado. No entanto, mais estudos focados e aprofundados precisam ser feitos.

Falta de Justificativas para uma Predição

Falta de Justificativas é o principal entrave para o aceite de sistemas preditivos. Para uma faixa considerável de aplicações (e.g., carros autônomos, predição de impacto futuro de cientistas e antecipação de diagnósticos de doenças), um compreensão do porquê de um predição particular para uma dada entrada é fundamental (RIBEIRO; SINGH; GUESTRIN, 2016; CARVALHO; PEREIRA; CARDOSO, 2019). Vários estudos, por exemplo Ribeiro, Singh e Guestrin (2016), Carvalho, Pereira e Cardoso (2019) e Zhang, Wang et al. (2020), têm sido conduzidos sobre o assunto. Em sua análise crítica de métodos de interpretação (ou entendimento) de modelos caixas pretas, e.g., Kim, Khanna e Koyejo (2016) aponta para a necessidade de explicativas mais profundas, e.g., eles defendem que mais informações, e sobretudo críticas ao modelo devem ser disponibilizadas para um entendimento mais amplo das decisões dessas máquinas por humanos.

Portanto, existe ainda uma necessidade por explicativas claras de modelos de aprendizado de máquina (i.e., modelos inferidos a partir de dados) (RIBEIRO; SINGH; GUESTRIN, 2016). Técnicas de redes neurais profundas têm alcançado acurácias preditivas absurdamente altas, e em muitos casos, em igualdade com o desempenho humano (MONTAVON; SAMEK; MÜLLER, 2018). Contudo, isso é muitas vezes conseguido ao custo de interpretabilidade (ZHANG; WANG et al., 2020).

Um número de trabalhos anteriores (ZHANG; WANG et al., 2020; CARVALHO; PEREIRA; CARDOSO, 2019; GEVREY; DIMOPOULOS; LEK, 2003; EL HECHI et al., 2021) tem tentado contornar essa barreira, para conseguir modelos mais claros eles têm, por exemplo:

1 - incorporado interpretabilidade diretamente na estrutura do modelo, o que resulta eventu-

almente em perdas de acurácia.

2 - buscado caracterizar o comportamento caixa preta do modelo, sem tentar elucidar seu funcionamento interno.

O Esquema na Figura 2.5 ilustra como se consegue interpretabilidade sem modificar o modelo. Um interpretador g destaca as características (relaçando ela) contribuindo mais (a intensidade da cor significa mais importante) para um predição a partir da entrada.

Finalmente, se planejamos tomar decisões acionáveis a partir de modelos de aprendizado de máquina, então nós precisamos entender o raciocínio por trás de uma decisão do modelo.

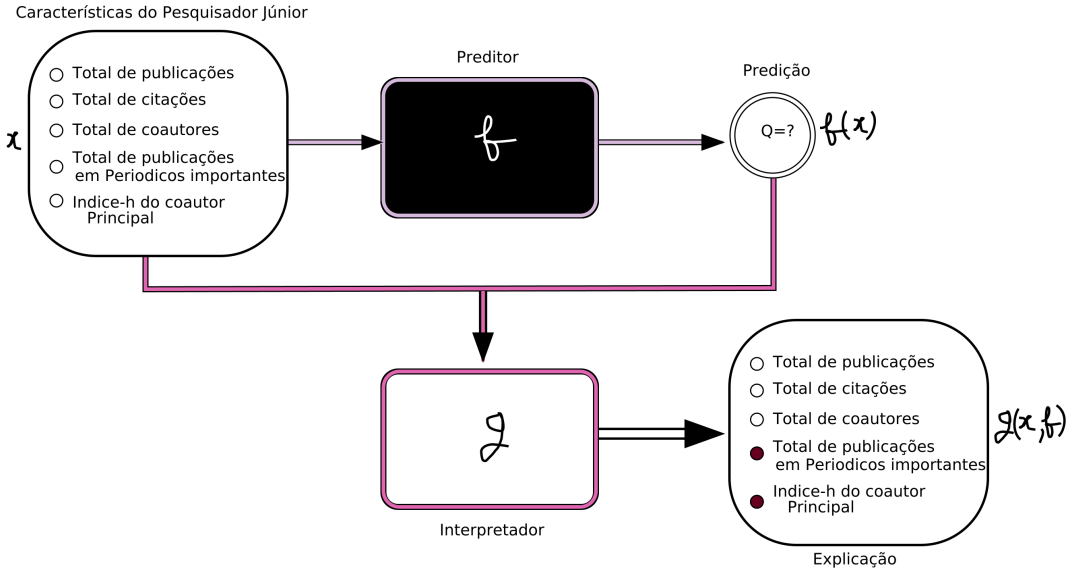


Figura 2.5: Fluxo de trabalho de um sistema de aprendizado de máquina interpretável.

Eventos Atípicos e Manipulação de Indicadores

Modelos de aprendizado de máquina em geral falham quando em contato com exemplos muito específicos. Os grandes eventos nunca são captados pelos modelos. Eventos atípicos como Covid, gravidez, guerras podem acelerar certas pesquisas, mas também paralisar outras. Definitivamente, modelos não lidam bem com tais eventos. Modelo nenhum poderia ter previsto uma pandemia como a Covid e nem as consequências disso. O investimento direcionado para certas pesquisas relacionadas a covid pode em teoria prejudicar severamente certos grupos, principalmente jovens cientistas. Muitos podem não mais continuar a carreira. Exemplos como esses, certamente, não poderiam fazer parte do conjunto de teste antes da Pandemia da Covid porque não se sabia da sua existência. Portanto, qualquer modelo falharia em tese porque a suposição de que o conjunto de teste é uma aproximação para exemplos não vistos falharia.

Uma outro caso de evento atípico é o uso intencional de autocitação por pesquisadores para melhorar seus indicadores. Modelos têm como entrada indicadores do cientista e como saída estimativas de indicadores futuros. Portanto, indicadores inflados causam distorções nas predições. Método de detecção de autocitação (BARTNECK; KOKKELMANS, 2011) devem

ser aplicados aos indicadores passados para o modelo como entrada. O funcionamento correto do modelo depende disso.

Nesse contexto, algo também afetando os indicadores de entrada passados para o modelo é a divisão dos créditos de uma pesquisa. Métodos de alocação de créditos em ciência (SHEN; BARABÁSI, 2014) devem ser aplicados para evitar problemas de autoria honorária e fantasma. A autoria honorária é aquela dada a um indivíduo apesar da falta de contribuições substanciais para o projeto de pesquisa. A autoria fantasma é essencialmente o oposto da autoria honorária, i.e., um autor dá uma contribuição significativa para um manuscrito e não tem ela reconhecida.

2.4 Considerações Finais

Trabalho complicado é o trabalho da previsão, sobretudo quando envolve decidir o futuro de alguém. Por outro lado, a disputa por recursos de pesquisa tende a aumentar nos próximos anos. Diante disso, os juízes, muito criticados, precisam de mais elementos para subsidiar suas decisões sobre a divisão dos recursos. Por isso, universidades e agências de fomento buscam por métricas mais completas para avaliar pesquisadores, i.e, métricas que vão além do aspecto científico, i.e., também provenha algo sobre o futuro do cientista quanto a outros aspectos, como potencial de ensino, de inovação, etc.

Neste capítulo, discutiu-se as muitas variáveis complexas que podem causar uma predição errada sobre a carreira de alguém. Avaliadores lidam com dados limitados, por exemplo, nenhuma base de dados reúne informações completas (sobre pesquisa, ensino e impacto social) sobre um pesquisador. Pensando nisso, este trabalho tenta quantificar o quanto do sucesso futuro do cientista iniciante pode ser explicado com a informação disponível nessas bases de dados.

Devido as especificidades de cada campo científico, um modelo único é desaconselhado. Dentro da própria ciência existem muitas outras. Cientistas das Ciências Humanas, Sociais Aplicadas e Linguística, Letras e Artes citam mais livros do que artigos científicos. Portanto, modelos do impacto futuro inferidos a partir de dados de cientistas da computação não são adequados para eles, porque cientistas da computação comportam-se de forma bem diferente, citando mais artigos do que livros.

Em ciência, o próprio cientista define a sua agenda de trabalho. Por isso, não é algo simples prever a sua agenda futura. Além disso, cientistas podem mudar de área. Ser área migratória pode gerar interferências. Por exemplo, um biólogo que mudou de área e agora fazendo pesquisa em Cientometria pode ter seu impacto afetado simplesmente porque as duas áreas têm um padrão de citação diferente. Mas ainda pode ser mais complexo, para um estatístico, na área de medicina, a sua contribuição na estatística pode não ser considerada como principal contribuição.

Estudos preliminares apontam diversos ganhos em ser área migratória. No entanto, é muito difícil precocemente identificar que alguém deve mudar de área. O fato é que a troca de área pode afetar substancialmente os rumos de uma carreira. Consequentemente, modelos do impacto futuro podem prever um futuro muito diferente da realidade para esses casos, e outros, que

destoam da maioria dos cientistas.

Capítulo 3

Conjuntos de Dados

3.1 Introdução

A comunicação acadêmica tem sofrido grandes mudanças ao longo das últimas décadas. Agora, cientistas têm à disposição novos canais de publicação e disseminação de suas descobertas científicas (BORGMAN; FURNER, 2002).

Essa informatização têm mudado a ciência. A comunicação escrita científica, na forma de artigos científicos, está agora sendo registrada *online* em uma escala massiva (LETCHFORD; MOAT; PREIS, 2015), gerando grandes conjuntos de dados. A disponibilidade desses dados para a comunidade científica tem oferecido um amplo leque de oportunidades de pesquisa e resultado em mais descorbertas, como o modelo do índice h futuro de Acuna e seus colegas (ACUNA; ALLESINA; KORDING, 2012) e muitos outros modelos (SARIGÖL et al., 2014; VAN DIJK; MANOR; CAREY, 2014) que suportam a ideia de que o sucesso futuro de um cientista depende em parte de fatores associados as suas primeiras publicações.

Além disso, diversas hipóteses têm sido confirmadas, por exemplo: se o sucesso de artigos pode ser parcialmente predito por seus sucessos de curto prazo (ABRAMO; D'ANGELO; FELICI, 2019), e se o tamanho do título de um artigo pode ser um bom preditor de sucesso futuro (LETCHFORD; MOAT; PREIS, 2015) .

Adicionalmente, eles também tem sido usados em outros estudos, como por exemplo por Brembs, Button e Munafò (2013) e Saha, Saint e Christakis (2003) para verificar se o uso de *rankings* de periódicos em avaliações de cientistas como um critério para promoção docente é válido.

Esse capítulo dá uma visão geral dos dois conjuntos de dados usados nesse trabalho: ACM e APS. Os dois conjuntos têm sido muito utilizados na literatura. Eles são usados nesta tese para:

1. reconstruir carreiras científicas (Capítulo 4 e 5), e

2. reconstruir comportamentos de publicação de pesquisadores por periódico (Capítulo 6).

Por último, algumas conclusões são tiradas e apresentadas na última seção.

3.2 ACM

A *Association for Computing Machinery* (ACM), fundada em 1947, é uma associação reconhecida internacionalmente por profissionais da área da computação por contribuir com o desenvolvimento e difusão do conhecimento da computação.

O conjunto de dados da ACM contém os relacionamentos de citação de artigos publicados pela ACM. Ele é altamente longitudinal com informação completa sobre as publicações de cientistas da computação de todo o mundo. O conjunto de metadados descrito por esse conjunto de dados inclui o título, a lista de autores, o ano de publicação, o local de publicação, a lista de referências e o resumo de cada publicação. Eles apresentam uma alta qualidade, uma vez que são providos pela própria ACM ao invés de serem extraídos unicamente a partir das próprias publicações.

Este conjunto de dados tem quase 2.5 milhões de artigos científicos abrangendo praticamente todas as áreas de pesquisa da computação. Eles foram escritos por aproximadamente 1.6 milhões de autores, e estão distribuídos entre 1935 e 2018.

O conjunto de dados tem sido usado para estudar diversos tópicos de pesquisa que vão desde a influência do tamanho do título de um artigo em seu sucesso futuro (LETCHFORD; MOAT; PREIS, 2015) até a recomendação de colaboradores acadêmicos (LIU; XIE; CHEN, 2018) e a predição de novas ocorrências de citações para um artigo (ABRAMO; D'ANGELO; FELICI, 2019).

A Figura 3.1 mostra um crescimento explosivo da literatura de ciência da computação nas últimas décadas, em particular nas mais recentes. Por exemplo, o número de publicações em 2010 é quase três vezes mais do que aquele de 10 anos antes. Um grande volume de publicações por ano tem criado esse cenário, e essa tendência de crescimento é esperada continuar dentro de um futuro previsível (YAN et al., 2011). Tem sido demonstrado que o aumento no número de pesquisadores em todo o mundo não é o único causador disso (KANELLOS et al., 2019). Também, a competição crescente que pressiona cientistas a continuamente produzirem resultados publicáveis, também conhecida como "publish or perish"e outros fatores têm contribuído para esse crescimento explosivo.

Na Figura 3.2 é mostrada a distribuição de citações na base de dados da ACM. A partir da figura, nota-se que ela possui uma cauda pesada, e.i., uma inclinação para a direita, indicando que um número muito pequeno de artigos são muito citados enquanto que a grande maioria das publicações são pouco referenciadas por seus pares. Esse padrão é visto por todas as áreas do conhecimento.

Como pode ser visto na Figura 3.3, enquanto a distribuição do total de artigos coautorados por Y autores, $N(Y)$, também segue o padrão observado para distribuição de citações, onde pouquíssimos artigos têm sido coautorados por um número maior do que 15 autores, a grande maioria deles têm sido coautorados por até 5 autores.

Cresceu nos últimos anos o número de publicações científicas assinadas por mais de mil autores, fenômeno conhecido como hiperautoria. Isso não é comum na Ciência da Computação. No entanto, como pode ser notado na Figura 3.3, existe um número expressivo de artigos

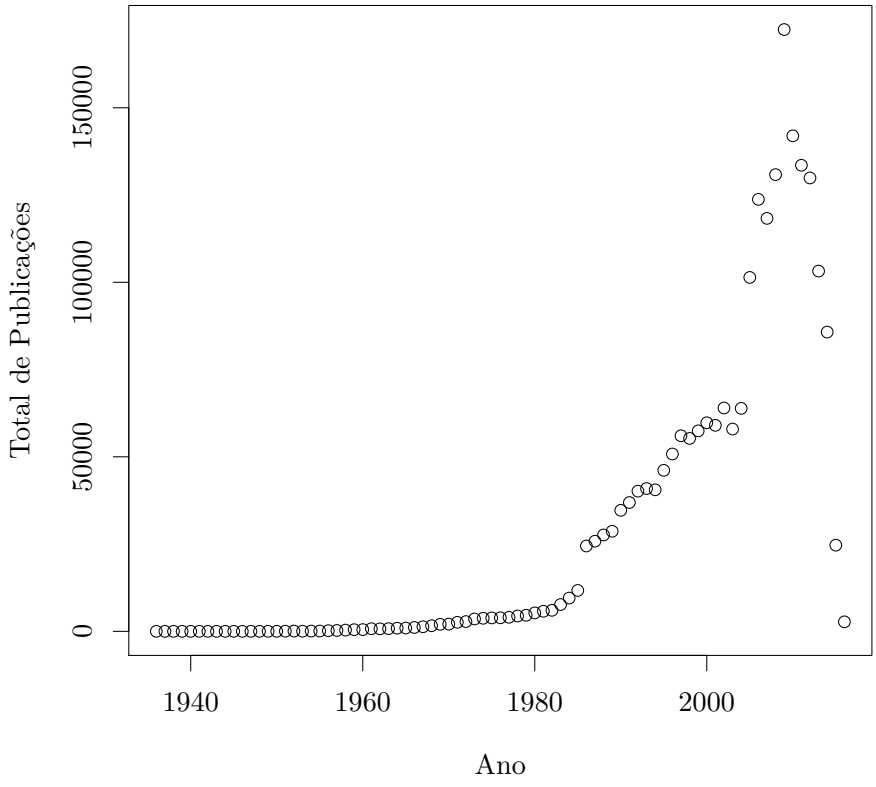


Figura 3.1: Crescimento da literatura de Ciência da Computação.

assinados por mais 50 autores.

Descobrir a função matemática que melhor descreve a distribuição de citações de publicações científicas é crucial. Estudos anteriores, por Seglen (1992), Redner (2005) e Perc (2010), têm sugerido diferentes funções. E, embora, na Figura 3.4, uma *lognormal* ajusta-se a distribuição de citações melhor do que uma lei da potência, o resultado de um teste estatístico de Vuong (VUONG, 1989) é inconclusivo, como pode ser visto na Figura 3.5. O dado sugere que por via de regra um modelo não é melhor do que o outro. A nula do teste não é rejeitada, A nula do teste verifica se as duas funções candidatas estão igualmente distantes no sentido *Kullback-Leiber* da distribuição de dados verdadeira. O teste de Vuong rejeita a nula se a estatística do teste, ratio log-likelihood, tende a +/- infinito.

E, apesar de saber que existe de fato uma grande quantidade de artigos que não são citados por várias razões, e.g, simplesmente por falta de qualidade ou originalidade, esse número impressiona na base de dados da ACM. Ele representa mais de 50 % porcento do dado, um total de aproximadamente 1 milhão e trezentos mil (1.379.133).

Por outro lado, um fenômeno geral em ciência é a tendência de crescimento do percentual de artigos coautorados e do número médio de autores por artigo. Esse crescimento é específico de campo. Na ciência da computação, como é mostrado na Figura 3.6, o percentual de artigos

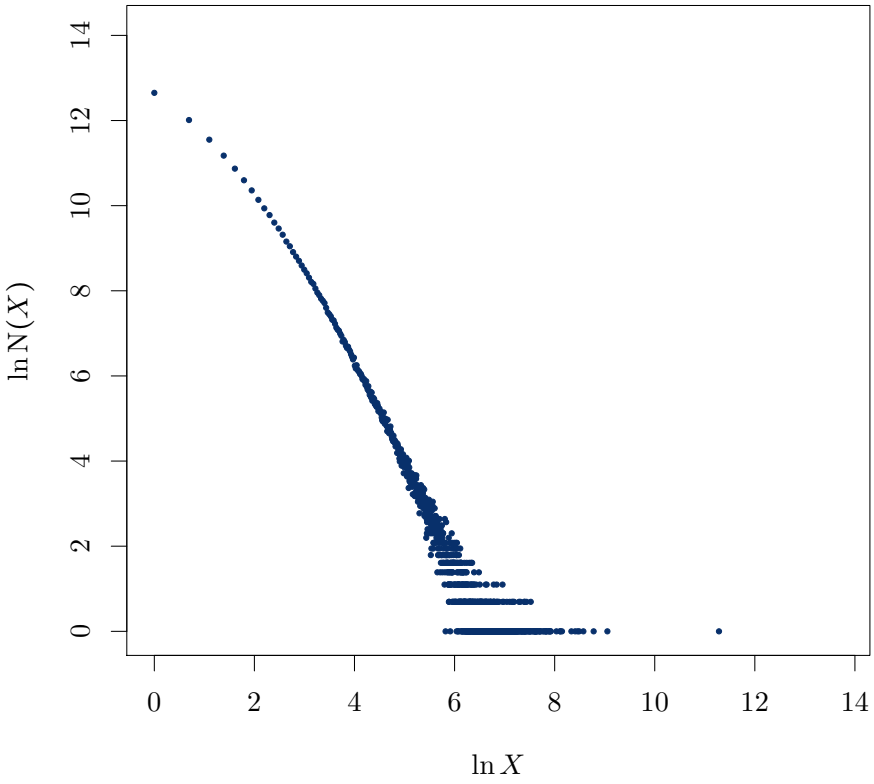


Figura 3.2: Distribuição de citação de 1.005.924 artigos com pelo menos uma citação na base de dados da ACM entre 1935 e 2017, em uma escala logarítmica dupla.

coautorados com 2, 3 ou 4 coautores tem sido cada vez mais comum, enquanto que o percentual de publicações solo tem desacelerado.

Como pode ser observado a partir da Figura 3.7, o número médio de citações por artigo por ano, ao longo de quase 7 décadas, tem aumentado. Esse aumento persistente da média, também conhecido como inflação de citação, afeta a análise longitudinal e a comparação de unidades (pesquisadores, periódicos, instituições de pesquisa) a partir de diferentes períodos (PETERSEN et al., 2019). Por isso, um número de trabalhos tem sido feito para corrigir o valor de uma citação, dado que o comportamento de publicação de um autor depende do período em que ele atuou ou atua, sendo muito desigual entre períodos.

Por último, para esse conjunto de dados não foi tratado o problema de resolução de nomes de autores. É usado rótulos numéricos disponíveis por Tang et al. (2008), que tem resolvido essa questão e provido identificadores únicos para os autores.

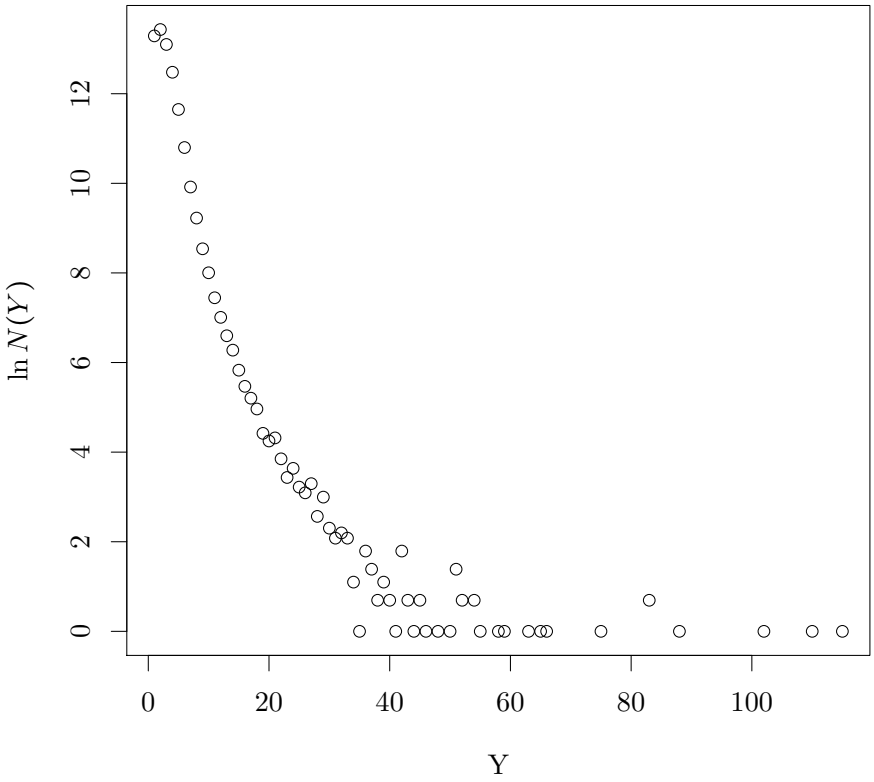


Figura 3.3: Distribuição do número de autores por publicação.

3.3 APS

A *American Physical Society (APS)*, fundada em 1899, é uma associação reconhecida internacionalmente por profissionais da área da física por contribuir com o desenvolvimento e difusão do conhecimento da área.

O conjunto de dados mantido pela APS* cobre mais de um século de artigos da física publicados em *Physical Review*, um coleção de periódicos revisados por pares publicando artigos científicos originais a partir de todas as áreas da física interdisciplinar, pura e aplicada. Em sua forma crua o conjunto de dados contem registros para artigos publicados nos periódicos de *Physical Review*, publicados de 1893 até 2018, cada um identificado com um rótulo numérico único. Para cada artigo os seguintes dados estão disponíveis: o título do artigo, a data de publicação, os nomes e afiliações de cada um dos autores, e a lista dos rótulos numéricos de artigos sendo citados (lista de referências).

