Como citar este artículo: D. A. da Silva, J. C. Marques and P. M. Tasinaffo, "Mapping the Asymmetries of Graduate Programs in Brazil: Modeling, Visualization and Reporting of Estimates," in IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 18, no. 3, pp. 268-277, Aug. 2023, doi: 10.1109/RITA.2023.3301510.

Mapping the asymmetries of graduate programs in Brazil: modelling, visualization and reporting of estimates

Daniela América da Silva, Johnny Cardoso Marques, and Paulo Marcelo Tasinaffo

***Abstract*—The graduate programs in Brazil have shown numerical growth and its evaluation system is approved by the national and international academic community. However, even with the growth and improvement of Brazilian graduate programs, the country has continental dimensions and coexists with regional asymmetries. Additionally, until nowadays has not been presented a comprehensive study on the variations in the distribution and growth of graduation programs by small geographic regions in Brazil. This study consists of a graduate program mapping by mesoregions, and the meta-model elaborated here aims to map the existing asymmetries in Education in the country to respond to scientific and technological development.**

***Index Terms*—Graduate programs, asymmetries, machine learning, mesoregions, modelling, data visualization.**

I. Introdução

A Pós-graduação stricto sensu no Brasil tem apresentado crescimento numérico desde a sua implantação e seu sistema de avaliação é aprovado pela comunidade acadêmica nacional e internacional [1]. Contribuindo para a consolidação da pós-graduação no país, sabe-se que a CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) tem duas preocupações: sua regulamentação e seu aprimoramento constante. O sistema incluiu revisão por pares, vinculou avaliação com classificação e recursos e, estabeleceu um formato de avaliação visando atingir um padrão de qualidade estabelecido e conhecido por todos. O padrão, na CAPES, foi estabelecido em comum acordo com diferentes áreas do conhecimento da comunidade acadêmica.

Por outro lado, mesmo com o aprimoramento do processo de avaliação dos programas de ensino, nosso país possui dimensões continentais e convive com assimetrias regionais. Por isso, o desafio de oferecer uma educação de qualidade requer que haja um conhecimento sobre estas assimetrias e a identificação de indicadores que apoiem a tomada de decisão e a coordenação das ações de todos os níveis governamentais [2]. A recente expansão da pós-graduação no país foi realizada em três ciclos: (i) expansão para o interior com criação de novas universidades através do SESU/MEC (Secretaria de Educação Superior/Ministério da Educação) no período de 2003 a 2006; (ii) reestruturação e expansão das universidades existentes através do Reuni (Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais) no período de 2007 a 2012 e, (iii) expansão com integração regional e internacional no período de 2008 a 2010 [3].

Além do programa Reuni, houve também o programa CsF (Ciência sem Fronteiras) e o programa ProEB (Programa de Mestrado Profissional para Professores da Educação Básica). O CsF foi fruto do esforço do MCTI (Ministério de Ciência e Tecnologia), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico)/MEC e CAPES/MEC no período de 2011 a 2014 que promoveu a consolidação, expansão e internacionalização da ciência e tecnologia, da inovação e da competitividade brasileira por meio do intercâmbio e da mobilidade internacional [4]. O ProEB foi lançado em 2011 e está vigente atualmente visando qualificar educadores em exercício na rede pública de ensino, estadual ou municipal [5].

Adicionalmente, a tecnologia é uma forma de modernizar o governo e fornecer serviços sociais de forma mais eficiente e, começa a ser amplamente utilizada pelos governos. Porém há uma ideia de que, a utilização de um algoritmo é novo, interessante, diferente e inovador, entretanto é necessário entender o que esses algoritmos poderão fazer. É importante conhecer bem como os modelos trabalham antes de serem introduzidos [6].

Especificamente, este trabalho tentará convencer a comunidade científica internacional e o governo federal brasileiro de que as assimetrias em nosso ambiente de pós-graduação poderão ser observadas através de técnicas de inteligência artificial, utilizando dados de plataformas abertas e, utilizando guias práticos éticos de desenvolvimento de sistemas inteligentes. Em uma visão macro, este trabalho contribui para a análise de assimetrias e a utilização da inteligência artificial para simplificar a criação de modelos, facilmente compreendê-los e explicá-los em linguagem natural e, apoiar a tomada de decisão e coordenação de ações nos níveis governamentais.

II. Revisão de Literatura

Este estudo seguiu um método de mapeamento sistemático [7] [8] com o objetivo de apresentar uma visão geral de uma área de pesquisa para relatar a quantidade e tipo de literatura e resultados que são publicados nela. O processo de mapeamento sistemático compreende 3 etapas: (1) a identificação da literatura relevante, (2) a composição de um esquema de classificação e (3) o mapeamento da literatura [7]. O método foi usado para examinar o corpo da pesquisa existente e para

Daniela América da Silva, Johnny Cardoso Marques and Paulo Marcelo Tasinaffo are with The Electronic and Computer Engineering Program, Area: Informatics, Instituto Tecnológico de Aeronáutica e-mail:damerica@ita.br.

Manuscript received May 18, 2023; revised Jul 05, 2023.

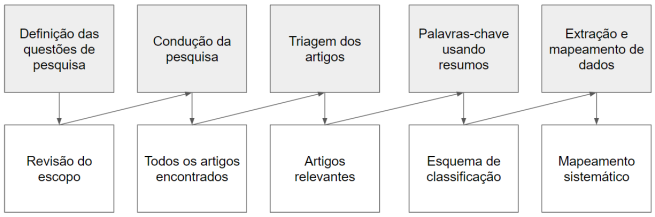


Fig. 1: Mapeamento Sistemático de Literatura

compreender a natureza da pesquisa realizada na área de assimetrias na pós-graduação e a utilização de aprendizado de máquina. O objetivo deste estudo foi apresentar uma visão geral das publicações disponíveis pertinentes à pós-graduação, assimetrias e o aprendizado de máquina. Seguindo o método de estudo de mapeamento sistemático apresentado na Fig. 1, o estudo foi guiado por um conjunto de questões de mapeamento. A Tabela I apresenta 6 questões de mapeamento (QMs) e a justificativa para a realização deste estudo. A estratégia de busca do estudo, bem como os critérios de inclusão e exclusão, foram baseados nessas 6 QMs.

Depois será apresentado na Fig. 2 o fluxograma com um resumo do processo de