Para o uso dessa base de dados, os autores precisaram ser extraídos a partir dos meta-dados dos artigos, e na fase de extração, o algoritmo 3, inicialmente desenvolvido por Martin et al. (2013) e melhorado por Sinatra et al. (2016), para resolução de nomes de autores foi aplicado. O

*https://journals.aps.org/datasets

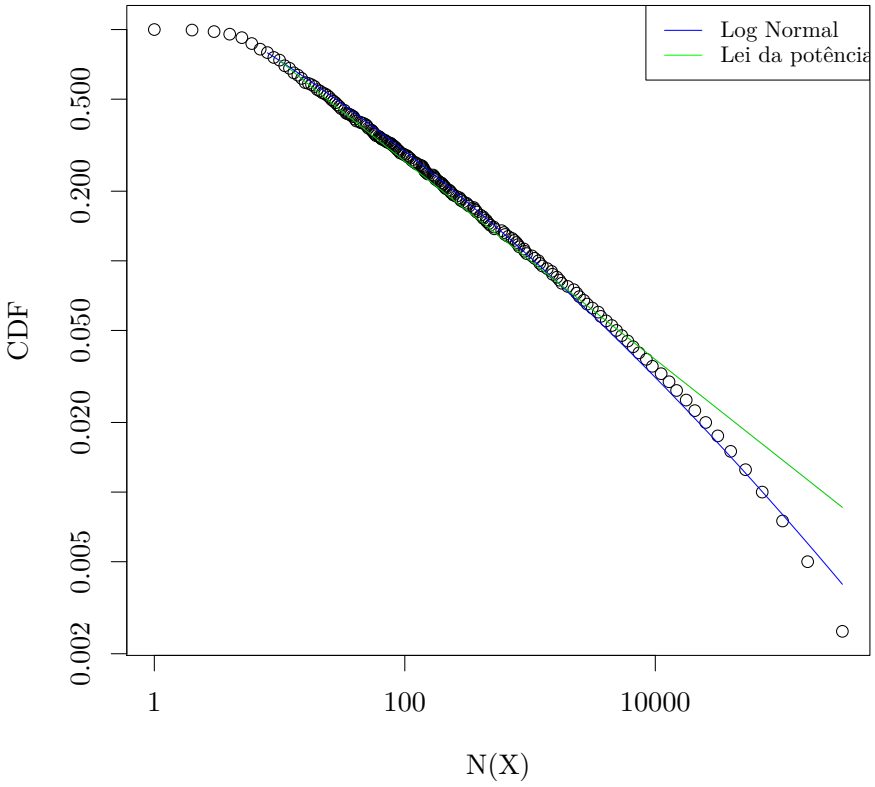


Figura 3.4: Comparativo das funções mais citadas na literatura para descrever N(X). N(X) é a quantidade de publicações com X citações. CDF é a função de distribuição acumulada.

algoritmo considera dois autores a mesma pessoa se possuem nomes completos correspondentes idênticos e compartilham uma afiliação, ou têm coautores em comum, ou citaram um ao outro.

Algoritmo 3 Desambiguador de nomes de autores.

Entrada: A lista A, contendo todos os nomes de autores encontrados no conjunto de dados da APS.
Saída: A lista L, contendo todos os pares de autores $\{P, Q\} \subset$ A, em que P e Q tratam-se da mesma pessoa

```
1   Inicialize a Lista L com o Conjunto Vazio
2   enquanto Existir Pares de Autores {P,Q} ⊂ A faça
3       se (P.UltimoNome = Q.UltimoNome e
            P.PrimeiraLetraDoPrimeiroNome = Q.PrimeiraLetraDoPrimeiroNome e
            (CompartilhamPeloMenosUmCoautor(P,Q) ou
            CompartilhamPeloMenosUmaAfiliação(P,Q) ou
            CitamUmAoOutroPeloMenosUmaVez(P,Q))) então
4               Adicione o Par {P,Q} a Lista L
5       fim se
6   fim enquanto
7   retorne L
```

O algoritmo não somente depende dos próprios nomes dos autores, mas também de padrões

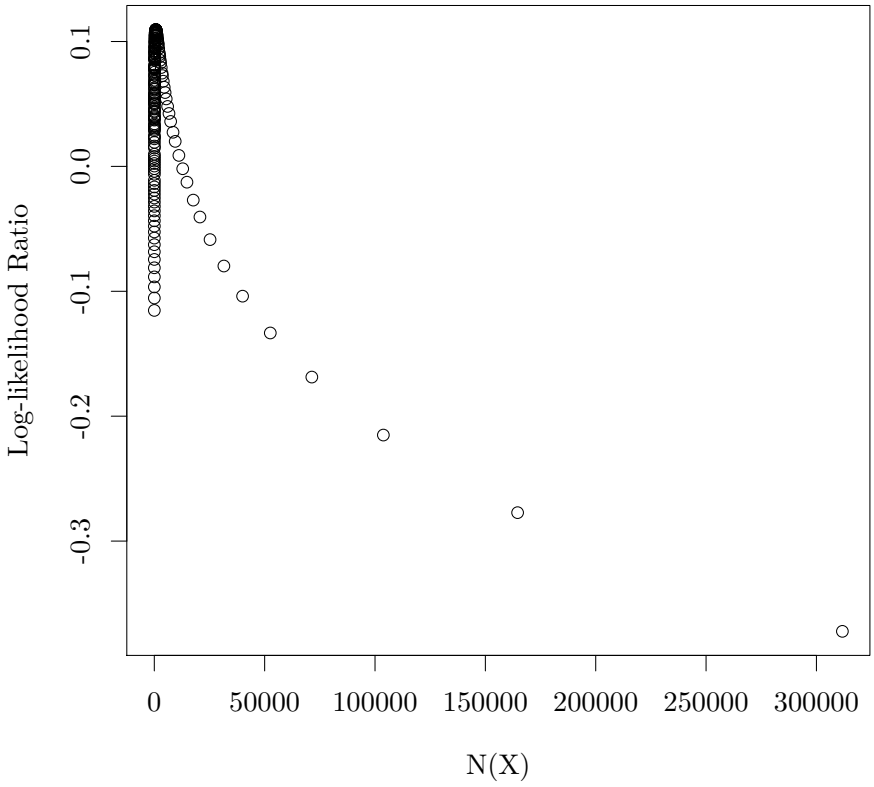


Figura 3.5: A estatística *log-likelihood ratio* não tende ao infinito.

de colaboração e afiliação, uma vez que autores com nomes similares que têm muitos dos mesmos colaboradores ou que estão na mesma instituição são mais prováveis de serem a mesma pessoa.

Um estudo anterior, por Martin et al. (2013), analisou minuciosamente esse conjunto de dados de 1983 até 2013, e concluiu que o conjunto de dados parece crescer exponencialmente, dobrando praticamente a cada 12 anos, e o número de citações por artigo dentro de *Physical Review* também parece crescer exponencialmente.

Parece que o comportamento de publicação de 2013 pra cá não mudou significativamente para alterar os achados de Martin et al. (2013), como mostrado na Figura 3.8. O volume de publicações por ano dentro do periódico *PRL* continua aumentando.

Zeng e Rong (2021) notaram que apesar da rede de citação do PRL está aumentando, a velocidade de crescimento desacelerou nos últimos anos. A Figura 3.9 mostra uma tendência clara de crescimento da média de citações recebidas por artigo por ano, insinuando que mais pesquisadores estão citando mais e recebendo mais citações anulamente.

Alem disso, nota-se a partir da Figura 3.10 que a probabilidade de novas coautorias múltiplas envolvendo mais de 10 autores é quase zero.

Finalmente, Letchford, Moat e Preis (2015) têm notado um desequilíbrio nas bases de dados de redes de citação em geral. Um vasto número de publicações é publicado por ano, alguns pou-

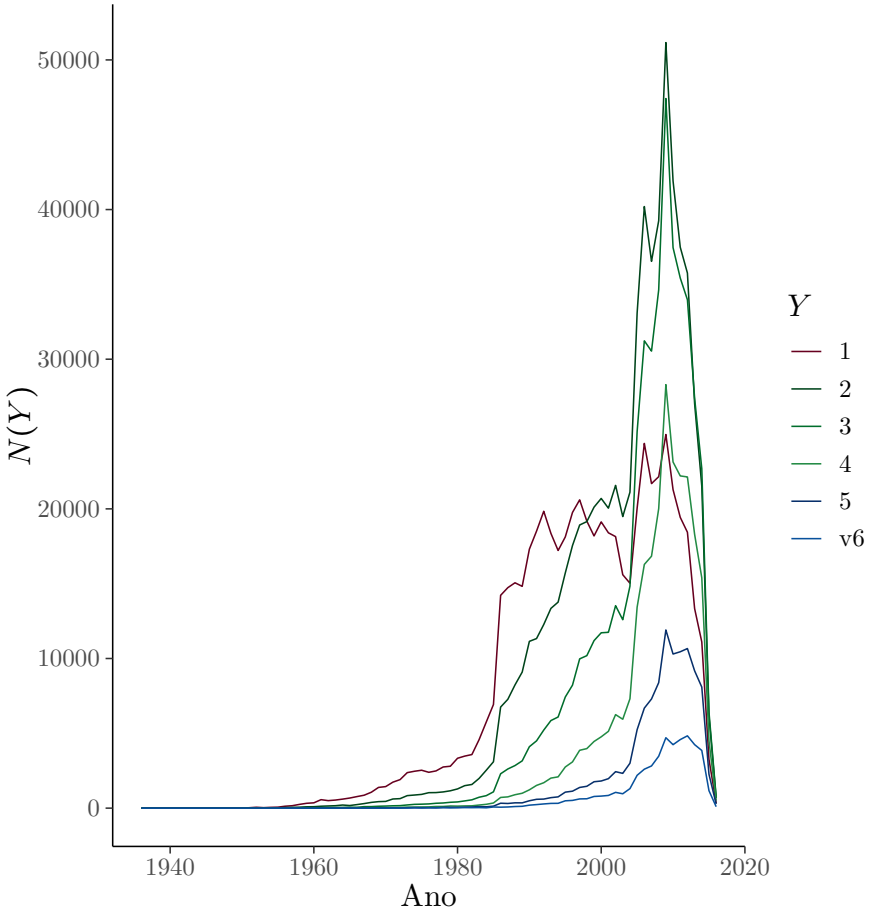


Figura 3.6: Tendência de coautoria múltlipa. N(Y) é o número de artigos com Y autores.

cos atraem atenção considerável, enquanto que outros (a maioria) passam quase despercebidos.

3.4 Considerações Finais

Progresso considerável tem sido feito com relação à predição de impacto futuro de entidades individuais (pesquisadores, artigos). Em grande parte, isso só tem sido possível graças à disponibilização de bases de dados de publicações/citações. Nesse capítulo, uma análise detalhada das bases de dados da ACM e APS foi apresentada. Elas foram usadas por um número de trabalhos no passado. E, tem sido usadas para suportar essa pesquisa.

As duas bases de dados apresentadas claramente têm algumas limitações. Apenas controlam citações de artigos indexados pelas próprias bases.

Em um nível mais amplo, mais pesquisa é necessária para determinar as razões de uma não citação, em uma tentativa de explicar o porquê de mais da metade de toda pesquisa publicada não ser citada.

Nós consideramos que essas bases de dados podem ser úteis para pesquisas em aprendizado de máquina, aprendizado profundo, processamento de linguagens naturais, grafos de conhecimentos, inteligência artificial, e sobretudo em análise de redes sociais acadêmicas e na construção

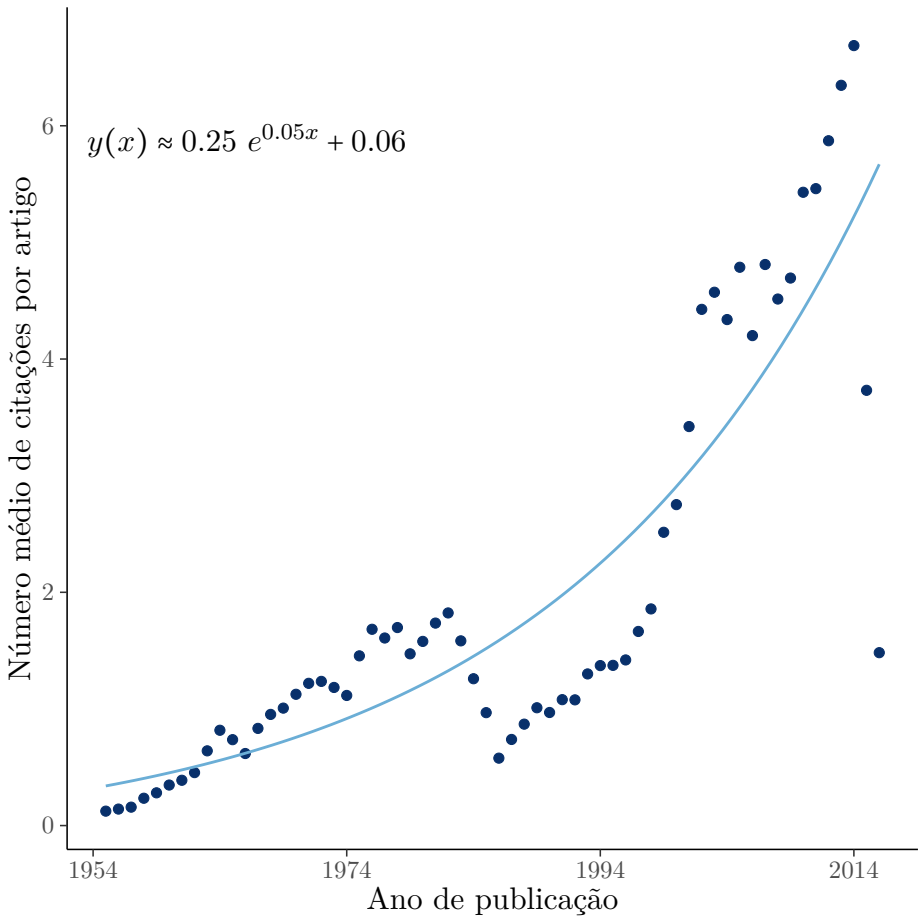


Figura 3.7: Número médio de citações por artigo por ano. Como indica a função exponencial $y(x)$ em que x é o ano, número tem crescido exponencialmente.

de sistemas baseados em aprendizado de máquina para auxiliar avaliações de pares.

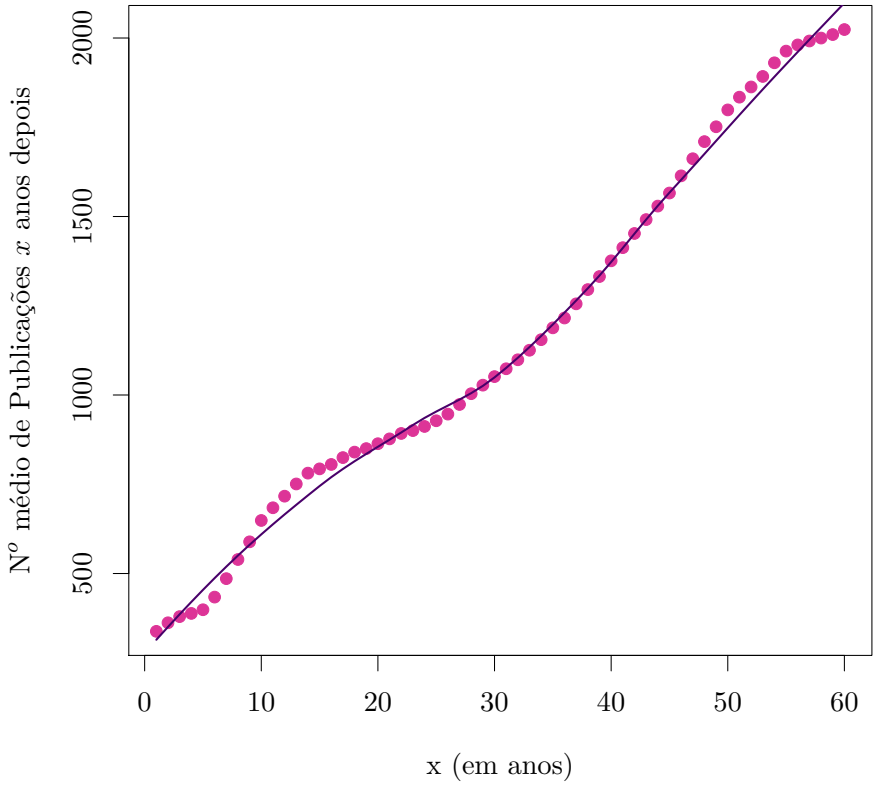


Figura 3.8: Crescimento do número de publicações no periódico PRL.

Figura 3.9: Crescimento da quantidade de citações recebidas por artigo por ano no periódico PRL. Para o cálculo da média considerou-se as citações recebidas por um artigo nos três primeiros anos de vida (depois da publicação).

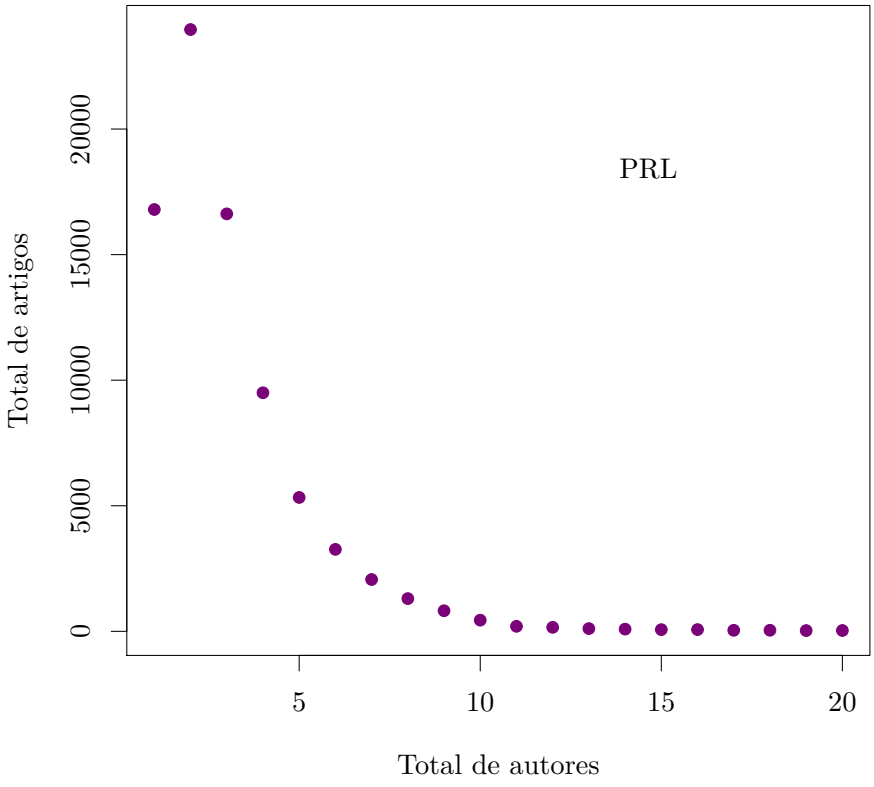


Figura 3.10: Distribuição do número de artigos pelo número de autores.

Capítulo 4

Equações para Prever o Sucesso Futuro de Pesquisadores Júnior

4.1 Introdução

Nas últimas duas décadas, a avaliação de pesquisadores júnior tem sofrido muitas mudanças, enfatizando julgamentos baseados em conteúdo e boas práticas científicas (SERRA et al., 2021; BARTNECK; KOKKELMANS, 2011), contrariamente ao uso de métricas baseadas no prestígio do periódico (GONZÁLEZ-PEREIRA; GUERRERO-BOTE; MOYA-ANEGÓN, 2010), tal como fatores de impacto. Pesquisadores oferecem explicações conflitantes sobre a utilidade de métricas.

Por um lado, estudos recentes (BORNMANN; WILLIAMS, 2017; LINDAHL, 2018) que examinaram muitos milhares de perfis de publicação de pesquisadores individuais, concluíram que o fator de impacto do periódico permitiu diferenciar entre pesquisadores com diferentes sucessos de publicação mais tarde e foi sim útil para identificar pesquisadores excepcionais (excelentes). Por outro lado, um outro estudo mais recente, por Brito e Rodríguez-Navarro (2019), concluiu que a métrica fator de impacto contradiz impacto de citação.

Comumente, agências de fomentação à pesquisa e comissões examinam a produção científica passada de pesquisadores júnior atrás daqueles com a potencialidade de se tornar um pesquisador proeminente (PENNER et al., 2013). E, apesar da significância disso, nós temos testemunhado um número reduzido de trabalhos.

Motivados por isso e pela disponibilidade cada vez maior de grandes bases de dados sobre publicações científicas, trabalhos anteriores (MAZLOUMIAN, 2012; DUCTOR et al., 2014; DONG; JOHNSON; CHAWLA, 2015) têm tirado proveito desse cenário e proposto novas medidas alternativas para estimar o potencial para impacto futuro de cientistas individuais. Neste contexto, Acuna, Allesina e Kording (2012) introduziram o modelo do índice h futuro de neurocientistas. No entanto, um exame detalhado desse modelo por Penner et al. (2013) mostrou que ele é enviesado para pesquisadores sênior, produzindo estimativas menos precisas para pes-

quisadores júnior.

Devido a suas presenças curtas na academia, pesquisadores júnior têm os históricos curtos (poucas publicações de sua autoria). Tem sido demonstrado que isso tem causado estimativas menos precisas dos seus impactos passados (e.g., o Q corrente de pesquisadores júnior) (ZENG; SHEN et al., 2017) e futuros (e.g., os índices-h futuros de pesquisadores júnior) (AYAZ; MASOOD; ISLAM, 2018). Especialmente, medidas em que a precisão aumenta com o acúmulo de citações para trabalho anterior do cientista, e.i, que dependem de tempo, como é o caso do Q (SINATRA et al., 2016) do cientista, têm a confiança nela reduzida devido a essa questão. No entanto, é importante saber o Q de pesquisadores jovens o quanto antes, uma vez que tem sido, conclusivamente, mostrado que o Q é uma medida não cumulativa, e.i, o seu valor é o mesmo por toda a carreira do pesquisador. Além de que, ele captura, em um número real, a habilidade de um cientista de transformar suas ideias em descobertas com um dado impacto medido por citações para elas.

O objetivo deste trabalho é apresentar novas equações estimadas para predizer o Q final de pesquisadores júnior, usando somente dados de publicações dos primeiros cinco anos de suas carreira. Usando exclusivamente métricas baseadas em publicações (e.g., o índice-h, o número de publicações), e informação inferida a partir dos meta-dados dessas publicações (e.g., o ano da primeira publicação de um autor como uma aproximação para sua idade acadêmica, seus coautores), estimar seu Q definitivo.

O diferencial deste trabalho está em considerar o impacto do colaborador principal do pesquisador júnior como relevante para explicar o seu impacto futuro. O uso de características indiretas (e.g, ligadas ao coautor principal) é completamente justificado por resultados de pesquisas recentes (LIU; TANG et al., 2018; LI; ASTE et al., 2019). Por exemplo, um estudo anterior, por Liu, Tang et al. (2018), concluiu que orientados guiados por orientadores com maior habilidade para pesquisa têm melhor desempenho acadêmico do que o resto, e ser orientado por um pode aumentar seu índice-h. Conclusão parecida foi dada por Li, Aste et al. (2019), que concluíram que pesquisadores júnior que coautoram trabalho com cientistas reconhecidos desfrutam de uma vantagem competitiva por todo o resto de suas carreiras, comparado com pares com perfis de início de carreira similares, mas sem coautores renomados.

A principal contribuição deste trabalho é uma fórmula para predizer o Q estável de pesquisadores júnior usando dados de seus cinco primeiros anos depois do início de suas carreiras de pesquisa. Adicionalmente, nós quantificamos a importância do coautor principal na carreira de cada pesquisador júnior.

4.2 Materiais e Método

Nós adotamos uma abordagem de aprendizado supervisionado padrão, Equação 5.2, para aprender a função $f_{\theta^*} : \mathbb{R}^5 \to \mathbb{R}$, para estimar o Q estável de um pesquisador júnior do cientista alvo, em termos das entradas descritas na Tabela 4.2. As entradas foram computadas no último ano da fase de pesquisador júnior. D é uma base de dados, em que cada linha contem as características de i como indicado na Tabela 4.1. Nós usamos o índice i para nos referirmos ao

pesquisador júnior e c_i para seu colaborador. Na subseção 4.2.1, nós propomos um algoritmo para buscar por c_i em uma base de dados de publicações.

$$\theta^* = \arg\min_{\theta} \sum_{(x,y)\in \mathrm{D}} (y - \mathrm{f}_\theta(x))^2 \tag{4.1}$$

Tabela 4.1: Conjunto de dados D criado. n é o número de cientistas.

		Ano do cálculo da medida				
Cientista		Ano da predição (ano 5)				Depois do Ano 15 da carreira
i	h_i	Q_i	Q_{ci}	N_{ci}	a_{ci}	Q_i
1	h_1	Q_1	Q_{c1}	N_{c1}	a_{c1}	Q_1
2	h_2	Q_2	Q_{c2}	N_{c2}	a_{c2}	Q_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$n-1$	h_{n-1}	Q_{n-1}	$Q_{c_{n-1}}$	$N_{c_{n-1}}$	$a_{c_{n-1}}$	Q_{n-1}
n	h_n	Q_n	Q_{c_n}	N_{c_n}	a_{c_n}	Q_n

Tabela 4.2: Variáveis preditoras.

Variável	Símbolo	Descrição
Índice h	h_i	O índice h de i
Valor Q	Q_i	O valor Q de i
Valor Q	Q_{c_i}	O valor Q do colaborador de i
Número de artigos	N_{c_i}	O n^o de artigos do colaborador de i
Idade acadêmica	a_{c_i}	A idade do colaborador de i

Visando escolher o melhor modelo que revele os relacionamentos no dado, nós testamos uma rede neural profunda (aqui chamada de modelo *Deep*) com d camadas escondidas $\mathrm{f}_{\mathrm{A}_1,\cdots,\mathrm{A}_d,\vec{b}}(\cdot)$ e um modelo de regressão linear (RL) com cinco variáveis independentes $\mathrm{f}_{\mathrm{c}_1,\cdots,\mathrm{c}_5}(\cdot)$, apresentadas anteriormente na Tabela 4.2. No caso do modelo *Deep*, os parâmetros treináveis θ são as matrizes $\mathrm{A}_1, \mathrm{A}_2, \cdots, \mathrm{A}_d$, e o vetor *bias* $\vec{b}$. E, no caso do modelo RL, os parâmetros treináveis são os coeficientes para as cinco características $\mathrm{c}_1, \mathrm{c}_2, \cdots, \mathrm{c}_5$.

Nós tomamos uma abordagem de avaliação *Holdout* (SAMMUT; WEBB, 2010). O propósito dessa abordagem é testar um modelo em dado diferente daquele em que ele foi aprendido (treinado). Portanto, o conjunto de dados D, na Tabela 4.1, foi particionado em um conjunto de treino (80%) e um conjunto de teste (20%).

Para avaliar a acurácia da predição de modelos, nós reportamos a raiz dos erro quadrático médio (RMSE em Inglês) sobre o conjunto de dados de teste, definida na equação 4.2:

$$\mathrm{RMSE} = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - \mathrm{f}_\theta(x_i))^2} \tag{4.2}$$

E, para avaliar a consistência e o viés de modelos, nós comparamos valores preditos $\hat{y}_i$ com valores observados y_i. Nós analisamos a linha de regressão dada pela equação 4.3.

$$y_i = \beta_0 + \beta_1 \hat{y}_i + \epsilon_i, \tag{4.3}$$

Nós comparamos a linha de regressão (equação 4.3) contra a linha 1:1 (linha diagonal). Nós testamos se valores preditos e observados são consistentes (ambos crescendo ao longo da reta de regressão 4.3), a nula desse teste estatístico é que são consistentes (i.e., inclinação da reta de regressão 4.3 é igual a 1 - β_1 = 1). Nós também testamos se os valores preditos são muito maiores ou muito menores do que os valores observados, i.e., se o modelo é enviesado. A nula desse teste estatístico é que o modelo não é enviesado (β_0 = 0). Caso ambas as nulas não sejam rejeitadas, então o desacordo entre as predições do modelo e o dado observado pode ser totalmente atribuído a variância não explicada.

4.2.1 Identificação do Colaborador Chave do Pesquisador Júnior

O colaborador chave do pesquisador júnior é aquele com mais publicações com maior impacto com quem o pesquisador júnior tem escrito um artigo. Para extrair o colaborador chave de cada pesquisador júnior, nós temos escrito o algoritmo 4. A parte central desse procedimento é descrita pelo algoritmo 5. O algoritmo varre todas as publicações de um cientista pertencentes a sua fase de pesquisador júnior, e encontra todos os seus coautores nesse período. Ele considera todos os coautores com um certa idade acadêmica como elegíveis, e garante retornar um deles, aquele com maior índice h. E, em caso de empates, usa como primeiro critério aquele com melhor desempenho de publicação (número de artigos). Notem que pode ser que o pesquisador júnior não tenha um coautor elegível (ou ter escrito artigos como o único autor nesse período), nesse caso o algoritmo retorna o próprio id (identificador) do pesquisador júnior. A linha do tempo na Figura 4.1 mostra a ideia por trás do algoritmo. Por último, notem que o colaborador chave é uma informação induzida a partir da base de dados de citação, e portanto, por alguma razão, pode ser que o pesquisador júnior não o consideraria como o seu maior influenciador no período. No entanto, em muitos casos pode se confirmar o relacionamento orientado-orientador.

Algoritmo 4 Identifica na base de dados de publicações aqueles pesquisadores com colaboradores elegíveis

Entrada: Um conjunto de dado de publicações A
Saída: Uma lista L de pesquisadores com colaboradores elegíveis

```
1   L ← ∅
2   para cada autor i pertencente ao conjunto de dado A faça
3       se Algoritmo 5 encontrar um coautor c_i para i então
4           Adicione i a lista L
5   retorne L
```

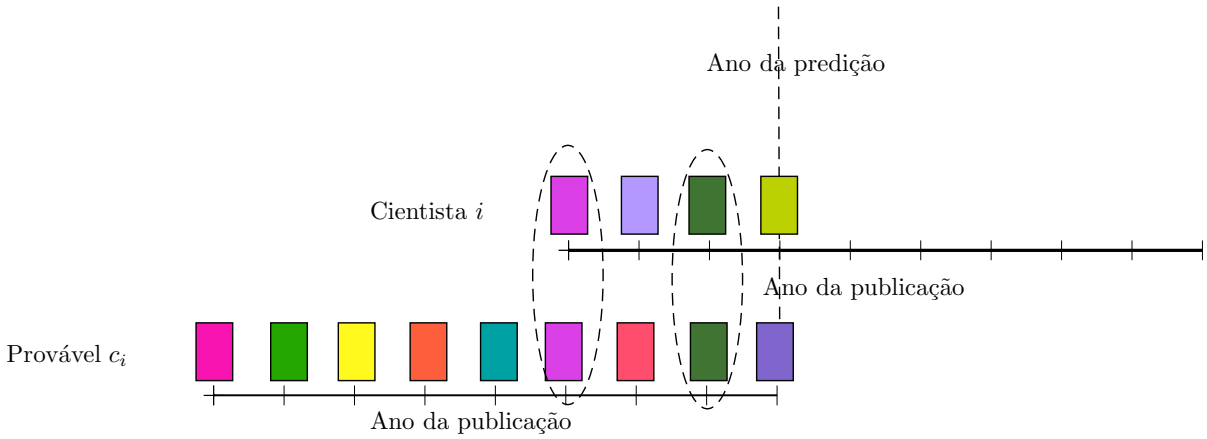


Figura 4.1: Linha do tempo mostrando publicações sobrepostas do histórico de publicação do pesquisador júnior i e o do coautor principal c_i. Pode existir mais de um coautor do pesquisador júnior elegível, mas o principal é aquele com um maior índice-h no ano da predição.

Algoritmo 5 Busca o provável colaborador principal do pesquisador júnior

Entrada: o identificador único de um autor i, período $T = Fase\_De\_Pesquisador\_Jr$ para buscar pelo coautor c_i e D, o número mínimo de anos na academia que seu coautor c_i deve ter desde que publicou pela primeira vez.
Saída: c_i o coautor de i encontrado ou i caso não encontre um coautor elegível.

```
1   c_i ← i
2   para cada publicação p de i faça
3       se p for uma publicação do período T então
4           para cada coautor e dessa publicação p faça
5               se e tiver pelo menos D anos desde que publicou pela primeira vez então
6                   se seu índice h (de e) for maior do que o índice h de c_i então
7                       c_i ← e
8                   senão se seu índice h (de e) for igual ao índice h de c_i então
9                           c_i ← aquele com mais artigos publicados.
10  retorne c_i
```

4.3 Resultados

4.3.1 Configuração dos Experimentos

Os dados usados para esse trabalho foram tirados do conjunto de dados de rede de citação da ACM (TANG et al., 2008) pela *AMiner*, um serviço online usado para realizar tarefas de mineração de dados, indexação e busca web. A base de dados utilizada indexa quase 2.5 milhões de artigos científicos publicados em periódicos e conferências da ciência da computação, escritos por aproximadamente 1 milhão e seiscentos mil cientistas, a maioria absoluta da computação. A rede de citação tem nela inseridos artigos publicados de 1936 até 2017, o ano da coleta desse dado. Finalmente, nós não avaliamos o problema de desambiguação de nomes uma vez que outros autores (TANG et al., 2008) disponibilizaram essa base de dados pré-processada com nomes(identificadores) únicos de autores. Uma análise detalhada desses dados é feita no Capítulo 3.

Configuração 1

Em um primeiro momento, nós analisamos as carreiras de 8.150 cientistas (i.e., cada linha da nossa base de dados de treino/teste representa um desses cientistas através de suas características.). Os critérios para selecionar essa amostra foram:

- ter um mínimo de 40 artigos, indexados na base de dados.
- Durante os 5 primeiros anos de trabalho, ter colaborado com um cientista com uma idade acadêmica de 10 no ano da predição, pelo menos.

Nós escolhemos essa faixa de pesquisadores para estudo porque ela inclui acadêmicos com carreiras fecundas (produtivas), mas com um impacto final muito diferente.

Uma abordagem longitudinal com relação ao conjunto de dados (consideramos cientistas de diferentes épocas) foi selecionada para extração de cientistas júnior com um coautor elegível. Nós escolhemos essa abordagem particular por causa dos critérios adotados.

Foi decidido que a melhor abordagem para esse estudo foi a abordagem *Holdout*: Nós selecionamos 20 % da amostra (1631) para compor nosso conjunto de teste, e o restante (6519) o conjunto de treino.

Configuração 2

Em um segundo experimento, nós flexibilizamos o primeiro critério de seleção de cientistas júnior, a configuração usada foi:

- ter entre 20 e 40 artigos (inclusive 20 e 40), indexados na base de dados.
- Durante os 5 primeiros anos de trabalho, ter colaborado com um cientista com uma idade acadêmica de 10 no ano da predição, pelo menos.

Foi encontrado 24.063 cientistas nessa faixa. Desse total 5.801 preenchem os requisitos (nossos critérios). Os resultados para essa configuração e a anterior são apresentados nas subseções seguintes.

As discussões seguintes consideram os primeiros cinco anos de trabalho do pesquisador júnior (idade acadêmica (ou tamanho de carreira) igual a 5) e a configuração inicial. A Tabela 4.3 mostra o resultado para segunda configuração 2 (em dado de treino). O modelo RL resultou em um $RMSE$ = 1.90 e um R^2 = 0.38 em dado não visto (i.e., em dado de teste). Por outro lado, na Tabela 4.3, nós apresentamos os resultados da regressão para os outros tamanhos de carreira, inclusive o tamanho 5.

Tabela 4.3: Resultados das regressões para os tamanhos de carreira 2,3,4, e 5. Configuração 1.

	Variável Dependente			
	Q_i			
	2	3	4	5
Q_{ci}	0.184*	0.194*	0.198*	0.130*
h_i	1.142*	0.889*	0.773*	0.515*
Q_i	0.0001	0.0003	0.0003	0.284*
Constante	2.881*	2.548*	2.245*	1.930*
Observações	5,854	6,753	7,512	8,150
R^2 ajustado	0.101	0.129	0.162	0.246
Note:				*p<0.01

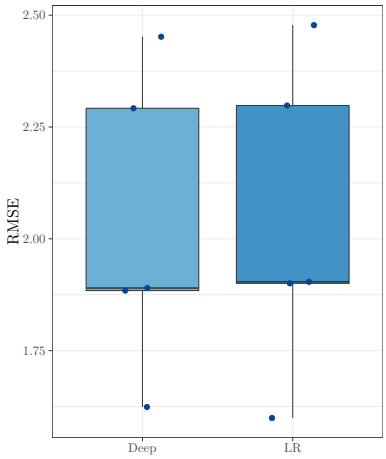
Tabela 4.4: Resultados da regressão (do modelo RL) para o tamanho 5 e a configuração 2.

	Variável Dependente:
	Q_i
Q_{c_i}	0.103* (0.081, 0.125)
h_i	0.558* (0.511, 0.606)
Q_i	0.387* (0.361, 0.413)
Constante	1.195* (1.086, 1.303)
Observações	4,516
R^2	0.350
R^2 Ajustado	0.349
Note:	*p<0.01

4.3.2 Acurácias para as abordagens *Deep* e de Regressão Linear (RL)

Os *boxplots* na Figura 4.2 mostram que as acurácias para a abordagem Deep e a abordagem RL são praticamente as mesmas. Como a acurácia da Deep não diferiu significativamente da acurácia da abordagem RL, nós inferimos que a Deep é uma função aproximada a linha de regressão RL. Também é possível dizer que, esse achado insinua a ausência de relacionamentos não lineares significantes no dado. Nós escolhemos o modelo RL ao invés do Deep porque ele é um modelo com excelente compreensibilidade, i.e, as razões para uma decisão do modelo podem ser facilmente compreendidas.

Nós também notamos uma melhora no valor do RMSE quando usando o modelo RL ao invés do modelo *baseline*, que considera os valores de Q atuais (menos precisos) dos pesquisadores júnior como seus valores definitivos (i.e, a serem observados na fase de pesquisadores sênior). Com um RMSE de 2.18, o modelo RL é aproximadamente 32% melhor do que o modelo *baseline*, com um RMSE de 3.22 (Tab. 4.2).



(a) Boxplots do resultado da validação cruzada 5-folds.

Modelo	RMSE	R^2 Ajustado
RL	2.18	0.28
Deep	2.22	0.35
Baseline	3.22	0.15

(b) Desempenho dos modelos. Baseline = ao Q correntes dos pesquisadores JR

Figura 4.2: Acurácias para os modelos testados.

4.3.3 Efeito da curta presença na academia do pesquisador júnior no seu Q

Os gráficos (*Scatter plots*) na Figura 4.3 mostram os resultados das regressões dos valores de Q de cientistas calculados em três momentos diferentes. O resultado confirma o efeito da dependência do tempo no cálculo do Q de cientistas. As estimativas dos cálculos para pesquisadores júnior são menos precisas.

4.3.4 Resultados da regressão de valores Observado x Predito

O resultado da regressão de valores observados versus preditos para o modelo RL na Tabela 4.5 mostra que o *slope* (i.e, o coeficiente de $\hat{y}_i$) não diferiu significativamente de um, sugerindo que os valores observados e preditos variam juntos consistentemente ao logo de suas

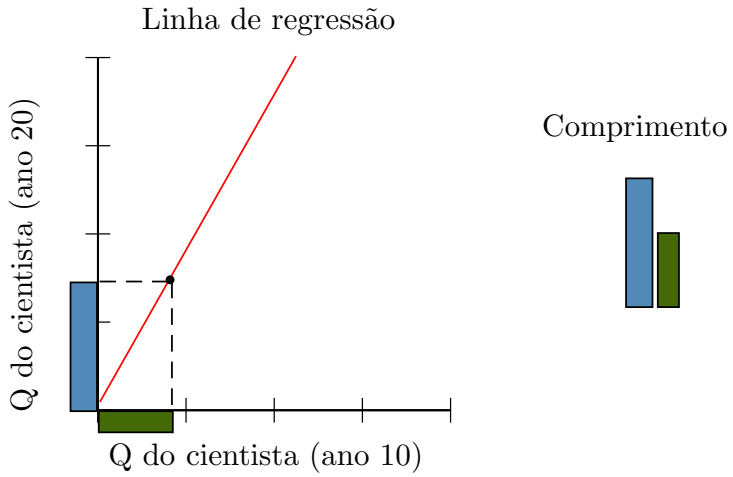


Figura 4.3: Resultados das regressões de valores de Q de um conjunto de cientistas medidos em dois momentos, no ano 10 e 20 da carreira (eixo y). Claramente, o Q desses pesquisadores medido no ano 10 é inferior ao Q deles medido no ano 20.

faixas. O valor da constante também não diferiu significativamente de zero, indicando que a média dos valores preditos (4.06) não apresenta um grau de enviesamento significativo com relação a média dos valores observados (4.13). Isso pode ser visto visualmente no gráfico da Figura 4.4.

Por outro lado, as predições do modelo *baseline* não tem consistência com os valores observados uma vez que o *slope* desviou-se significativamente de um (3.048). Isso visualmente pode ser visto na Figura 4.3, no gráfico da direita, Q_{20} versus Q_{10}.

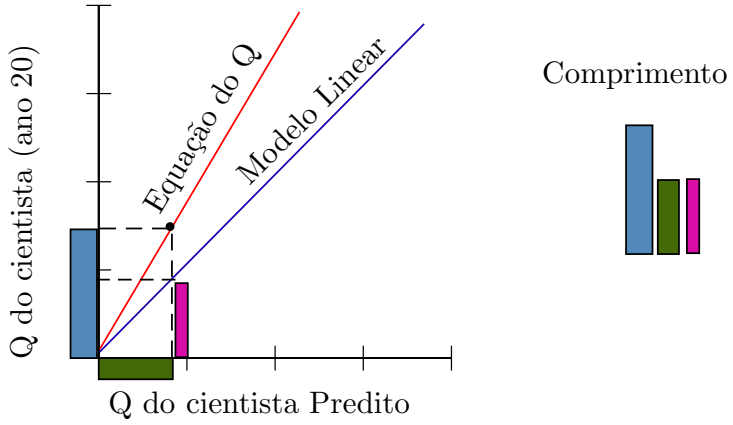


Figura 4.4: Plot no mesmo gráfico da regressão de valores preditos e observados do modelo linear (cor azul, Observado = 0.009 + 1.016Predito) e do modelo Q (cor vermelha) - O valor do Q corrente do cientista júnior como suposto ser seu Q definitivo.

Notem que a aproximação da linha de regressão da linha diagonal não indica boa acurácia, mas que o modelo não contribui significantemente para a perda de acurácia que tem sido observada.

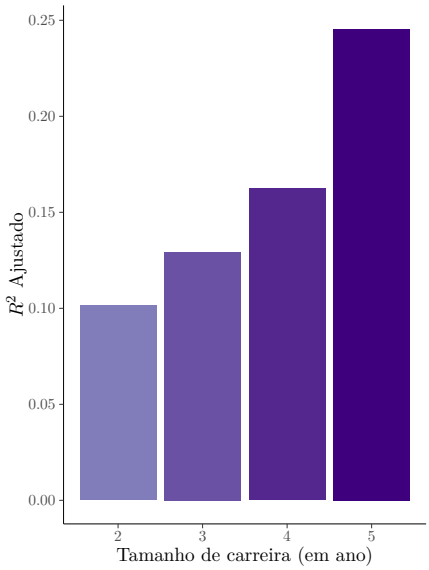
Como pode ser visto na Figura 6.3a, a quantidade de anos como um pesquisador júnior tem

Tabela 4.5: Resultado da regressão de valores Observados versus Preditos.

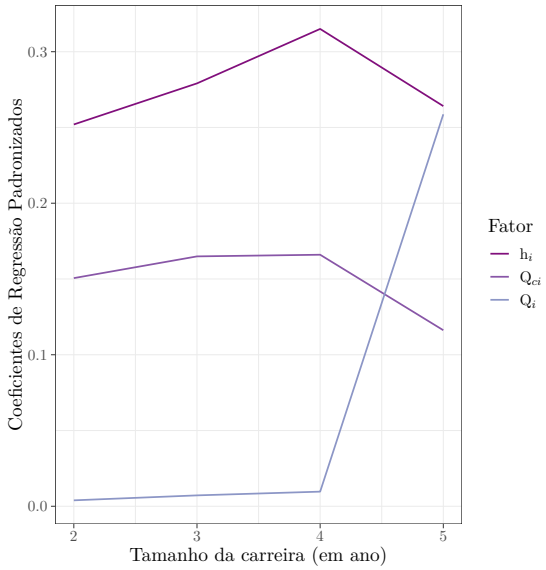
	Variável Dependente:	
	y_i (Obs. Q value)	
	LR model	Baseline
$\hat{y}_i$ (Pred. Q value)	1.016^{*} (0.938, 1.094)	0.486^{*} (0.430, 0.542)
Constante	0.009 (−0.327, 0.344)	3.048^{*} (2.878, 3.218)
Observações	1,631	1,631
R^2 Ajustado	0.283	0.149
Note:		*p<0.01

efeitos sobre o desempenho dos modelos. Para um tamanho de carreira de 2, o valor do R^2 ajustado é aproximadamente 59% pior do que para um tamanho de 5.

Nós observamos a partir da Figura 6.3b que as três medidas do pesquisador contribuindo mais para uma predição mudam em importância para predizer o seu Q com o passar do tempo. Enquanto que o Q do pesquisador júnior torna-se cada vez mais influente, o Q do seu coautor chave e o seu índice h perdem força.



(a) Efeito sobre a acurácia do modelo.



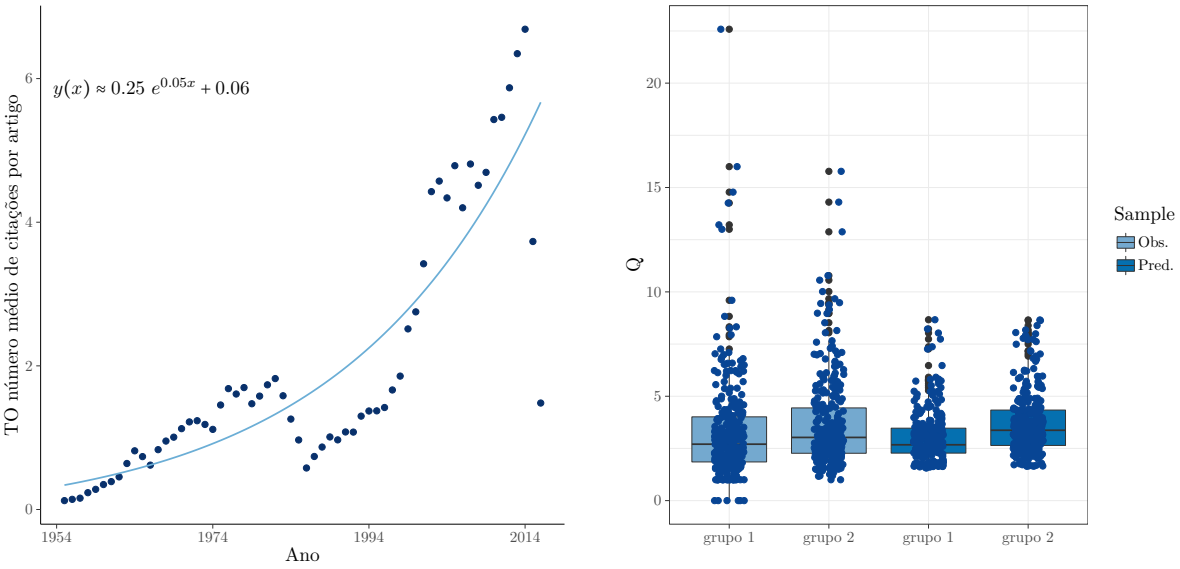
(b) Efeito sobre as explicativas do modelo.

Figura 4.5: O tamanho da carreira do pesquisador júnior tem efeito significante sobre o desempenho do modelo, e também sobre suas explicativas (o quanto cada fator contribui para uma predição) para uma predição.

Como mostrado na Figura 4.6a, a média de citação tem aumentado constantemente. O dado mostra que a produção de citações nessa base tem crescido a uma taxa anual de 5% e dobrado a cada 13 anos, aproximadamente.

Para avaliar o efeito desse viés temporal na acurácia do modelo, nós comparamos o desempenho do modelo em dois grupo separados, o grupo 1 contendo pesquisadores com publicações antes de 1999, e o grupo 2 nos anos seguintes. Devido a inflação de citação, o grupo 1 tem recebido menos citações em média do que o grupo 2.

Os *boxplots* comparativos na Figura 4.6b mostram que nenhuma diferença significativa na ordem de disposição das medianas observadas e preditas foi encontrada (i.e., as medianas (observadas e preditas) para o grupo 1 são menores do que para o grupo 2, como esperado), pelo contrário, as duas amostras apresentam o mesmo padrão. Isso sugere que nossos modelos são modelos robustos para lidar com esse viés. O erro de predição observado pode ser atribuído à outros fatores como as imperfeições na coleta e o desequilíbrio da amostra de treino.



(a) O crescimento da literatura de Ciência da Computação.

(b) Predito versus Observados

Figura 4.6: O padrão visual percebido nos *plots* de caixas de observados também é visto naqueles de preditos, mostrando que nossos modelos são modelos robustos para lidar com o aumento observado da média de citação ao longo do tempo.

4.4 Discussão e Conclusões

Avaliar a competência e o potencial de pesquisadores júnior de se tornarem bons pesquisadores é essencial em muitos contextos acadêmicos, e.g., em aplicações de seleção de quadros editoriais. Devido a suas presenças curtas na academia, pesquisadores júnior têm históricos de publicações curtos e em evolução. Por isso, as avaliações baseadas neles não são confiáveis por via de regra (i.e., as métricas calculadas nesse período não refletem seus valores reais mais tarde, mesmo o Q que é estável por toda a carreira de um cientista.). Nesse trabalho, nós propomos equações estimadas para predizer o Q estável (definitivo) de pesquisadores júnior em função de seus valores de Q instáveis, índices h e o valor Q estável de seu coautor principal (mais produtivo

ou prolífico) na época.

As predições de nossas equações (para cada tamanho de carreira) são melhores do que as predições do modelo tomado como uma base de comparação, que toma valores de Q correntes de pesquisadores júnior, e assume que eles corresponderão aos seus valores de Q definitivos. Nós acreditamos que as explicativas do modelo (equações) para uma predição sua auxiliará julgadores em suas decisões, e.g., nós encontramos que quando predizendo impacto futuro de cientistas com no máximo dois anos como pesquisador júnior, olhar para o desempenho (ou impacto) de seus coautores, sobretudo o seu coautor chave, é importante. Essa é uma parte da explicativa do modelo para alguns casos particulares de cientistas nesse conjunto como impacto alto predito.

Além disso, nós temos aconselhado um novo algoritmo para identificar (inferir) o coautor influente de um pesquisador júnior a partir de qualquer conjunto de dados de citação. Até onde sabemos, essa é a primeira tentativa de quantificar o impacto da inflação de citação sobre a confiabilidade de modelos de aprendizado de máquina, aprendidos a partir de um conjunto de dados de uma rede de citação abrangendo um período muito longo, apresentando esse viés temporal.

Atualmente, existe pouca evidência sistemática de quais fatores (e.g., a qualidade da educação ou da instituição corrente, o tamanho da comunidade de pesquisa, gênero, dinâmicas de subcampo, hábitos de publicação) determinam o Q de um pesquisador, sendo que esse estudo é o primeiro passo para melhorar nosso entendimento dessa questão. Nossos resultados apontam para uma correlação positiva entre os valores do Q dos colaboradores chaves e os valores do Q definitivos de pesquisadores júnior. Entretanto, mais trabalho focado nisso necessita ser realizado.

Também, mais pesquisa precisará ser feita para aumentar a confiabilidade do uso de sistemas baseados em aprendizado de máquina (e.g., poder confiar em predições de modelos), apesar de nossas equações proverem explicativas para predições. Nesse contexto, nós pensamos que, integrando mais características da pesquisa do pesquisador em uma predição pode melhorar essa questão. Pensando nisso, nós especulamos que um trabalho publicado recentemente (KAYAL et al., 2019) sobre mineração de informação de financiamento de pesquisa, e um outro (DI IORIO et al., 2018) sobre a qualidade de uma citação (classificação da intenção de uma citação) podem melhorar as explicativas de modelos para uma predição.

E, embora diferenças de gênero significantes não tenham sido encontradas na cobertura de citação nas bases de dados, *Google Scholar*(GS) e Web of Science (WoS) (ANDERSEN; NIELSEN, 2018), nós consideramos que avaliações críticas de comitês de avaliação são indispensáveis para prevenir algoritmos de perpetuarem quaisquer tipos de viés.

Por último, nós esperamos que nossa pesquisa será benéfica em resolver o desafio para avaliadores (e.g., membros de bancas de concursos de docentes) lerem cada artigo a partir de vários candidatos para fazerem julgamentos qualitativos e pessoais. Combinado com boas práticas de avaliação por pares, essa ferramenta pode funcionar como suporte e sem opacidade para aplicantes.

Capítulo 5

Arcabouço Computacional para Selecionar Pesquisadores Importantes ainda em suas Fases de Pesquisadores Júnior

5.1 Introdução

A disputa por recursos de pesquisa tem, nas últimas décadas, se intensificado, e ao tentar prever o sucesso científico futuro do pesquisador, comissões avaliadoras têm feito isso de qualquer modo. Métricas tradicionais, como fator de impacto de periódicos e total de citações, continuam a ser usadas, embora elas não tenham sido eficazes para esse propósito.

Neste contexto, por exemplo, o índice-h futuro do cientista (HIRSCH, 2005) seria de grande valia para essas comissões. O índice H é um índice que tenta representar a importância relativa de um cientista, comparável em sua área de pesquisa, considerado robusto, por avaliar de forma simultânea, os aspectos relativos à produção (quantidade de artigos produzidos) e ao impacto (número de citações).

Como um ponto de partida para observar o quadro amplo de estudos sobre a predição do sucesso futuro de cientistas, nós podemos olhar nos artigos que citaram o modelo do índice-h futuro de Acuna, Allesina e Kording (2012). Usando o *VOSviewer* v. 1.6.15, um software de cientometria desenvolvido pela Universidade de Leiden, nós podemos explorar as coocorrências de termos em seus títulos e resumos. Depois de remover automaticamente as terminologias não informativas, pode-se observar em um gráfico de sobreposição os termos mais frequentes em uma escala de cores indicando o ano em que cada termo foi mais referenciado. Um total de 149 termos foram encontrados no conjunto de dados da *Scopus* citando o artigo de Acuna, Allesina e Kording (2012), e considerando unicamente os termos que ocorreram seis ou mais vezes e

os sessenta por cento mais frequentes deles segundo a pontuação de relevância do software, chegou-se ao grafo na Figura 5.1 com 51 termos.

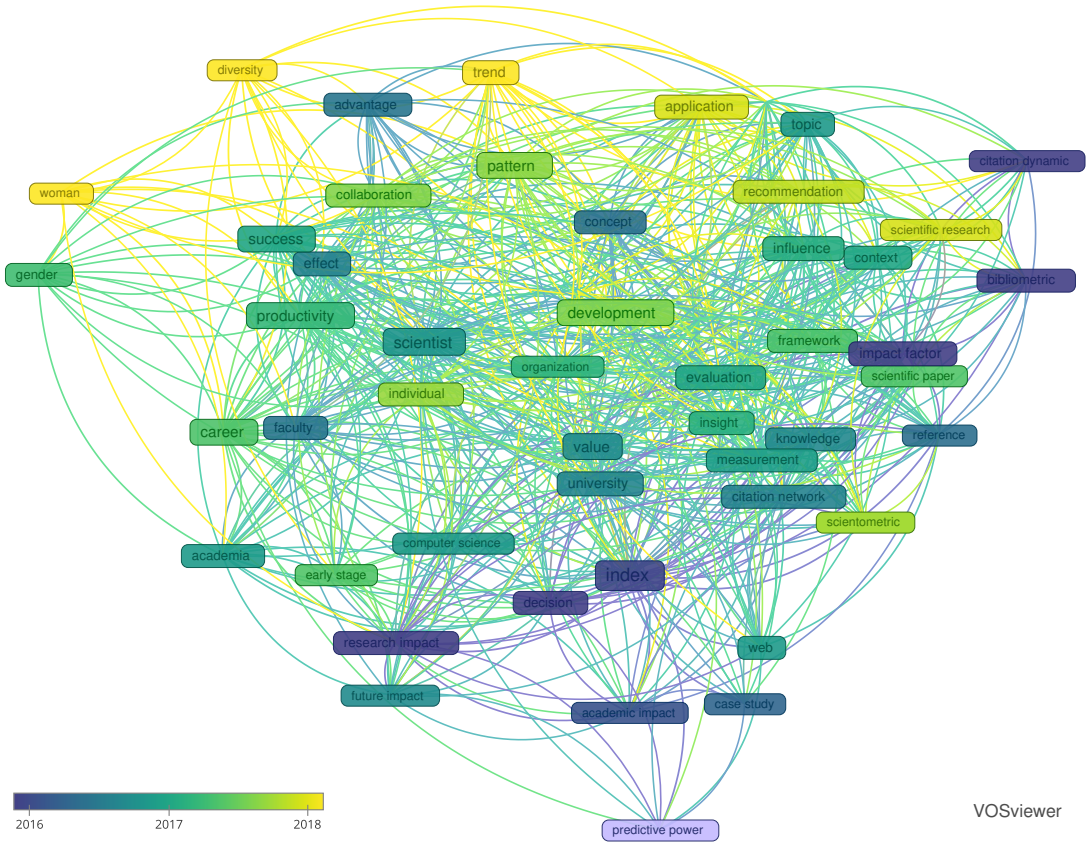


Figura 5.1: Gráfico de sobreposição de termos mais frequentes encontrados nos títulos e resumos de artigos citando o trabalho de Acuna, Allesina e Kording (2012) gerado pelo software *VOSviewer*.

Como pode ser observado, temáticas relacionadas ao gênero, à mulher e à diversidade são mais recentes do que aquelas focadas na ideia de impacto. Pelo grafo, pode-se notar que avaliações de desempenho acadêmico que levem em conta as diferenças históricas entre os avaliados têm ganhado importância. Devido à produção e avaliação de conhecimento serem tão relevantes para o papel de docente (HOLDEN; ROSENBERG; BARKER, 2005), cabe aos membros de bancas de seleção se esforçarem para fazerem tais avaliações de desempenho acadêmico.

Com a tendência de crescimento do número de candidatos por vaga disponível, avaliar cada candidato tem se tornado um desafio. Para aliviar isso, Frank (2019) sugere que bancas de seleção avaliem somente os principais achados dos candidatos (i.e., seus artigos representativos) e seus planos para os próximos anos, visando avaliar não só seus trabalhos correntes, mas também os prováveis impactos de suas futuras contribuições científicas. No entanto, o próprio autor admite que mesmo essa estratégia pode não ser escalável diante de um número muito grande de candidatos.

Por outro lado, uma abordagem em dois passos, proposta por Bornmann, Guns et al. (2021b), recomenda o uso de indicadores, focando em excelência científica, para selecionar em

uma primeira rodada, os prováveis melhores candidatos, e em uma segunda rodada, consultores *ad hoc* avaliam novamente o resultado da busca anterior. Esse trabalho somente tem testado a força preditiva de indicadores de impacto (desempenho) passado (e.g., índice-h do pesquisador).

Apesar desse interesse, o que torna um pesquisador júnior importante no futuro não é claro. Esse trabalho propõem uma nova abordagem de heurísticas *fast-in-frugal* para identificá-los antes de se tornarem um. Heurísticas *fast-in-frugal* são estratégias de decisão que usam parte da informação disponível (alguns poucos indicadores) e ignoram o resto. O diferencial desse trabalho é a experimentação de indicadores alternativos (e.g., o total de citações futuras que é esperado para um artigo representativo de um candidato) capturando o potencial para impacto futuro de jovens cientistas nesse contexto.

Nossa principal contribuição é um arcabouço para auxiliar comitês de seleção em suas tarefas diárias, em especial, a tarefa de avaliar o potencial para impacto futuro de cientistas individuais. Esse arcabouço pode ser adaptado para outras carreiras (mundo corporativo, esporte,etc).

5.2 Revisão de Literatura

Métricas de sucesso e desempenho científico na ciência já afetam a avaliação de pesquisadores e o financiamento de propostas, e consequentemente o futuro da ciência e as carreiras futuras de pesquisadores na ciência (SCHWEITZER, 2014).

Preocupada com as consequências negativas de indicadores mal construídos para o futuro da ciência, a comunidade de pesquisa deixou de promover medidas tais como o fator de impacto de periódicos e o índice h para criticá-los. Para a comunidade, eles podem reforçar vieses estruturais tais como racismo, sexismo e classicismo (SUGIMOTO, 2021). Em uma nota crítica sobre o fator de impacto, Garfield (1999) comparou a métrica à energia nuclear mostrando uma visão cautelosa sobre seu uso. Quando em boas mãos, métricas poderiam ser construtivamente usadas e, em mãos erradas, abusadas.

Tem sido demonstrado que grupos bem representados de certas disciplinas das ciências exatas continuam a dominar os concursos para cargos docentes em universidades (GIBBS KENNETH D et al., 2016; WRIGHT; VANDERFORD, 2017). Por isso, Peters (2017) tem aconselhado focar em vários critérios para sucesso, ao invés de olhar somente para o sucesso passado como o único indicador de sucesso futuro, sem considerar as barreiras sistêmicas existentes para certos grupos.

Um outro problema foi relatado por Fernandes et al. (2020) que concluíram que acima de um certo valor limitante, os indicadores tradicionalmente usados para medir o sucesso da pesquisa são inefetivos quando diferenciando entre candidatos contratados e não contratados.

Para Sugimoto (2021), ao guiar-se exclusivamente pela abordagem "publique ou pereça" para sucesso, diferentes atores da ciência (instituições de pesquisa, agências de fomento) têm perpetuado o "efeito mateus" * (MERTON, 1968) e alargado as diferenças já existentes em produtividade e reconhecimento entre acadêmicos.

*Efeito Mateus é um fenômeno social descrevendo as recompensas desproporcionais colhidas por aqueles em posições privilegiadas.

Vários estudos, por exemplo (BATISTA-JR; GOUVEIA; MENA-CHALCO, 2021), (ACUNA; ALLESINA; KORDING, 2012), (SARIGÖL et al., 2014), (WEIHS; ETZIONI, 2017), têm sido conduzidos sobre predição de sucesso futuro de autores. Uma revisão recente da literatura sobre o assunto (HOU et al., 2019b) encontrou que a distância entre acadêmicos citando o trabalho de outro acadêmico em redes de colaboração, o número de artigos em periódicos renomados, o índice h corrente, o número de diferentes eventos científicos em que um cientista publicou, o número de anos desde que um cientista publicou pela primeira vez, e a posição dentro de redes de colaboração foram os fatores mais testados até então para predizer o impacto futuro de autores.

O estudo de predição de impacto futuro de um cientista, medido através de seu índice h futuro, foi primeiramente realizado por Acuna, Allesina e Kording (2012). Desde sua publicação, o modelo de Acuna, Allesina e Kording (2012) para predizer o índice h futuro de um cientista tem sofrido duras críticas (GARCÍA-PÉREZ, 2013; PENNER et al., 2013). A principal delas é que o modelo é inclinado para pesquisadores sênior, sendo menos preciso para a faixa de pesquisadores júnior.

Trabalho anterior (PENNER et al., 2013; AYAZ; MASOOD; ISLAM, 2018) falhou ao tentar predizer o índice h futuro de um cientista para muito anos depois, usando características relacionadas ao início de carreira.

Por outro lado, Lee (2019) concluiu que o número de publicações em periódicos ou eventos científicos, durante a fase de pesquisador júnior, foi o fator que contribuiu mais para o desempenho de pesquisa nos anos seguintes (número de publicações nos 4 anos seguintes) e para o impacto de pesquisa nos anos seguintes (número de citações recebidas nos 4 anos seguintes).Também estudando os fatores relacionados ao início de carreira afetando os cientistas, Zhang e Yu (2020) concluiu que uma boa estratégia de publicação para pesquisadores em início de carreira é publicar alguns de seus artigos no mesmo periódico.

Um estudo recente, por Sinatra et al. (2016), que analisou as carreiras individuais de 3 mil físicos, encontrou que diferentes cientistas possuem diferentes talentos inerentes, chamado fator Q. Uma medida que quantifica a habilidade de um pesquisador de transformar suas ideias em descobertas com dado impacto para seu campo de pesquisa, medido pelo número de referências para elas feita por seus pares após elas terem sido comunicadas. Em um trabalho anterior (BATISTA-JR; GOUVEIA; MENA-CHALCO, 2021), nós descobrimos que o fator Q estável esperado para um pesquisador sênior é em grande medida previsível a partir de características relacionadas à sua fase de pesquisador júnior.

Apesar da importância dada às questões relacionadas aos desdobramentos futuros de uma carreira de pesquisa, ainda não sabemos quais fatores relacionados ao início de carreira mais contribuem para o sucesso futuro de um pesquisador, sucesso este medido através seus indicadores futuros. Nesse trabalho, nós propomos uma nova abordagem de heurísticas *fast-and-fruagal* para identificar pesquisadores prolífico no futuro quando eles ainda estão nas fases iniciais de suas carreiras. Baseado nas carreiras iniciais de 1631 pesquisadores amostrados a partir do conjunto de dados da ACM, e usados como entrada para nossos modelos de seleção, nós comparamos as classificações de pesquisadores preditas por nossos modelos de seleção contra a classificação

observada, isto é, a lista de pesquisadores ordenada pelo valor de Q estável do pesquisador.

5.3 Abordagem

Seja $S = \{1, 2, ..., n\}$ o conjunto de pesquisadores júnior recuperados da base de dados A e r_x o ranking induzido pelo indicador x, e suposto ser o ranking futuro deles. Um indicador é classificado em indicador corrente e futuro. Indicador futuros são estimados por modelos de aprendizado de máquina (e.g., Redes neurais profundas, Máquinas de vetores de suporte (SVM), Regressão linear) e indicadores correntes por indicadores tradicionais (e.g., índice-h, o total de citações, o Q). A Tabela 5.1 classifica eles quanto ao critério anterior. Um pesquisador é classificado como um pesquisador júnior pelos cinco primeiros anos desde a sua primeira publicação. Nesse trabalho, i representa um pesquisador júnior e c_i o seu influenciador principal (orientador). Como pode ser visto na Figura 5.2, os valores de indicadores futuros são estimados a partir da força preditiva de indicadores de pesquisadores júnior.

Com o objetivo de selecionar pesquisadores competentes a partir de um grande conjunto de candidatos, diferentes heurísticas são testadas e avalidas quanto à sua confiabilidade. As heurísticas avaliadas são do tipo uma única pista.

Uma heurística é composta de dois passos. Primeiramente, ordena-se os candidatos usando um indicador (e.g., um dos indicadores listados na Tabela 5.1). E em seguida, o ranking induzido anteriormente é avaliado novamente por um comitê de especialistas. Notem que o ranking induzido pelo indicador explora a força preditiva de indicadores.

Formalmente, consiste em classificar n candidatos $1, 2, i, \ldots, n$ por seus desempenhos (impactos), $P(1), P(2), P(i), \ldots, P(n)$, usando simplesmente um indicador de impacto (corrente ou futuro). Se a posição no ranking de $P(i) \leq valor$, em que $valor$ é o número total de aplicantes que pode ser selecionado, um comitê de seleção irá avaliar i (possivelmente mais detalhes da carreira serão considerados pelos avaliadores); senão, o candidato i é eliminado. Vale destacar que de qualquer forma (e.g., usando nossas heurísticas ou não) sempre existirão eliminados porque o número de vagas é muito inferior ao total de candidatos.

Por último, notem que o processo de eliminação envolve um único indicador (uma única pista) que visa um objetivo considerado central pelo comitê de seleção (e.g., o impacto futuro do pesquisador). O impacto futuro do cientista júnior é medido pelo seu Q estável (indicador futuro), computado quando eles já são pesquisadores sênior. A seguir discute-se a fundamentação por trás dos indicadores (futuros e correntes) testados.

O Q estável (futuro) do cientista júnior e o total aproximado de citações para seu artigo representativo, i.e., o mais citado nos cinco primeiros anos de carreira, são calculados a seguir. Notem que esta sendo usado apenas dados dos cinco primeiros anos de publicação dos candidatos como dado de entrada para inferir os modelos.

5.3.1 Cálculo dos Indicadores Futuros

- Cálculo do Q Futuro do Candidato

Tabela 5.1: Indicadores usados para induzir os rankings futuros.

Heurística	Indicador	Descrição	Futuro	Corrente
H1	Q_i	Valor Q	x	
H2	c_{r_i}	Quantidade de citações*	x	
H3	Q_i	Valor Q		x
H4	Q_{c_i}	Valor Q		x
H5	h_i	Índice h		x
H6	h_{c_i}	Índice h		x

*r_i é o artigo representativo do cientista júnior i (i.e., seu melhor artigo nos cinco primeiros anos da carreira).

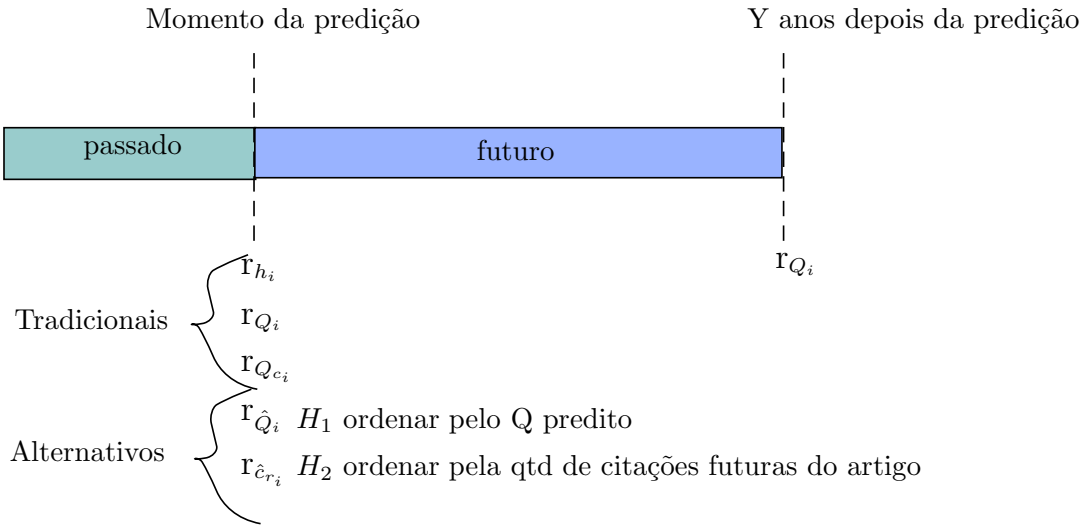


Figura 5.2: Indicadores tradicionais versus futuros.

O modelo (de regressão) expresso na equação 5.1 proposto em um trabalho anterior nosso (BATISTA-JR; GOUVEIA; MENA-CHALCO, 2021) foi usado para estimar esse parâmetro.

$$Q_i^{\text{Fut}} = 1.930 + 0.284Q_i^{\text{Curr}} + 0.13Q_{c_i}^{\text{Curr}} + 0.515h_i^{\text{Curr}} \tag{5.1}$$

em que Q_i^{Curr}, $Q_{c_i}^{\text{Curr}}$ e h_i^{Curr} são o Q corrente do aplicante júnior, o Q corrente de seu coautor chave (maior índice h), e o índice-h corrente do aplicante júnior, respectivamente.

O Q é um parâmetro único para o cientista i que quantifica a sua habilidade de melhorar um projeto de potencial p_α selecionado randomicamente, e consequentemente, publicar um artigo de impacto $\text{c}_{t,\alpha i} = \text{Q}_i\text{p}_\alpha$ depois de t anos desde a publicação do artigo α. Q_i^{Curr} é o Q do cientista i nos primeiros cinco anos de pesquisador. Ele foi calculado usando o modelo 5.2 dado em Sinatra et al. (2016)

$$Q_i^{\text{Curr}} = e^{(\frac{1}{\text{N}}\sum_{\alpha=1}^{\text{N}} \log_e \text{c}_{t,\alpha i}) - \mu_\text{P}}, \tag{5.2}$$

Nesta equação, $c_{t,\alpha i}$ é a quantidade de citações recebida pelo artigo α passados t anos de sua publicação (nesse trabalho, nos cinco primeiros anos de trabalho do pesquisador), e μ_p é uma variável dependente de campo científico e comum para todos os cientistas em um mesmo campo científico. Neste trabalho, μ_p é uma constante. Por último, $\log_e x$ é o logaritmo natural de x.

- Cálculo da quantidade de citações futuras do artigo do candidato

  O número de citações de um artigo científico ao longo da sua vida, representado por c_a^∞, é estimado através do modelo 5.3 dado em Wang, Song e Barabási (2013), definido em termo do 'fitness' λ_a do artigo. Esse parâmetro relaciona-se com a popularidade do artigo logo depois da sua publicação em um periódico científico (ou conferência científica), medido pelo tamanho do pico de citação no dois ou três anos seguintes.

$$c_a^\infty = m(e^{\lambda_a} - 1) \tag{5.3}$$

  Nesta equação, m representa o número médio de referências que cada novo artigo contém, e λ_a é a inclinação da linha de regressão dada na equação 5.4, i.e., $\lambda_a = \beta_1$.

$$c_a^t = \beta_0 + \beta_1 t + \epsilon_t, \tag{5.4}$$

  A variável c_a^t representa o número cumulativo de citações do artigo a e β_0, β_1 são o intercepte e o coeficiente de regressão para o número de anos t depois da publicação, respectivamente. ϵ_t é o erro.

Por último, o índice-h foi calculado como definido por Hirsch (HIRSCH, 2005), i.e., o número de artigos com o número de citações maior ou igual a h.

5.4 Avaliação

O objetivo dessa seção é apresentar as metodologias usadas na avaliação.

5.4.1 Configuração experimental

Para as discussões seguintes, considera-se 1.631 pesquisadores tirados do conjunto de dados da ACM (ver o Capítulo 3, para mais detalhes sobre o dado usado.). Eles constituem os candidatos, os quais desejamos ordená-los (classificá-los) por seus impactos futuros. Para todos eles, temos considerado somente a produção científica dos primeiros cinco anos de trabalho. Os critérios para selecionar essa amostra foram:

- ter um total de 40 artigos, indexados na base de dados, pelo menos.
- Durante os 5 primeiros anos de trabalho, ter colaborado com um cientista com uma idade acadêmica de 10 pelo menos.

Notem que está sendo usado o conjunto de teste (i.e., em dado não visto) para medir o desempenho do modelo na equação 5.1 (i.e., 20% do dado, o restante foi usado para treinamento de modelos) porque queremos comparar as acurácias dos rankings futuros produzidos por este modelo (heurística 1) contra o observado, e também contra os rankings preditos pelas demais heurísticas (heurística 2,3,4,5 e 6) na Tabela 5.1. Nós encontramos que o índice h médio deles é de 1.93 nos primeiros cinco anos de carreira. por último, nós configuramos o valor de m para 30, representando o número médio de referências que cada novo artigo é esperado ter.

Métricas de avaliação

Existem vários métodos para comparar dois rankings (FAGIN; KUMAR; SIVAKUMAR, 2003), entre eles o método de *Kendall* (KENDALL; GIBBONS, 1990) e de *Spearman* (DIACONIS; GRAHAM, 1977). Nesse trabalho, nós medimos o grau de concordância entre o ranking de candidatos predito por uma heurística e o ranking futuro deles observado usando o método de *Kendall*, e a acurácia do ranking predito, nós usamos precisão *top-k*. Esses dois métodos têm sido usados por Bar-Ilan, Levene e Lin (2007), Letchford, Moat e Preis (2015), Kanellos et al. (2019), em pesquisas anteriores.

- Métrica τ de *Kendall*

  A métrica τ de *Kendall* é computada baseada no número de pares de itens ordenados concordantes entre duas listas (WANG; SHAO et al., 2020). Um coeficiente de zero significa nenhum acordo entre os rankings, e o valor de 1 ou -1 significa perfeito acordo ou perfeito acordo inverso, respectivamente.

  Precisamente, seja r_i e s_i as classificações do i-ésimo candidato de acordo com a qualidade x e y, respectivamente. O coeficiente de correlação de *Kendall* é definido como:

$$\tau = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ij}}{\sqrt{(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2)(\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}^2)}} = \frac{n^o \text{ de pares concordantes - } n^o \text{ de pares discordantes}}{\binom{n}{2}}$$

  em que, $a_{ij} = \text{sgn}(r_j - r_i)$, $b_{ij} = \text{sgn}(s_j - s_i)$ e

$$\text{sgn}(\theta) = \begin{cases} 1 & \text{if } \theta > 0 \\ 0 & \text{if } \theta = 0 \\ -1 & \text{if } \theta < 0 \end{cases}$$

  Qualquer par de observações (r_i, s_i) e (r_j, s_j), em que $i \neq j$, é concordante se as classificações de ambos os candidatos, i e j, concordarem uma com a outra, i.e. se $r_i > r_j$ e $s_i > s_j$ ou $r_i < r_j$ e $s_i < s_j$. Elas são discordantes se $r_i > r_j$ e $s_i < s_j$ ou $r_i < r_j$ e $s_i > s_j$. Se $r_i = r_j$ ou $s_i = s_j$, o par não é nem concordante, nem discordante.

- Métrica *top-k*

Por outro lado, a precisão *top-k* fornece a porcentagem de itens compartilhados entre os itens *top-k* classificados em ambos os rankings. Nós selecionamos essa métrica porque é preciso avaliar a precisão do ranking predito com relação ao observado após a redução do total de candidatos para um número gerenciável.

5.4.2 Resultados

Para mostrar o grau de concordância entre os *rankings* de fato e predito, nós testamos a hipótese nula que τ é zero (nenhuma concordância) contra a alternativa que τ é diferente de zero. Um p-valor menor do que 0.01 significa estatisticamente significante e indica forte evidência contra a hipótese nula.

A partir da Tabela 5.2, nós podemos notar um correlação monotônica positiva, moderada ou fraca entre o ranking predito por uma heurística e o observado. Por outro lado, como mostrado na Figura 5.3, nenhuma concordância ($\tau = 0$) existe entre um *ranking* randômico e o observado.

Tabela 5.2: Coeficientes de correlação τ de Kendall entre a classificação observada e as que foram preditas pelos modelos de seleção.

Modelo usando a heurística	Kendall's τ	Correlação
H1	0.38	Fraca
H2	0.47	Moderada
H3	0.36	Fraca
H4	0.25	Fraca
H5	0.34	Fraca
H6	0.21	Fraca

.00-.19 muito fraco, .20-.39 fraco, .40-.59 moderado, .60-.79 forte, .80-1.0 muito forte.

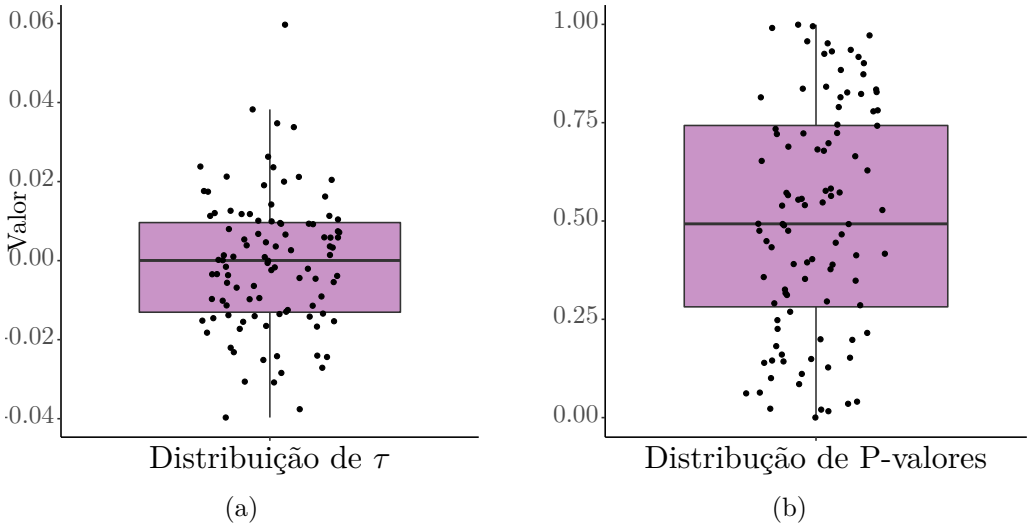


Figura 5.3: Distribuição de coeficientes Kendall' τ e p-valores para 100 comparações de rankings- o randômico e o observado.

Nós observamos a partir da Figura 5.4 que a escolha randômica de candidatos resultou em uma precisão Top-30% menor do que 40%. Como esperado, sua precisão é inferior à qualquer

uma de nossas heurísticas.

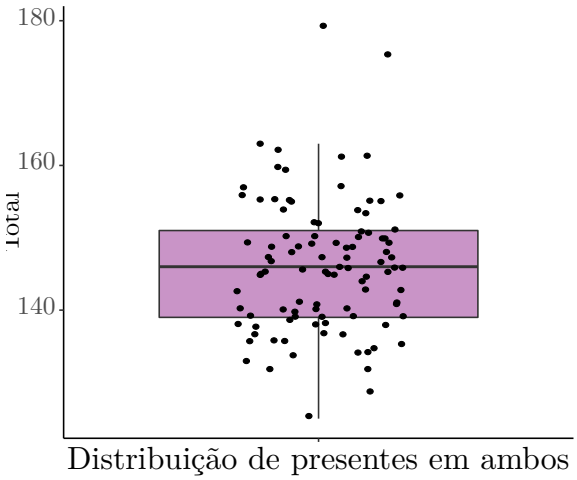
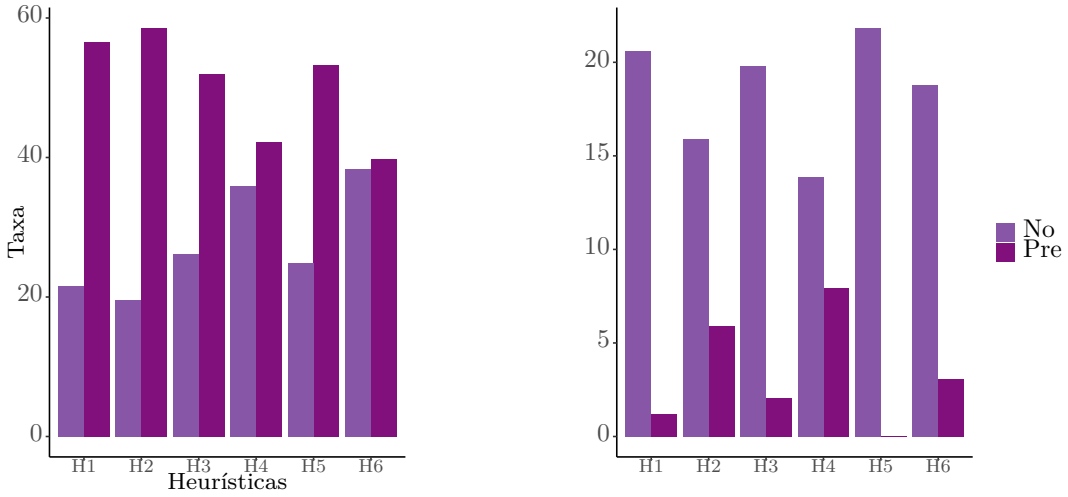


Figura 5.4: Distribuição de totais de pesquisadores entre os top-30% em cada ranking - o randômico e o observado.

Nós notamos a partir da Figura 5.5 que a precisão Top-30% das heurísticas no grupo B é relativamente muito menor do que no grupo A. O grupo A inclui pesquisadores com índice-h > *média* = 1.93, enquanto que O grupo B índice-h ≤ *média*. A disparidade sugere que todos os indicadores testados nas heurísticas discriminam contra pesquisadores no grupo B em algum grau.



(a) Grupo A

(b) Grupo B

Figura 5.5: Precisão Top-30% por grupo. Na cor lilás escuro é a precisão Top-30% (Pre) e na cor lilás claro é o erro (No). O erro significa que o candidato apareceu na lista observada, mas não foi predito.

5.4.3 Discussão

Por um lado, o número de candidatos por vaga aberta em concursos docentes tem aumentado muito. Por outro lado, como identificar um docente habilidoso ainda nas fases iniciais da carreira não é claro. Tradicionalmente, tomadores de decisão têm olhado para o sucesso passado como um principal indicador de sucesso futuro, sem considerar as barreiras sistêmicas, como o acesso desigual a recursos de pesquisa. Nesse contexto, um arcabouço computacional para identificar executores excepcionais bem cedo poderá auxiliar comissões avaliadoras (de promoção docente, concurso docente) em suas tarefas.

Nesse trabalho, nós propomos uma nova abordagem de heurísticas *fast-and-frugal* para identificar pesquisadores proeminentes antes de se tornarem um. Nossa abordagem de heurísticas baseada em indicadores alternativos (e.g., Q futuro, quantidade de citação futura) supera a abordagem tradicional baseada em indicadores de desempenho passado. Nossos achados são consistentes com resultados anteriores (ACUNA; ALLESINA; KORDING, 2012; LAURANCE et al., 2013; SARIGÖL et al., 2014; VAN DIJK; MANOR; CAREY, 2014; WEIHS; ETZIONI, 2017; BATISTA-JR; GOUVEIA; MENA-CHALCO, 2021).

Embora seja perigoso fazer quaisquer interpretações causais desses resultados, nós podemos especular que o que pode estar por trás do melhor desempenho de modelos de seleção baseados em predições está relacionado com a habilidade de máquinas de melhorar seu desempenho a partir da experiência. Mais pesquisa focada e aprofundada seria necessária para confirmar isso.

Também, nós notamos que pesquisadores júnior com indicadores abaixo da média são bem menos prováveis do que outros de serem notados por quaisquer modelos de seleção como pesquisadores proeminentes no futuro. Essa descoberta está em perfeito acordo com estudos anteriores (GIBBS KENNETH D et al., 2016; KÖCHLING; WEHNER, 2020; WRIGHT; VANDERFORD, 2017), que também tem encontrado vieses não intencionais em seleção de candidatos introduzidos por algoritmos e pessoas.

Por outro lado, algumas limitações podem provavelmente ter influenciado nossos resultados. A primeira delas é a abordagem horizontal tomada sobre o conjunto de dados com relação a dimensão tempo. Pesquisadores, trabalhando em diferentes épocas, foram igualmente tratados quando considerando os métodos de seleção. Essa escolha ignora o fato que o comportamento de publicação tem mudado ao longo do tempo, hoje um pesquisador publica mais artigos do que a trinta anos atrás. Isso poderia afetar a acurácia dos modelos. Mais pesquisa focada e aprofundada seria necessária para confirmar isso.

Também, pode ser, que devido a nossa decisão de somente ter incluído em nossa amostra pesquisadores com, pelo menos, um mínimo de 40 artigos publicados, isso tenha causado a exclusão de pesquisadores que iniciaram fortemente, mas não continuaram publicando no mesmo ritmo.

Por último, pode ser o caso, que a restrição de somente incluir pesquisadores que tenha um coautor com pelo menos dez anos de experiência, tenha implicitamente não considerado os coautores que principalmente publicam com seus pares.

Resumidamente, se ou não um pesquisador se tornará um pesquisador proeminente, pode

ser predito, em grande medida, pelo seu histórico de publicação, ainda que consideremos uma carreira curta de um pesquisador júnior.

5.5 Conclusão

Avaliar candidatos para promoção, financiamento de pesquisa demanda da predição precisa do desempenho futuro latente deles. Através de heurísticas baseadas em promessas futuras, e.g., através de estimativas do impacto futuro da publicação chave do cientista, avaliadores podem reduzir os esforços e o tempo para tomada de decisão. Neste trabalho, testou-se o desempenho dessas heurísticas contra heurísticas baseadas em indicadores tradicionais (e.g., o índice h).

Nossa evidência empírica tem nos levado a considerar que, em avaliação de candidatos para cargos de docentes, olhar para o desempenho (impacto) futuro dos candidatos é melhor do que olhar unicamente para seus desempenhos passados. Enquanto o último pressiona os pesquisadores a continuamente produzirem resultados publicáveis, o primeiro diminui isso porque o foco desloca para a qualidade do trabalho.

Nós temos notado que predizer a promessa futura de bons pesquisadores é difícil. No entanto, predizer o progresso de pesquisadores abaixo da média é ainda mais desafiador.

Trabalhos futuros devem se concentrar em melhorar a qualidade das predições das heurísticas e testar novas, e.g., novos modelos preditivos. Por último, nós esperamos que nossa pesquisa sirva como uma base para estudos futuros focados em diversidade, inclusão, e justiça social, no contexto de avaliação das promessas futuras de candidatos para financiamento, composição de quadros de universidades e promoção docente.

Capítulo 6

Q para Períodicos

6.1 Introdução

Publicar descobertas inovadoras em periódicos revisados por pares não só é importante para avançar a ciência, mas também a carreira de um autor (BREMBS, 2018). Publicar naqueles entre os melhores classificados em rankings de periódicos pode render a um pesquisador um vantagem, e.g, maior visibilidade para suas pesquisas, convites para participar de grandes projetos científicos, novos colegas ou até uma promoção na carreira. A posição do periódico nesses rankings é importante, porque pesquisadores utilizam essa informação quando decidindo onde publicar seus trabalhos ou quando buscando por trabalhos relacionados a sua pesquisa. Também, rankings de periódicos são relevantes para governos, porque servem como uma referência para avaliar o desempenho de pesquisa internamente em seus países.

Os rankings de periódicos são baseados em medidas de impacto. Tais medidas medem a influência (prestígio, visibilidade) de um periódico, levando em conta as citações entre artigos dentro do periódico. O *SCImago Journal Rank* (SJR) (GONZÁLEZ-PEREIRA; GUERRERO-BOTE; MOYA-ANEGÓN, 2010) é uma dessas medidas de impacto usada para produzir a classificação de periódicos indexados pelos conjuntos de dados da *Web of science* (WoS) e Scielo. Uma outra é o fator de impacto do periódico.

Complementarmente às medidas de impacto de um periódico, Braun, Glänzel e Schubert (2006) introduziram o índice h para periódico, um índice cumulativo (EGGHE, 2007; JIN et al., 2007; PENNER et al., 2013) inicialmente proposto por Hisch, para medir a importância relativa de um pesquisador, e agora, também a importância relativa de um periódico.

Diversos fatores podem afetar a confiabilidade de um ranking. Entre eles estão o uso de medidas cumulativas para indução do ranking e a amplitude da cobertura do conjunto de dados. Por exemplo, um estudo anterior, por Bar-Ilan, Levene e Lin (2007), concluiu que existe um considerável desacordo entre os rankings gerados a partir do Google Scholar e da ISI Web of Science.

Recentemente, Sinatra et al. (2016) desenvolveram um novo modelo, denominado modelo Q, para definir e predizer carreiras individuais, e concluíram que uma carreira científica é um resultado de sorte (i.e., envolve escolher randomicamente diversos projetos de pesquisa, cada

um com um potencial impacto, a partir de uma distribuição de probabilidade comum a todos os cientistas) e talento nato capturado por uma medida não cumulativa - o Q. Os autores confirmaram a natureza estacionária (não cumulativa) do Q para um contexto específico.

O Q, assim como o índice-h, possui cálculo simples baseado no impacto das publicações (medido pela quantidade de citações recebidas por uma publicação) e pode ser aplicado a qualquer nível de agregação, como sugere o dado na Figura 6.1. Neste trabalho, nós introduzimos o Q para periódicos. Nós avaliamos a similaridade entre os rankings induzidos pelo Q e pelo SJR. A vantagem do ranking produzido pelo Q (uma medida não cumulativa) é demonstrada entre os principais jornais de revisão de física, entre eles *Physical Review Letters*, *Physical Review*, e *Reviews of Modern Physics*. Adicionalmente, nós apresentamos evidência contra a natureza não cumulativa do Q, i.e, nós notamos que as métricas média e variância têm aumentado constantemente ao longo de algumas décadas.

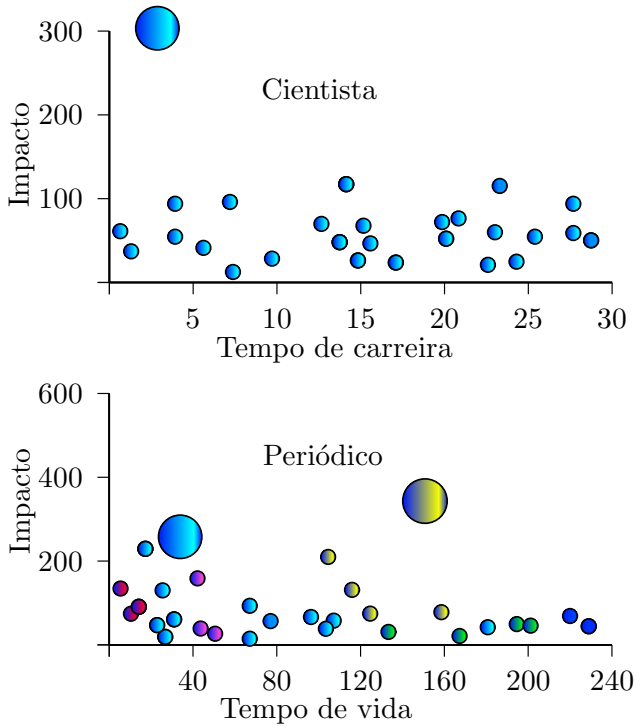


Figura 6.1: O procedimento para o cálculo do Q para um periódico é como para um cientista. Basta ver o periódico como um cientista e o tempo de vida do periódico como o tempo de carreira do cientista.

Neste trabalho, buca-se respostas para as seguintes questões de pesquisa:

Q1. O Q de um periódico é uma medida não cumulativa como tem sido declarado para o Q de um pesquisador?

Q2. Existe uma correlação positiva entre o Q do periódico e o Q do pesquisador?

Q3. Existe uma correlação positiva entre os rankings induzidos pelo Q e pelo SJR?

6.2 Método

Calcula o Q do periódico

O Q do periódico é calculado sobre uma amostra das publicações (ou a lista inteira) publicada nele. A fórmula usada é essencialmente a mesma como aquela proposta originalmente por Sinatra et al. (2016). O algoritmo 6 computa o Q de um periódico. Por comparar exclusivamente periódicos específicos de um campo (da física), notem que não consideramos a constante μ da fórmula original, dependente de cada campo, porém o algoritmo pode ser facilmente adaptado. O uso desse parâmetro permite comparar periódicos a partir de diferentes campos científicos.

Algoritmo 6 Cálcula o Q do periódico.

Entrada: a lista L = $\{1, 2, \cdots, n\}$ contendo identificadores únicos da amostra de artigos do periódico (ou da lista completa), e um número natural $t > 0$, representando o número de anos após a publicação para um artigo acumular citações.
Saída: um número real representando Q do periódico

```
1   m ← 0
2   soma ← 0
3   media ← 0
4   para cada artigo α na lista L faça
5       qtdC ← o número de citações que o artigo α recebeu, passados t anos da sua publicação
6       se qtdC > 1 então
7           soma ← soma + log_e qtdC                    // e é o número natural.
8           m ← m + 1
9   se m > 0 então
10      media ← soma/m
11      retorne exp(media)                              // exp(x) = e^x
12  senão
13      retorne m
```

Conjunto de dados

Nós consideramos uma seleção arbitrária de 6 periódicos revisados por pares mantidos pela American Phisical Society. Os periódicos selecionados têm pelo menos 4 décadas, portanto um histórico longo de publicações, e todos eles estão em plena atividade. Nós denotamos por $\mathcal{J} = \{PRL, \cdots, PRD\}$ o conjunto de periódicos indexados pelo conjunto de dados, mas somente selecionamos seis. A Tabela 6.2 mostra o ano de criação e a quantidade de publicações contabilizadas por esses periódicos.

A partir dos metadados dos artigos publicados em $\mathcal{J}$, nós identificamos unicamente os autores dos artigos. Nós usamos o algoritmo para resolução de nomes de autores proposto por Martin et al. (2013). O procedimento considera dois autores os mesmos se possuem nomes completos correspondentes idênticos e compartilham uma afiliação, ou têm coautores em comum, ou citaram um ao outro.

Tabela 6.2: Periódicos indexados pelo conjunto de dados APS.

Sigla	Descrição	Ano de criação	# Artigos
PRA	Physical Review A	1970	78664
PRB	Physical Review B	1970	187359
PRC	Physical Review C	1970	39889
PRD	Physical Review D	1970	87183
PRE		1993	57784
PR		1913	47940
PRAB		2016	637
PRAPPLIED		2014	1655
PRFLUIDS		2016	1236
PRMATERIALS		2017	1054
PRPER		2016	264
PRSTAB		1998	2384
PRSTPER		2005	368
PRI		1893	1469
PRX		2011	1334
RMP	Reviews of Modern Physics	1929	3348
PRL	Physical Review Letters	1958	123447

Para cada periódico $J \in \mathcal{J}$, nós identificamos todos os artigos publicado nele. Para cada um computamos o seu Q usando o Algoritmo 6. Nós usamos uma abordagem longitudinal do dado.

Dentro de cada periódico $J \in \mathcal{J}$, nós identificamos todos os autores que publicaram nele, e para cada autor $i = 1, 2, \cdots, N_J$, nós identificamos as suas publicações até 2018 e, para cada publicação, nós computamos o seu impacto quantificado por citações. Notem que um autor i pode ter publicado em mais de um desses periódicos. Nós somente consideramos citações de itens indexados pelo conjunto de dados da APS.

Avaliação da natureza estacionária do Q

Para avaliar a natureza não cumulativa do parâmetro Q, nós selecionamos o método de correlogramas para modelos de auto-regressão não estacionários introduzido por Nielsen (2006). O método compara correlogramas parciais obtidos a partir de duas abordagens para os seus cálculos: uma delas baseada em autocorrelações r_u, e a outra baseada em autocovariâncias g_u. O autores concluíram que para séries não cumulativas, os correlogramas induzidos a partir dos dois métodos anteriores praticamente se sobrepõem, enquanto que para séries temporais exibindo características não estacionárias os correlogramas divergem consideravelmente.

Para uma dada série $X_1, \cdots, X_T$, o cálculo de cada ponto do correlograma (u, r_u) ou (u, g_u), para $0 \leq u < T$, é dado por:

$$\text{Abordagem 1} \quad r_u = \frac{\sum_{t=u+1}^{T}(X_t - \bar{X}_{u+1}^{T})(X_{t-u} - \bar{X}_{1}^{T-u})}{\sqrt{(\sum_{t=u+1}^{T}(X_t - \bar{X}_{u+1}^{T})^2 \sum_{t=u+1}^{T}(X_{t-u} - \bar{X}_{1}^{T-u})^2)}}$$

$$\text{Abordagem 2} \quad g_u = \frac{\sum_{t=u+1}^{T}(X_t - \bar{X}_1^T)(X_{t-u} - \bar{X}_1^T)}{\sum_{t=1}^{T}(X_t - \bar{X}_1^T)^2}$$

em que $\bar{X}_v^{T-w}$ é a média da amostra $X_v, \cdots, X_{T-w}$.

Nós também aplicamos o teste estatístico de *Kwiatkowski-Phillips-Schmidt-Shin* (KPSS) (KWIATKOWSKI et al., 1992). A nula do teste estatístico corresponde à estacionariedade de uma série, a alternativa à sua não estacionariedade. Os autores produziram tabelas de valores críticos por meio de simulações. A Tabela 6.3 contém essas referências. Se o valor da estatística de teste for maior do que o valor crítico, então a hipótese nula é rejeitada, a serie é não estacionária.

Tabela 6.3: Tabelas de valores críticos com constante.

Níveis críticos	10%	5%	1%
Valores críticos	0.347	0.463	0.739

Medindo a correlação de dois rankings de anos diferentes

Para verificar se o ranking induzido pelo indicador x tem sofrido mudanças de um ano para outro, as correlações entre os rankings dos últimos quinze anos foram computadas. Medidas de correlação, por exemplo Kendall (KENDALL; GIBBONS, 1990) e footrule de Spearman, para computar o grau de acordo entre rankings têm sido recentemente estudadas em diferentes contextos por Fagin, Kumar e Sivakumar (2003), Bar-Ilan, Levene e Lin (2007) e Kanellos et al. (2019).

Kendall foi a métrica de correlação selecionada por nós. A métrica é computada baseada no número de pares de itens ordenados concordantes entre dois rankings. Um coeficiente de zero significa "nenhuma correlação“ entre eles e um valor de +1 ou −1 significa "perfeito acordo“ ou “ perfeita inversão dos rankings“, respectivamente.

6.3 Resultados

Na Figura 6.2 é mostrado o Q de cada periódico em função do período de tempo t fixado para cada artigo acumular citações, e.g., para cada artigo, somente as referências (citações) para ele ocorridas dentro do período t são levadas em conta. O resultado desse experimento mostra que para um $t = 20$ o Q do periódico é bem próximo a sua média.

O dado na Figura 6.3 sugere que a o Q dos periódicos está crescendo constantemente. No entanto, apesar da métrica ter uma variância crescente, ela é relativamente pequena.

Por outro lado, a classificação de periódicos pelo indicador SJR ao longo dos últimos 16 anos (2005-2020) apresentou classificações variadas (houve mudança nas posições do ranking) enquanto pelo Q não houve mudanças, como mostrado no dado da Figura 6.4. Por tanto, a classificação induzida pelo Q é menos afetada pelo fator tempo do que a classificação induzida pelo indicador SJR.

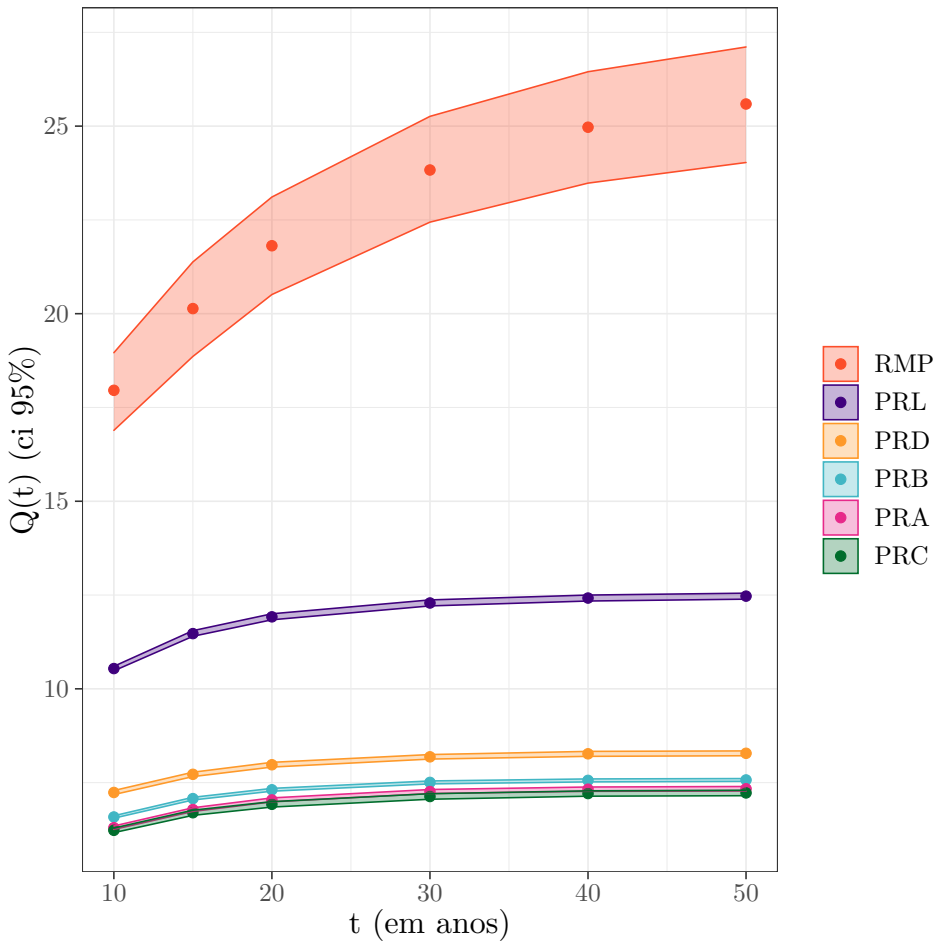


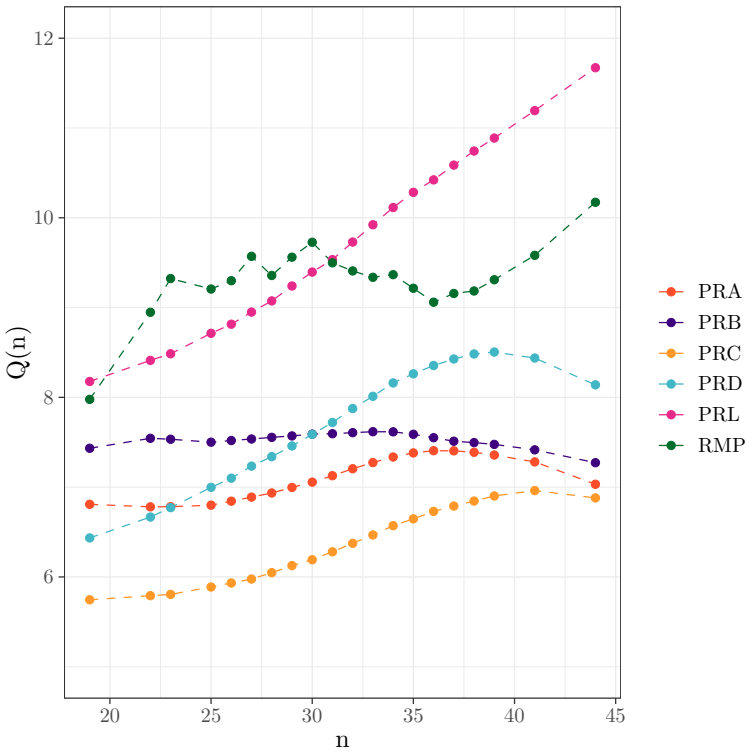
Figura 6.2: Classificação dos seis periódicos induzido pelo Q. Nós consideramos todos os artigos publicados em cada periódico até o final de 2018. Q(t) é o Q do periódico, e t o tempo para cada artigo acumular citações. A quantidade de citações recebida no período t mede o impacto do artigo.

Os boxplots de valores de Q de autores por periódico, na Figura 6.5, revelam que existe uma correlação positiva fraca entre periódicos de Q alto e autores de Q alto.

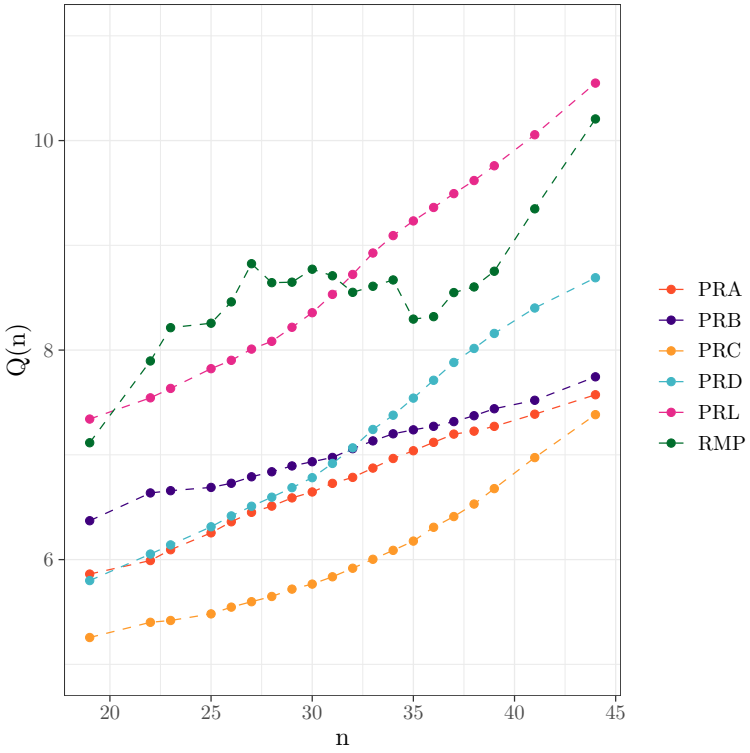
A Figura 6.6 mostra uma clara tendência de afastamento dos correlogramas gerados a partir dos dois métodos de criação. A não sobreposição dos correlogramas é característico de processos não estacionários (i.e., média e variância mudando ao longo do tempo).

Por outro lado, quando misturados os valores de Q, os correlogramas praticamente se sobrepõem, como pode ser visto na Figura 6.7, algo que é esperado para séries randômicas (média e variância constantes).

A não estacionariedade do Q de um periódico também é sugerida pelo teste estatístico de *Kwiatkowski-Phillips-Schmidt-Shin*. Como pode ser visto na Tabela 6.4, o valor da estatística de teste para as séries contendo valores de Q ano a ano é maior do que o valor crítico (0.739) definido pelos autores (Tabela 6.3), indicando que devemos rejeitar a hipótese nula de série estacionária. Diferentemente do resultado anterior, ao misturar os valores de Q de cada série, o valor da estatística de teste para cada série indica não rejeitar a hipótese nula.



(a) O Q do periódico, n anos depois da sua criação, considerando todas as citações recebidas pelos artigos publicados dentro do período na base de dados.

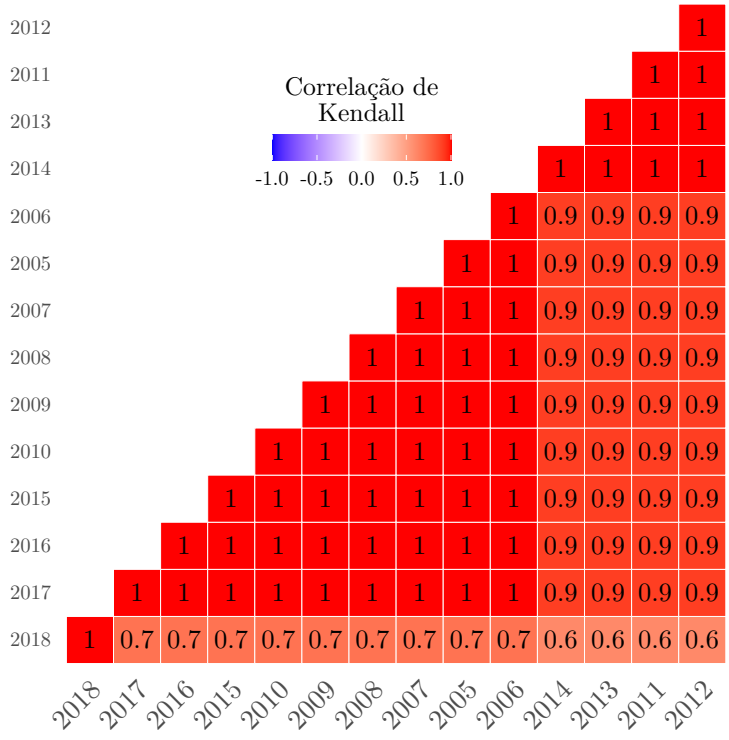


(b) O Q do periódico, n anos depois da sua criação, considerando somente as citações recebidas dentro do período.

Figura 6.3: Evolução do Q dos periódicos.

Raking 2018 pelos Indicadores		
Posição	SJR	Q
1	RMP	RMP
2	PRL	PRL
3	PRD	PRD
4	PRB	PRB
5	PRC	PRA
6	PRA	PRC

(a) Comparativo entre os rankings induzidos pelo SJR e pelo Q para o ano de 2018. Coeficiente de kendall (0.87, arredondado para 0.9).



(b) Grau de correlação dos rankings anuais dos seis periódicos (RMP, PRL, PRD, PRB, PRA,PRC) induzidos pelo SJR. Cada elemento da matriz triangular mostra a correlação entre dois rankings em anos diferentes. Um coeficiente 1 significa que o ranking não sofreu mudanças.

Figura 6.4: Correlação entre os rankings induzidos pelo SJR e pelo Q. Diferentemente do ranking induzido pelo SJR que sofreu mudanças nas posições entre 2004 e 2019, o ranking pelo Q se manteve sem alterações, e apresenta boa correlação com o ranking induzido pelo SJR recente (ranking 2018).

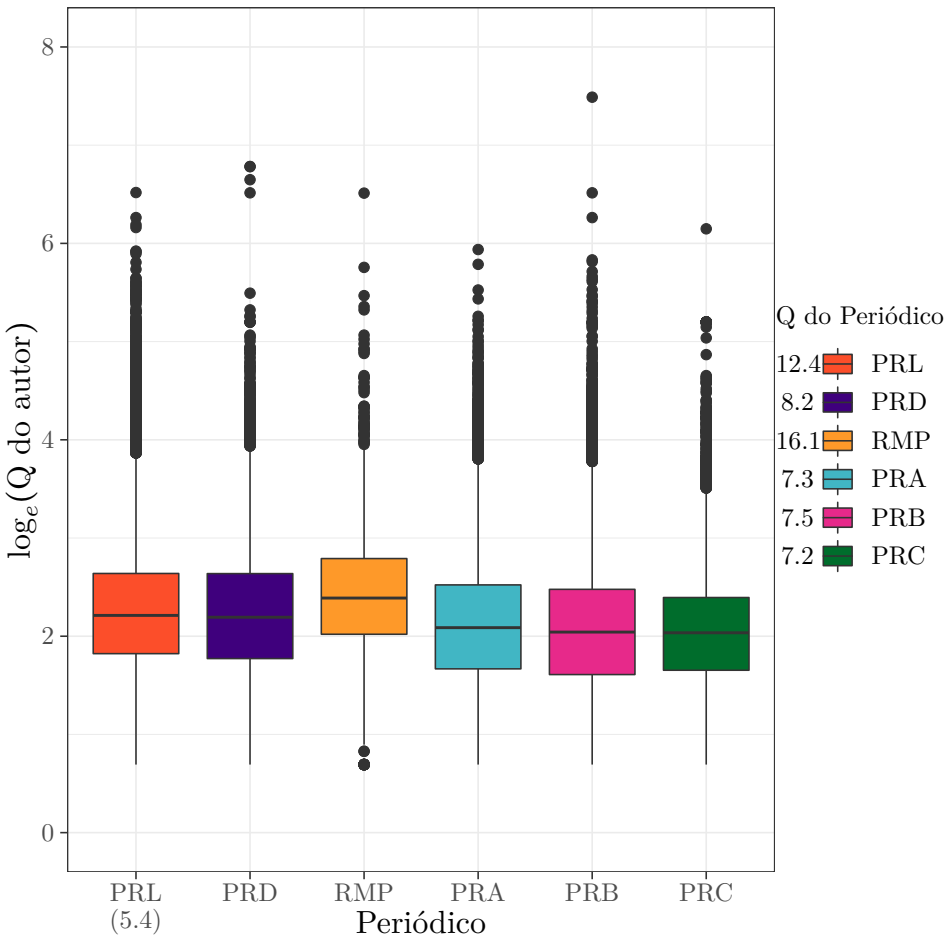


Figura 6.5: Boxplots comparativos de valores de Q de autores por periódico. Existe uma correlação positiva fraca entre o Q do periódico e o Q do autor.

Tabela 6.4: Valor da estatística KPSS para as séries randomizada e observada. A série observada contem o Q do periódico medido para 19, 22, 23, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41, 44 anos após a sua criação. A série randomizada é composta desses mesmos valores, porém misturados 39, 25, 19, 35, 28, 29, 26, 37, 34, 30, 33, 27, 32, 41, 38, 31, 22, 36, 44, 23.

Para o periódico	Série misturada	Serie como observada
PRL	0.2102	0.777*
PRC	0.2391	0.7518*
PRD	0.1948	0.7727*
PRB	0.2387	0.7887*
RMP	0.3006	0.4934
PRA	0.1781	0.7751*

Hipótese nula - a série é estácionária (média e variância constante).
* A nula é rejeitada em um nível de 1% de significância (valores da estatística acima de 0.739).

6.4 Discussão e Conclusão

Rankings de periódicos são um elemento chave quando indivíduos ou instituições avaliam pesquisas e selecionam periódicos. Nesse trabalho, nós propomos o Q para periódico. O ranking produzido pela medida Q é muito estável, porque ele é menos impactado pela cobertura da base de dados (e, também pelo tempo) do que o ranking induzido pela métrica SJR, e consequentemente, pelo fator de impacto. As métricas SJR e o fator de impacto produziram classificações fortemente correlacionadas (GONZÁLEZ-PEREIRA; GUERRERO-BOTE; MOYA-ANEGÓN, 2010).

O Q do autor é uma medida não cumulativa que captura a sua habilidade de transformar suas ideias em descobertas com um certo impacto (SINATRA et al., 2016). No entanto, surpreendentemente, para periódicos, esta medida parece aumentar ao longo do tempo (média e variância não constantes). Nós especulamos que a inflação de citação presente no dado da APS afeta as estatísticas média e variância do Q. Mais pesquisa focada e aprofundada é necessária para confirmar isso.

Por outro lado, embora o número médio de publicações por cientista (e referências por artigo) tem aumentado constantemente ao longo do tempo, o ranking induzido pelo Q não tem sido afetado por isso - o ranking se manteve inalterado. A tendência de desaceleração do Q sugere uma classificação única. O Q do periódico define a sua posição relativa no ranking.

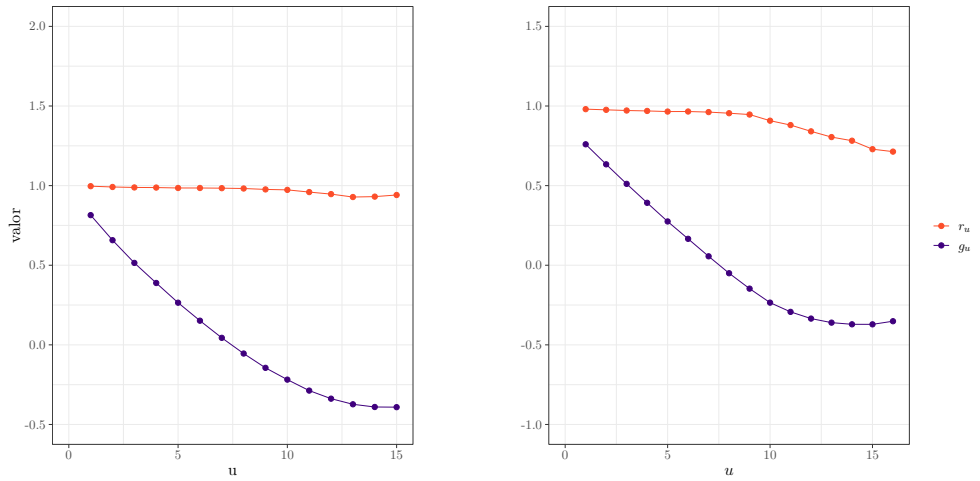
Quando selecionando pessoal para vagas abertas de pesquisadores em universidades é crítico identificar aqueles que farão pesquisas de qualidade, assim atraindo novas colaborações e conseguindo apoio para seus projetos futuros. Desenvolvimentos recentes nessa área têm mostrado a necessidade por modelos preditivos de impacto futuro confiáveis. Um tal modelo que combine diferentes fatores relacionados a uma carreira em uma predição (e.g, fatores ligados ao local de publicação) ajudará comitês de seleção a tomar decisões mais rápidas e menos subjetivas.

Por outro lado, interessantemente, para valores altos de Q_j e Q_i, o Q do periódico e o Q do pesquisador, respectivamente, foi encontrado uma correlação positiva. A descoberta anterior é consistente com pesquisas anteriores, e.g, (ACUNA; ALLESINA; KORDING, 2012; WEIHS; ETZIONI, 2017; LEE, 2019), que encontraram que, entre outros variáveis, aquelas relacionadas ao local de publicação, i.e., o número de publicações em periódicos importantes em que o cientista publicou e o número de publicações em diferentes periódicos em que ele já publicou são variáveis preditoras de desempenho (ou impacto) futuro de um autor, i.e, predizem seu índice-h futuro.

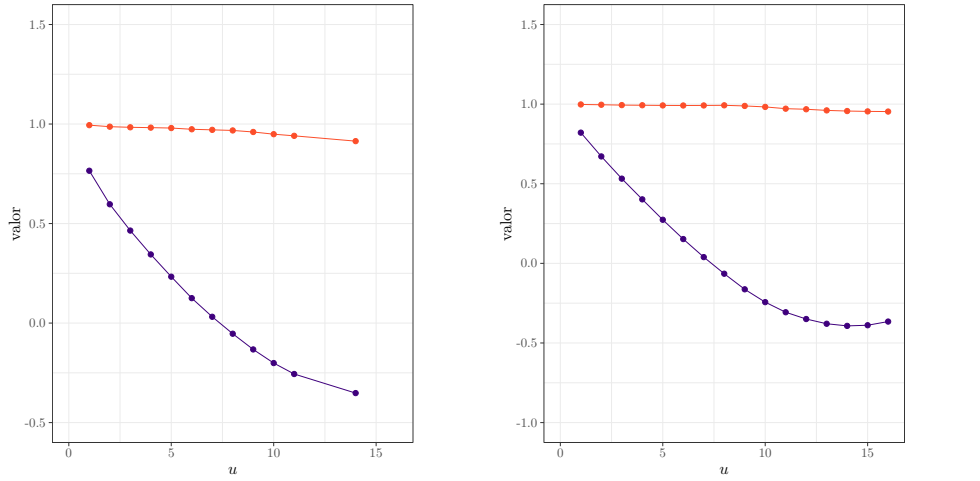
Nossa descoberta de que potenciais artigos influentes (artigos muito citados) são mais prováveis de serem publicados em periódicos na parte de cima dos rankings reforça a importância para um pesquisador de publicar em periódicos de maior impacto.

Por fim, o total de publicação e o índice-h de Hisrch são as métricas de desempenho e de impacto mais comumente usadas para avaliar pesquisadores individuais, respectivamente. As duas são medidas cumulativas, e portanto não são adequadas para comparação de pesquisadores em diferentes estágios da carreira, pois favorecem pesquisadores com carreias longas e com longas listas de publicações e prejudicam pesquisadores júnior com uma presença curta na academia.

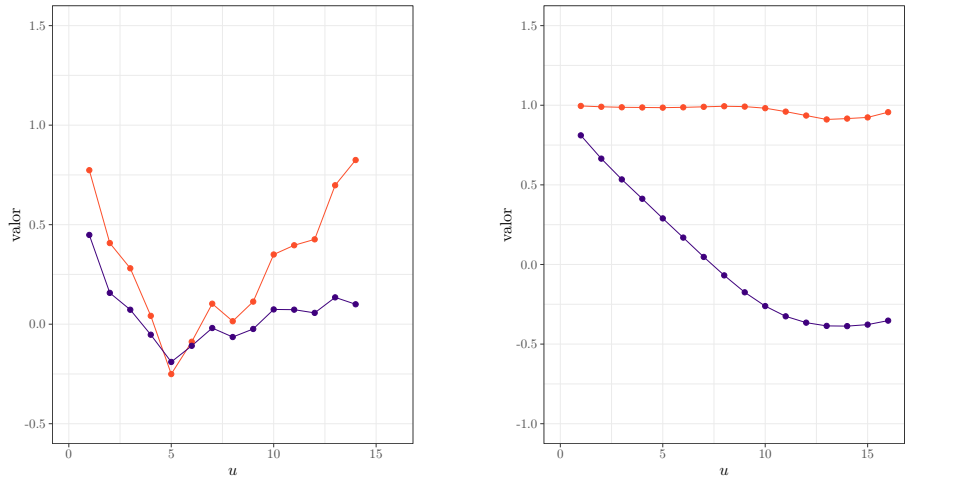
No contexto de periódicos, a avaliação de impacto recente e a comparação de periódicos em diferentes estágios de maturidade também são uma questão importante. Para um pesquisador selecionando um periódico para publicar, independentemente do tamanho e do tempo de vida do periódico, é importante conhecer não só seu impacto recente, mas também o futuro.



(a) PRA (b) PRB

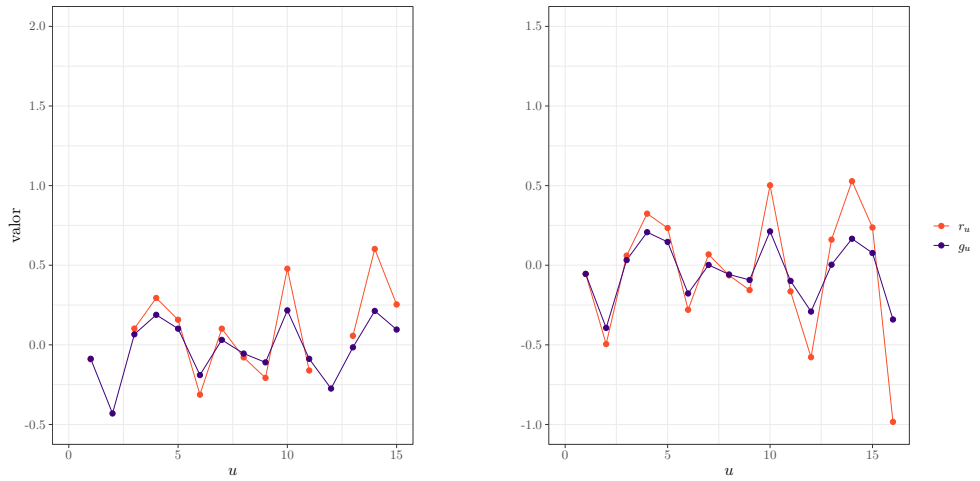


(c) PRC (d) PRD

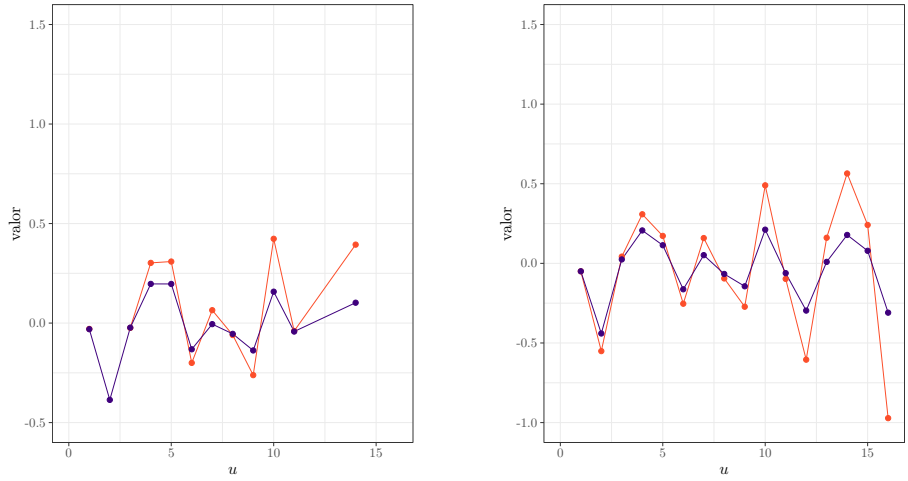


(e) RMP (f) PRL

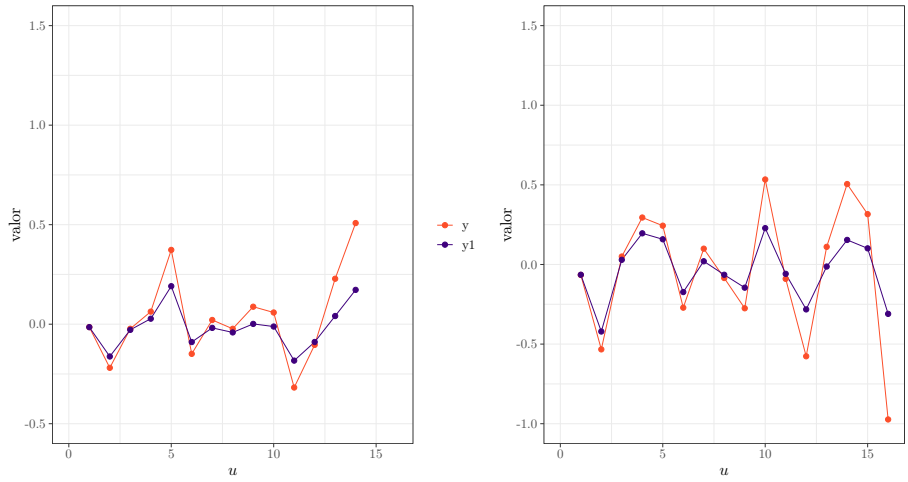
Figura 6.6: Os correlogramas das séries de valores de Q de um periódico gerados pelos métodos de autocorrelação (r_u) e autocovariância (g_u) divergem, indicando que a séries são geradas por um processo não estacionário. Cada série contém os valores de Q do periódico após t=19, 22, 23, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41, 44 anos depois da sua criação.



(a) PRA (b) PRB



(c) PRC (d) PRD



(e) RMP (f) PRL

Figura 6.7: Ao misturar os valores de Q das séries, os correlogramas gerados pelo método de autocorrelação (r_u) e autocovariância (g_u) praticamente caminham juntos, sugerindo que as séries são geradas por um processo estacionário.

Capítulo 7

Conclusões

Devido às muitas aplicações do mundo real que podem notavelmente se beneficiar de predições corretas do impacto (ou desempenho) futuro de pesquisadores individuais, o tópico tem sido muito explorado nos últimos anos. Contudo, ainda não é claro, quando um pesquisador têm uma maior probabilidade de desenvolver suas ideias e transformá-las em publicações de sucesso ao longo da carreira. Nessa tese, tentou-se identificar quais fatores relacionados aos cinco primeiros anos de carreira de um cientista contribuem mais para seu desempenho individual futuro.

As equações suportam a ideia de que diferentes fatores relacionados ao início de carreira de um cientista afetam o seu impacto futuro (e.g., o Q definitivo, o índice h futuro), e que cada um tem um peso diferente que pode aumentar ou diminuir com o tempo. Essa descoberta é importante porque saber que um determinado indicador pouco contribui para o desempenho individual futuro de um cientista pode servir de justificativa para o não uso dele em novas avaliações. (o objetivo específico 1, seção 1.1, foi alcançado.)

Ao experimentar com medidas alternativas (e.g., o Q predito do cientista júnior) e compará-las contra medidas tradicionais (e.g índice-h corrente do cientista júnior) quanto à tarefa de produzir o ranking observado de pesquisadores mais tardiamente, nós encontramos que elas são confiáveis. A implicação prática disso é a possibilidade de que comissões avaliadoras (de promoção docente, estágios probatórios, concurso) passem a embasar suas decisões nessas predições e usem o entendimento por trás de uma decisão do modelo para suportar uma decisão da comissão. As novas medidas são usadas em um arcabouço computacional para selecionar bons cientistas precocemente. (os objetivos específicos 2 e 3, seção 1.1, foram alcançados.)

É importante destacar que, apesar da elevada precisão das equações, confiar em predições sempre envolve algum risco, existem muitas armadilhas nessa abordagem e atenção deve ser redobrada para evitar injustiças. Além disso, todas as medidas anteriores (medidas de impacto corrente e medidas de potencial para impacto futuro) favoreceram pesquisadores júnior com melhor desempenho nos primeiros cinco anos de carreira. Isso justifica o uso de explicativas de modelos como um elemento adicional para ajudar comissões avaliadoras a formarem suas próprias convicções. Explicativas de modelos não substitui o juízo dos membros da comissão.

Por último, essa tese contribui com um arcabouço para suportar as decisões de governos, agências de governo, instituições de pesquisa e cientistas, quando preocupados em avaliar o provável desempenho individual futuro de cientistas júnior.

7.1 Resumo de Contribuições

Primeiramente, nossos esforços focaram em quantificar o impacto futuro de cientistas com até cinco anos de carreira, como medido pelo Q dos cientistas após o ano 15 da carreira. Nós testamos a abordagem *Deep* e a abordagem de regressão linear, e comparamos suas acurácias. Nós obtivemos resultados confiáveis mostrando que os valores preditos, por modelos inferidos a partir ambas as abordagens, são melhores do que os estimados pelo próprio modelo Q (o Q do cientista júnior, como o seu Q definitivo, i.e., após o ano 15 da carreira) quando usando somente dados dos primeiros cinco anos de publicação dos cientistas.

Depois, medimos o grau de concordância entre os rankings (de pesquisadores) preditos e observados. Especificamente, nós comparamos os rankings preditos produzidos por medidas alternativas focando em seus prováveis impactos futuros e medidas tradicionais (e.g., índice h) contra o ranking observado (o ranking de pesquisadores futuro). O estudo usou dados de 1,631 cientistas da computação extraídos do conjunto de dados da ACM, usados como aplicantes em nossos modelos de seleção. Nós encontramos medidas alternativas às medidas tradicionais. Além disso, mostramos que medidas tradicionais podem acentuar as desigualdades em desempenho e impacto entre cientistas júnior.

Por último, focou-se na construção de uma nova medida de impacto para periódicos que não fosse cumulativa. Nós decidimos estudar o Q, porque seu cálculo baseia-se somente nas citações de publicações de uma entidade individual (cientista, grupo de pesquisa, periódico ou país). Nós descobrimos que a natureza não cumulativa da medida Q de um periódico vem sendo afetada por fatores como, e.g., a mudança de comportamento de publicação dos cientistas ao longo do tempo, publicando e citando um número cada vez maior de trabalhos.

7.2 Lista de Publicações

As principais ideias deste trabalho de doutorado foram publicadas em veículos nacionais (eventos) assim como internacionais (capítulo de livro e artigo em periódico).

Lista das publicações referente à tese apresentadas em eventos científicos, e publicadas em periódicos indexados:

- BATISTA-JR, A. d. A.; GOUVEIA, F. C.; MENA-CHALCO, J. P. Identification of promising researchers through fast-and-frugal heuristics. In:_____. *Predicting the Dynamics of Research Impact*. Cham: Springer International Publishing, 2021. p. 195-207.
- BATISTA-JR, A. de A.; GOUVEIA, F. C.; MENA-CHALCO, J. P. Predicting the Q ofjunior researchers using data from the first years of publication. *Journal of Informetrics*, v. 15, n. 2, p. 101130, 2021.
- BATISTA-JR, A. de A.; MENA-CHALCO, J. P. Identificação de acadêmicos promissores de sucesso científico através de programação linear inteira. In: Anais do SBPO 2019. Limeira, SP, Brasil: [s.n.], 2019. p. 107692.
- BATISTA-JR, A. de A.; MENA-CHALCO, J. P. Assessing the future scientific impact of scientists. In: Poster apresentado em Terceiro Encontro Paulista de Pós-graduandos em Computação (EPPC). São Paulo, SP, Brasil, 2019.

Adicionalmente, foram realizados trabalhos em colaboração com diferentes pesquisadores relacionados a outras temáticas, mas que contribuíram para o aprendizado e o debate nas ideias discutidas no doutorado:

- SAMPAIO, R. B.; BATISTA-JR, A. de A.; FERREIRA, B. S.; BARRETO, M. L.; MENA-CHALCO, J. P. Scientometric Analysis of Research Output from Brazil in Response to the Zika Crisis Using e-Lattes. *Journal of Data and Information Science*, v. 5, n. 4, p.137–146, 2020.
- SAMPAIO, R. B.; BATISTA-JR, A. de A.; MENA-CHALCO, J. P. e-Lattes: um novo arcabouço em linguagem R para análise do currículo Lattes. In: Encontro Brasileiro de Bibliometria e Cientometria, v. 6, n. 6, p. 6^o Encontro Brasileiro de Bibliometria e Cientometria, 2018.
- ARAÚJO, P. V. de; BATISTA-JR, A. de A.; GAZZIRO, M. Esquema de votação seguro e transparente através de encriptação homomórfica. In: ANAIS do IV Workshop de Tecnologia Eleitoral. São Paulo: SBC, 2019. p. 25-36.

7.3 Limitações e Trabalhos Futuros

Prever o futuro de alguém é uma tarefa complicada e o seu uso exige cautela para não causar danos involuntários ao avaliado, por achar que o futuro dele será menor do que o de outros ou favorecê-lo. Dados incompletos, eventos atípicos, desigualdades de gênero e as especificidades de cada campo científico são algumas dificuldades enfrentadas neste trabalho. Para seus usos práticos, é de suma importância entender o raciocínio de modelos caixas pretas diante destas e outras questões.

Os tópicos seguintes são reservados para trabalho futuro. No lugar do modelo indicar um único valor do Q futuro do cientista alvo seria preferível apontar um intervalo de valores potenciais. Por exemplo, o cientista A terá um Q entre x e y. Avaliar um colega por um número parece simples de mais. Também, pretende-se estudar este problema a partir da perspectiva de sistemas complexos considerando calculo de erro. Adicionalmente, pretende-se dotar o modelo de capacidades para tratar eventos atípicos (covid, gravidez) usando teoria do caos.

Nessa tese, focou-se no desempenho individual (no número de artigos acadêmicos) de pesquisadores júnior e no impacto de suas publicações (no total de citações), porque nossos conjuntos de dados não contém outras informações comumente encontradas em bases de dados curriculares, e.g., em que instituição o pesquisador se formou. Entretanto, nós reconhecemos que essa abordagem é uma abordagem relativamente estreita. Em pesquisa futura, uma abordagem mais holística deve ser aplicada, levando em consideração outras aspectos de uma carreira em pesquisa científica, e.g., quanto tempo dedica-se para atividades de ensino, para participação em eventos científicos, bem como tempo de dedicação para pesquisa após o término do doutorado, informação de projetos de pesquisa concluídos e em quais periódicos têm publicado mais, em periódicos.

Esta pesquisa mostrou que diferentes tipos de discriminação contra minorias (pesquisadores júnior com arranques tardios, mas com sucesso futuro) são involuntariamente praticados por algoritmos de aprendizado de máquina. Estudos precisam ser feitos, que considere, *gap* de gênero e enviesamentos contra minorias.

Nós estamos agora investigando a utilização do nosso arcabouço de testes, em um contexto novo, a validação de métodos para interpretar e entender redes neurais profundas (MONTAVON; SAMEK; MÜLLER, 2018). Interessantemente, nosso arcabouço poderia ser útil nesse contexto. O treino de uma rede neural em um de nossos conjuntos de treinamento tem como resultado final esperado uma função aproximada da equação de regressão, isso tem sido mostrado em um trabalho nosso anterior. Como um resultado disso, temos que, o grau de concordância entre os coeficientes da equação de regressão e os valores para esses coeficientes correspondentes na visão dos métodos de interpretação pode ser medido. Um grau de correlação próximo significa que o método captura o entendimento por trás de uma dada predição, e zero o contrário. Nossos resultados preliminares parecem promissores.

Devido à amostra ter sido retirada a partir de bases de dados específicas indexando trabalhos de pesquisadores da Ciência da Computação e Física, replicações desse estudo a partir de outras áreas são necessárias. As descobertas notadas aqui podem ser específicas para as áreas estudas.

Uma outra importante questão é a abordagem longitudinal que foi tomada dentro da base de dados quanto à dimensão tempo. Nós pensamos que essa escolha poderia grandemente impactar a força preditiva das equações pelo fato que o comportamento de publicação varia expressivamente entre diferentes períodos de tempo.

Por último, as descobertas desse trabalho confirmam a importância de explicativas de modelos para aplicações que necessitam de decisões fundamentadas. E, embora tenhamos testemunhado um significante progresso em aprendizado de máquina em anos recentes, ainda há muito espaço para melhorias em relação a interpretabilidade de máquinas.

Referências

ABRAMO, Giovanni; D'ANGELO, Ciriaco Andrea; FELICI, Giovanni. Predicting publication long-term impact through a combination of early citations and journal impact factor. **Journal of Informetrics**, v. 13, n. 1, p. 32–49, 2019.

ACUNA, Daniel E; ALLESINA, Stefano; KORDING, Konrad P. Predicting scientific success. **Nature**, v. 489, n. 201, p. 201–202, 2012.

ANDERSEN, Jens Peter; NIELSEN, Mathias Wullum. Google Scholar and Web of Science: Examining gender differences in citation coverage across five scientific disciplines. **Journal of Informetrics**, v. 12, n. 3, p. 950–959, 2018.

ARORA, Sanjeev. Mathematics of Machine Learning: An Introduction. In_____. **Proceedings of the International Congress of Mathematicians (ICM 2018)**. Rio de janeiro: World Scientific, 2018. p. 377–390.

AYAZ, Samreen; MASOOD, Nayyer; ISLAM, Muhammad Arshad. Predicting scientific impact based on h-index. **Scientometrics**, v. 114, n. 3, p. 993–1010, 2018.

BAI, Xiaomei et al. Quantifying Success in Science: An Overview. **IEEE Access**, v. 8, p. 123200–123214, 2020.

BAR-ILAN, Judit; LEVENE, Mark; LIN, Ayelet. Some measures for comparing citation databases. **Journal of Informetrics**, v. 1, n. 1, p. 26–34, 2007.

BARTNECK, Christoph; KOKKELMANS, Servaas. Detecting h-index manipulation through self-citation analysis. **Scientometrics**, Springer Netherlands, v. 87, n. 1, p. 85–98, 2011.

BATISTA-JR, Antônio de Abreu; GOUVEIA, Fábio Castro; MENA-CHALCO, Jesús P. Predicting the Q of junior researchers using data from the first years of publication. **Journal of Informetrics**, v. 15, n. 2, p. 101130, 2021.

BAZELEY, Pat. Defining Early Career in Research. **Higher Education**, v. 45, n. 3, p. 257–279, 2003.

BORGMAN, Christine L.; FURNER, Jonathan. Scholarly communication and bibliometrics. **Annual Review of Information Science and Technology**, v. 36, n. 1, p. 2–72, 2002.

BORNMANN, Lutz; GUNS, Raf et al. Which aspects of the Open Science agenda are most relevant to scientometric research and publishing? An opinion paper. **Quantitative Science Studies**, v. 2, n. 2, p. 438–453, 2021.

BORNMANN, Lutz; GUNS, Raf et al. Which aspects of the Open Science agenda are most relevant to scientometric research and publishing? An opinion paper. **Quantitative Science Studies**, v. 2, n. 2, p. 438–453, 2021.

BORNMANN, Lutz; WILLIAMS, Richard. Can the journal impact factor be used as a criterion for the selection of junior researchers? A large-scale empirical study based on ResearcherID data. **Journal of Informetrics**, v. 11, n. 3, p. 788–799, 2017.

BRAUN, Tibor; GLÄNZEL, Wolfgang; SCHUBERT, András. A Hirsch-type index for journals. **Scientometrics**, v. 69, n. 1, p. 169–173, 2006.

BREMBS, Björn. Prestigious Science Journals Struggle to Reach Even Average Reliability. **Frontiers in Human Neuroscience**, v. 12, p. 37, 2018.

BREMBS, Björn; BUTTON, Katherine; MUNAFÒ, Marcus. Deep impact: unintended consequences of journal rank. **Frontiers in human neuroscience**, Frontiers Media S.A., v. 7, p. 291–291, jun. 2013.

BRITO, Ricardo; RODRÍGUEZ-NAVARRO, Alonso. Evaluating research and researchers by the journal impact factor: Is it better than coin flipping? **Journal of Informetrics**, v. 13, n. 1, p. 314–324, 2019.

CARVALHO, Diogo V.; PEREIRA, Eduardo M.; CARDOSO, Jaime S. Machine Learning Interpretability: A Survey on Methods and Metrics. **Electronics**, v. 8, n. 8, 2019.

CLAUSET, Aaron; LARREMORE, Daniel B.; SINATRA, Roberta. Data-driven predictions in the science of science. **Science**, v. 355, n. 6324, p. 477–480, 2017.

DI IORIO, Angelo et al. Investigating Facets to Characterise Citations for Scholars. In_____. **Semantics, Analytics, Visualization**. Cham: Springer International Publishing, 2018. p. 150–160.

DIACONIS, Persi; GRAHAM, R. L. Spearman's Footrule as a Measure of Disarray. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 39, n. 2, p. 262–268, 1977.

DONG, Yuxiao; JOHNSON, Reid A.; CHAWLA, Nitesh V. Will This Paper Increase Your h-Index? Scientific Impact Prediction. In: PROCEEDINGS of the Eighth ACM International Conference on Web Search and Data Mining. Shanghai, China: Association for Computing Machinery, 2015. (WSDM '15), p. 149–158.

DUCTOR, Lorenzo et al. Social Networks and Research Output. **The Review of Economics and Statistics**, v. 96, n. 5, p. 936–948, 2014.

EGGHE, Leo. Dynamic H-Index: The Hirsch Index in Function of Time: Brief Communication. **Journal of the American Society for Information Science and Technology**, John Wiley & Sons, Inc., USA, v. 58, n. 3, p. 452–454, 2007.

EINSTEIN, A.; PODOLSKY, B.; ROSEN, N. Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? **Physical Review**, American Physical Society, v. 47, p. 777–780, 10 1935.

EL HECHI, M. W. et al. Leveraging interpretable machine learning algorithms to predict postoperative patient outcomes on mobile devices. **Surgery**, v. 169, n. 4, p. 750–754, 2021.

FAGIN, Ronald; KUMAR, Ravi; SIVAKUMAR, D. Comparing Top k Lists. **SIAM Journal on Discrete Mathematics**, v. 17, n. 1, p. 134–160, 2003.

FERNANDES, Jason D. et al. A survey-based analysis of the academic job market. **eLife**, eLife Sciences Publications, Ltd, v. 9, e54097, 2020.

FORTUNATO, Santo et al. Science of science. **Science (New York, N.Y.)**, v. 359, n. 6379, eaao0185, 2018.

FRANK, Michael C. N-Best Evaluation for Academic Hiring and Promotion. **Trends in Cognitive Sciences**, v. 23, n. 12, p. 983–985, 2019.

GARCÍA-PÉREZ, Miguel A. Limited validity of equations to predict the future h index. **Scientometrics**, v. 96, n. 3, p. 901–909, 2013.

GARFIELD, Eugene. Journal Impact Factor: A Brief Review. **Canadian Medical Association Journal**, Canadian Medical Association, v. 161, n. 8, p. 979–980, 1999.

GEVREY, Muriel; DIMOPOULOS, Ioannis; LEK, Sovan. Review and comparison of methods to study the contribution of variables in artificial neural network models. **Ecological Modelling**, v. 160, n. 3, p. 249–264, 2003.

GIBBS KENNETH D, Jr et al. Research: Decoupling of the minority PhD talent pool and assistant professor hiring in medical school basic science departments in the US. Edição: Peter Rodgers. **eLife**, eLife Sciences Publications, Ltd, v. 5, e21393, 2016.

GOGOGLOU, Antonia. **Complex Networks and Machine Learning in Scientometrics**. 2017. Tese (Doutorado) – Faculty of Sciences, School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece.

GONZÁLEZ-PEREIRA, Borja; GUERRERO-BOTE, Vicente P.; MOYA-ANEGÓN, Félix. A new approach to the metric of journals' scientific prestige: The SJR indicator. **Journal of Informetrics**, v. 4, n. 3, p. 379–391, 2010.

HERMAN, Eti et al. The impact of the pandemic on early career researchers: what we already know from the internationally published literature. **Profesional de la Información**, v. 30, n. 2, 2021.

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

HOLDEN, G.; ROSENBERG, G.; BARKER, K. Bibliometrics: a potential decision making aid in hiring, reappointment, tenure and promotion decisions. **Social work in health care**, v. 41, n. 3-4, p. 67–92, 2005.

HOU, Jie et al. Prediction methods and applications in the science of science: A survey. **Computer Science Review**, v. 34, p. 100197, 2019.

HOU, Jie et al. Prediction methods and applications in the science of science: A survey. **Computer Science Review**, v. 34, p. 100197, 2019.

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

KANELLOS, I. et al. Impact-Based Ranking of Scientific Publications: A Survey and Experimental Evaluation. **IEEE Transactions on Knowledge and Data Engineering**, v. 8, n. 14, p. 1–18, 2019.

KAYAL, Subhradeep et al. A Framework to Automatically Extract Funding Information from Text. In_____. **Machine Learning, Optimization, and Data Science**. Cham: Springer International Publishing, 2019. p. 317–328.

KE, Qing et al. Defining and identifying Sleeping Beauties in science. **Proceedings of the National Academy of Sciences**, National Academy of Sciences, v. 112, n. 24, p. 7426–7431, 2015.

KENDALL, Maurice; GIBBONS, Jean D. **Rank Correlation Methods**. 5. ed. [S.l.: s.n.], 1990.

KIM, Been; KHANNA, Rajiv; KOYEJO, Oluwasanmi. Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability. In: PROCEEDINGS of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc., 2016. (NIPS'16), p. 2288–2296.

KÖCHLING, Alina; WEHNER, Marius Claus. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. **Business Research**, v. 13, n. 3, p. 795–848, nov. 2020.

KOTSIANTIS, S. B.; ZAHARAKIS, I. D.; PINTELAS, P. E. Machine learning: a review of classification and combining techniques. **Artificial Intelligence Review**, v. 26, n. 3, p. 159–190, 2006.

KWIATKOWSKI, Denis et al. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? **Journal of Econometrics**, v. 54, n. 1, p. 159–178, 1992.

LARIVIÈRE, Vincent; ARCHAMBAULT, Éric; GINGRAS, Yves. Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). **Journal of the American Society for Information Science and Technology**, v. 59, n. 2, p. 288–296, 2008.

LAUDEL, Grit; GLÄSER, Jochen. From apprentice to colleague: The metamorphosis of Early Career Researchers. **Higher Education**, v. 55, n. 3, p. 387–406, 2008.

LAURANCE, William F. et al. Predicting Publication Success for Biologists. **BioScience**, v. 63, n. 10, p. 817–823, 2013.

LEE, Danielle H. Predicting the research performance of early career scientists. **Scientometrics**, v. 121, n. 3, p. 1481–1504, 2019.

LEHMANN, Sune; JACKSON, Andrew D.; LAUTRUP, Benny E. Measures for measures. **Nature**, v. 444, n. 7122, p. 1003–1004, dez. 2006.

LETCHFORD, Adrian; MOAT, Helen Susannah; PREIS, Tobias. The advantage of short paper titles. **Royal Society Open Science**, v. 2, n. 8, p. 150266, 2015.

LEYDESDORFF, Loet; MILOJEVIĆ, Staša. Scientometrics. In: WRIGHT, James D. (Ed.). **International Encyclopedia of the Social & Behavioral Sciences (Second Edition)**. Second Edition. Oxford: Elsevier, 2015. p. 322–327.

LI, Jiang; YE, Fred Y. Distinguishing Sleeping Beauties in Science. **Scientometrics**, Springer-Verlag, Berlin, Heidelberg, v. 108, n. 2, p. 821–828, 2016.

LI, Weihua; ASTE, Tomaso et al. Early coauthorship with top scientists predicts success in academic careers. **Nature Communications**, v. 10, n. 1, p. 5170, nov. 2019.

LINDAHL, Jonas. **In search of future excellence : bibliometric indicators, gender differences, and predicting research performance in the early career**. 2020. f. 60. Tese (Doutorado) – Umeå University, Department of Sociology.

_____. Predicting research excellence at the individual level: The importance of publication rate, top journal publications, and top 10% publications in the case of early career mathematicians. **Journal of Informetrics**, v. 12, n. 2, p. 518–533, 2018.

LIU, Jiaying; TANG, Tao et al. Understanding the advisor–advisee relationship via scholarly data analysis. **Scientometrics**, v. 116, n. 1, p. 161–180, 2018.

LIU, Zheng; XIE, Xing; CHEN, Lei. Context-Aware Academic Collaborator Recommendation. In: PROCEEDINGS of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, United Kingdom: Association for Computing Machinery, 2018. (KDD '18), p. 1870–1879.

MARTIN, Travis et al. Coauthorship and citation patterns in the Physical Review. **Physical Review E**, American Physical Society, v. 88, p. 012814, 1 jul. 2013.

MAZLOUMIAN, Amin. Predicting Scholars' Scientific Impact. **PLOS ONE**, Public Library of Science, v. 7, n. 11, p. 1–5, nov. 2012.

MCELRATH, Karen. Gender, Career Disruption, and Academic Rewards. **The Journal of Higher Education**, Ohio State University Press, v. 63, n. 3, p. 269–281, 1992.

MERTON, Robert K. The Matthew Effect in Science. **Science**, American Association for the Advancement of Science, v. 159, n. 3810, p. 56–63, 1968.

MINGERS, John; LEYDESDORFF, Loet. A review of theory and practice in scientometrics. **European Journal of Operational Research**, v. 246, n. 1, p. 1–19, 2015.

MITCHELL, Thomas M. **Machine Learning**. 1. ed. USA: McGraw-Hill, Inc., 1997. ISBN 0070428077.

MJOLSNESS, E.; DECOSTE, D. Machine learning for science: state of the art and future prospects. **Science**, v. 293, n. 5537, p. 2051–2055, 2001.

MONTAVON, Grégoire; SAMEK, Wojciech; MÜLLER, Klaus-Robert. Methods for interpreting and understanding deep neural networks. **Digital Signal Processing**, v. 73, p. 1–15, 2018.

NIELSEN, Bent. Correlograms for Non-Stationary Autoregressions. **Journal of the Royal Statistical Society. Series B (Statistical Methodology)**, [Royal Statistical Society, Wiley], v. 68, n. 4, p. 707–720, 2006.

PENNER, Orion et al. On the Predictability of Future Impact in Science. **Scientific Reports**, v. 3, n. 3052, 2013.

PERC, Matjaž. Zipf's law and log-normal distributions in measures of scientific output across fields and institutions: 40 years of Slovenia's research as an example. **Journal of Informetrics**, v. 4, n. 3, p. 358–364, 2010.

PETERS, Gjalt-Jorn Ygram. Why not to use the journal impact factor as a criterion for the selection of junior researchers: A comment on Bornmann and Williams (2017). **Journal of Informetrics**, Elsevier Science, v. 11, n. 3, p. 888–891, 2017.

PETERSEN, Alexander M. et al. Methods to account for citation inflation in research evaluation. **Research Policy**, v. 48, n. 7, p. 1855–1865, 2019.

RAAN, Anthony F. J. van. Sleeping Beauties in science. **Scientometrics**, v. 59, n. 3, p. 467–472, 2004.

REDNER, Sidney. Citation Statistics from 110 Years of Physical Review. **Physics Today**, v. 58, n. 6, p. 49–54, 2005.

RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: PROCEEDINGS of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: Association for Computing Machinery, 2016. (KDD '16), p. 1135–1144.

SAHA, Somnath; SAINT, Sanjay; CHRISTAKIS, Dimitri A. Impact factor: a valid measure of journal quality? **Journal of the Medical Library Association : JMLA**, Medical Library Association, v. 91, n. 1, p. 42–46, 2003.

SAMMUT, Claude; WEBB, Geoffrey I. (Ed.). Holdout Evaluation. In: **Encyclopedia of Machine Learning**. Boston, MA: Springer US, 2010. p. 506–507.

SARIGÖL, Emre et al. Predicting scientific success based on coauthorship networks. **EPJ Data Science**, v. 3, n. 1, p. 9, 2014.

SCHWEITZER, Frank. Scientific networks and success in science. **EPJ Data Science**, v. 3, n. 1, p. 35, 2014.

SEGLEN, Per O. The skewness of science. **Journal of the American Society for Information Science**, v. 43, n. 9, p. 628–638, 1992.

SERRA, Mônica da Costa et al. Research integrity and scientific misconduct: International guidelines, national standards and cooperative research. Ethical and legal reflections. **Research, Society and Development**, v. 10, n. 9, e46110918219, jul. 2021.

SHEN, Hua-Wei; BARABÁSI, Albert-László. Collective credit allocation in science. **Proceedings of the National Academy of Sciences**, National Academy of Sciences, v. 111, n. 34, p. 12325–12330, 2014.

SINATRA, Roberta et al. Quantifying the evolution of individual scientific impact. **Science**, v. 354, n. 6312, 2016.

SUGIMOTO, Cassidy R. Scientific success by numbers. **Nature**, v. 593, n. 7857, p. 30–31, 2021.

TANG, Jie et al. ArnetMiner: Extraction and Mining of Academic Social Networks. In: PROCEEDINGS of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2008. p. 990–998.

VAN DIJK, David; MANOR, Ohad; CAREY, Lucas B. Publication metrics and success on the academic job market. **Current Biology**, v. 24, n. 11, r516–r517, 2014.

VUONG, Quang H. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. **Econometrica**, v. 57, n. 2, p. 307–333, 1989.

WANG, Dashun; BARABÁSI, Albert-László. The Q-Factor. In: THE Science of Science. [S.l.]: Cambridge University Press, 2021. p. 60–70.

WANG, Dashun; SONG, Chaoming; BARABÁSI, Albert-László. Quantifying Long-Term Scientific Impact. **Science**, American Association for the Advancement of Science, v. 342, n. 6154, p. 127–132, 2013.

WANG, Fan; SHAO, Wei et al. Re-evaluation of the Power of the Mann-Kendall Test for Detecting Monotonic Trends in Hydrometeorological Time Series. **Frontiers in Earth Science**, v. 8, p. 14, 2020.

WEIHS, Luca; ETZIONI, Oren. Learning to Predict Citation-Based Impact Measures. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL). [S.l.: s.n.], 2017. p. 1–10.

WILSON, Mark C.; TANG, Zhou. Noncumulative measures of researcher citation impact. **Quantitative Science Studies**, v. 1, n. 3, p. 1309–1320, 2020.

WOOLSTON, Chris. Pandemic darkens postdocs' work and career hopes. **Nature**, v. 585, n. 7824, p. 309–312, 2020.

WRIGHT, Charles B.; VANDERFORD, Nathan L. What faculty hiring committees want. **Nature Biotechnology**, v. 35, n. 9, p. 885–887, 2017.

YAN, Rui et al. Citation Count Prediction: Learning to Estimate Future Citations for Literature. In: PROCEEDINGS of the 20th ACM International Conference on Information and Knowledge Management. Glasgow, Scotland, UK: Association for Computing Machinery, 2011. (CIKM '11), p. 1247–1252.

YUCESOY, Burcu; BARABÁSI, Albert-László. Untangling performance from success. **EPJ Data Science**, v. 5, n. 1, p. 17, 2016.

ZENG, An; SHEN, Zhesi et al. The science of science: From the perspective of complex systems. **Physics Reports**, v. 714-715, p. 1–73, 2017.

ZENG, Xiangwei; RONG, Zhihai. Evolution of the Physics Citation Network with Motifs. In: 2021 40th Chinese Control Conference (CCC). [S.l.: s.n.], 2021. p. 776–780.

ZHANG, Xinyang; WANG, Ningfei et al. Interpretable deep learning under fire. English (US). In: PROCEEDINGS of the 29th USENIX Security Symposium. [S.l.]: USENIX Association, 2020. (Proceedings of the 29th USENIX Security Symposium), p. 1659–1676.

ZHANG, Yajie; YU, Qiang. What is the best article publishing strategy for early career scientists? **Scientometrics**, v. 122, n. 1, p. 397–408, 2020.

ZHOU, Zhi-Hua. Introduction. In: MACHINE Learning. Singapore: Springer Singapore, 2021. p. 1–24.

Apêndice A

Estruturas de Dados usadas na Tese

Neste apêndice são apresentadas representações das estruturas de dados utilizadas neste trabalho. Na Figura A.1, a tabela hash a2p é apresentada. Na Figura A.2, a tabela hash p2a é mostrada, e na Figura A.3, a tabela hash p2p é descrita. Finalmente, nas Figuras A.5 e A.4 são apresentadas as tabelas hash, p2v e year.

Exemplos da utilização destas estruturas de dados são dados no Apêndice C, e como carregá-las, no Apêndice B.

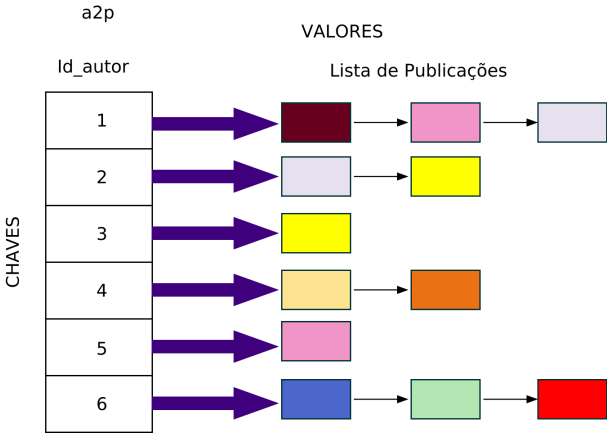


Figura A.1: Estrutura de dados representando a lista de publicações de cada autor identificado no conjunto de dados. a2p: nome do hash; Chave: o identificador do autor; Valor: a lista de identificadores das suas publicações.

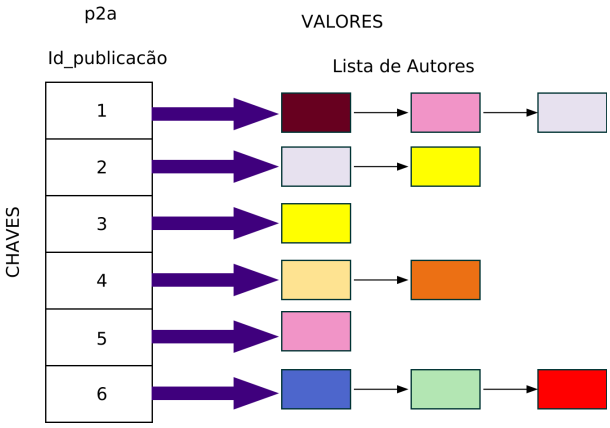


Figura A.2: Estrutura de dados mantendo a informação dos autores de cada publicação. p2a: nome do hash; Chave: o identificador da publicação ; Valor: a lista de identificadores dos autores da publicação.

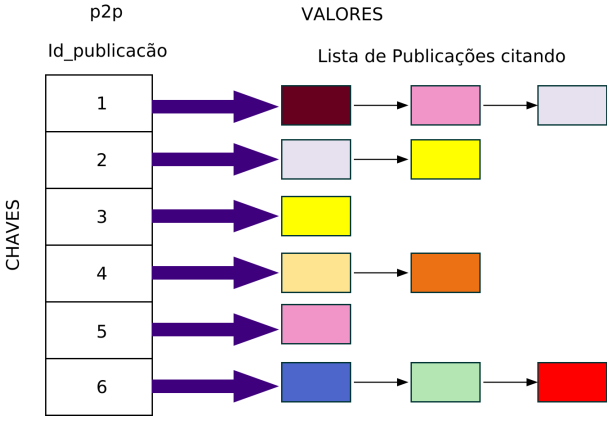


Figura A.3: Estrutura de dados *Hash* representando a lista de publicações citando uma publicação. p2p: nome do hash; Chave: o identificador da publicação ; Valor: a lista de identificadores das publicações citanto a publicação.

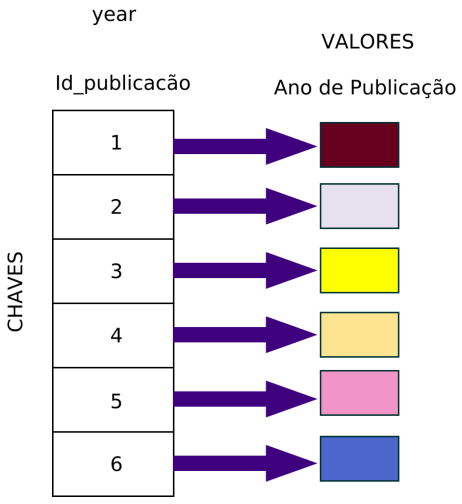


Figura A.4: Estrutura de dados mantendo a informação do ano de publicação de cada artigo no conjunto de dados. year: nome do hash; Chave: o identificador da publicação; Valor: o ano da publicação.

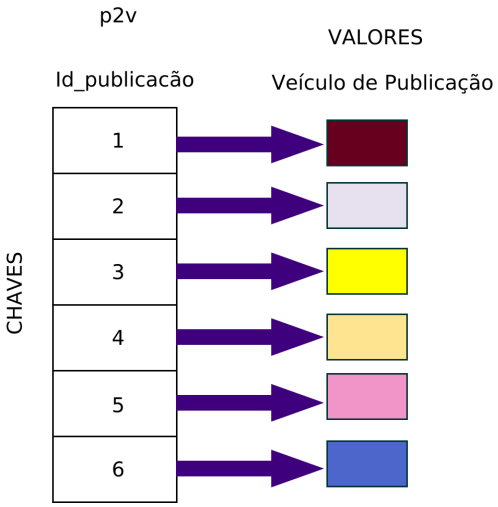


Figura A.5: Estrutura de dados mantendo a informação do veículo de publicação de cada artigo. p2v: nome do hash; Chave: o identificador da publicação; Valor: nome do veículo de publicação.

Apêndice B

Codigo R para Leitura de Dados

Neste apêndice é apresentado a parte do código fonte R que carrega, a partir do arquivo acm.txt, as Estruturas de Dados no Apêndice A. Rpacks e Acm são as pastas aonde estão os meus pacotes R e o arquivo acm.txt com os metadados dos artigos. O código completo pode ser acessado no endereço https://github.com/antoniodeabreu/tese.

```
install.packages("jsonlite", lib="/home/antonio.batista/antonio/Rpacks",
dependencies = TRUE)
install.packages("hash", lib="/home/antonio.batista/antonio/Rpacks",
dependencies = TRUE)

library("jsonlite", lib.loc="/home/antonio.batista/antonio/Rpacks")
library("hash", lib.loc="/home/antonio.batista/antonio/Rpacks")

a2p<-hash()
year<-hash()
p2a<-hash()
p2p<-hash()

p <- "/home/antonio.batista/antonio/Acm/acm.txt"
conn <- file(p,open="r")
lines <- readLines(conn)
close(conn)

cont<-0
cont1<-0
cont2<-0
for (i in 1:length(lines)){

  if(lines[i]!=""){
```

```
    if(regexpr("#index", lines[i],fixed = TRUE) ==1){
      id<- substr(lines[i], nchar("#index")+1, nchar(lines[i]))
      if(id!=""){cont1<-1}
    }else if(regexpr("#t", lines[i],fixed = TRUE) ==1){
      y<- substr(lines[i], nchar("#t")+1, nchar(lines[i]))
      if(y!=""){cont2<-1}
     }else if(regexpr("#@", lines[i],fixed = TRUE) ==1){
       a<-substr(lines[i], nchar("#@")+1, nchar(lines[i]))
       if(a!=""){
         e<-unlist(strsplit(a, split=",␣"))
         au<-list()
         for(i in 1:length(e)){
            au<-c(au, list(e[i]))
            if(e[i]!=""){
              .set(a2p,e[i],list())
            }
          }
          cont<-1
        }
      }
  }else{
      if(cont!=0 && cont1!=0 && cont2!=0){
      .set(year,id, y)
      .set(p2a,id, au)
      .set(p2p,id,list())
       cont<-0
       cont1<-0
       cont2<-0
      }
  }
}

cont<-0
cont1<-0
for (i in 1:length(lines)){

  if(lines[i]!=""){
    if(regexpr("#index", lines[i],fixed = TRUE) ==1){
      id<- substr(lines[i], nchar("#index")+1, nchar(lines[i]))
      if(id!=""){cont1<-1}
    }else if(regexpr("#@", lines[i],fixed = TRUE) ==1){
```

```
      a<-substr(lines[i], nchar("#@")+1, nchar(lines[i]))
      if(a!=""){
        e<-unlist(strsplit(a, split=", "))
        au<-list()
        for(i in 1:length(e)){
          if(e[i]!=""){
          au<-c(au, list(e[i]))
          }
        }
        cont<-1
      }
    }
  }else{

    if(cont!=0 && cont1!=0 ){
     for(a in au){
      .set(a2p,a,c(a2p[[a]],list(id)))
     }
     cont<-0
     cont1<-0
    }
  }

}
cont1<-0
citing<-list()
for (i in 1:length(lines)){
  if(lines[i]!=""){
    if(regexpr("#index", lines[i],fixed = TRUE) ==1){
      id<- substr(lines[i], nchar("#index")+1, nchar(lines[i]))
      if(id!=""){cont1<-1}
    }else if(regexpr("#%", lines[i],fixed = TRUE) ==1){
      citing<-c(citing, list(substr(lines[i], nchar("#%")+1, nchar(lines[i]))))
    }
  }else{
    if(cont1!=0){
      .set(p2p,id,citing)
      citing<-list()
    }
  }
}
```

Apêndice C

Codigo R para Cálculo de Métricas

Neste Apêndice encontram-se os códigos fonte das métricas de impacto (ou desempenho) de autores utilizadas nesta tese. Usa-se as estruturas de dados definidas no Apêndice

A função qtdCit retorna a quantidade de citações acumulada pela publicação p até o ano y. Exemplo de como fazer a chamada qtdCit("10.1103/PhysRevB.55.11552", 2020).

```
qtdCit<-function(p,y){  # => p identificador do artigo, y ano
    c<-0
    for(t in p2p[[p]]){
        if(!is.null( year[[t]])){
          if( as.integer(year[[t]])<=y){
            c<-c+1
          }
        }
    }
    return(c)
}
```

A função Q calcula o Q do cientista, em que a é o identificador de um cientista e n representa o limite de tempo para coletar citações. Exemplo de chamada da função Q("bernard barbara",20).

```
Q<-function(a,n){
        soma<-0
        r<-0
        for(p in a2p[[a]]){
                y<-year[[p]]
                qtd<-qtdCit(p, y+n)
                if(qtd>1){
                        soma<-soma+log(qtd,base=exp(1))
                        r<-r+1
                }
        }
        if(r>0){
                return(exp(soma/r))
        }else{
                return(r)
        }
}
```

A função indice.H calcula o índice-h do cientista, n anos após sua primeira publicação.

```
indice.H<-function(a, n){
        #encontrar o ano do primeiro artigo do autor
        pub<-1:length(a2p[[a]])
        cont<-1
        for(p in a2p[[a]]){
                pub[cont]<-as.integer(year[[p]])
                cont<-cont+1
        }
        psort<-sort(pub)
        #setar o numero de anos apos a priemira publicacao do
        #autor para o calculo do indice-h nesse periodo
        yearTwo<-psort[1]+n
        x<-rep(0,length(a2p[[a]]))
        cont<-1
        for(p in a2p[[a]]){
        #quantas citacoes ele recebeu ate yearTwo
                c<-0
                for(t in p2p[[p]]){
                        if(!is.null( year[[t]]) ){
                                if( as.integer(year[[t]])<=yearTwo){
                                        c<-c+1
                                }
                        }
                }
                x[cont]<-c
                cont<-cont+1
        }
        citation<-sort(x)
        h<-0
        for(i in length(citation):1){
                if(citation[i] <= h){
                        break
                }
                h<-h+1
        }
        return(h)
}
```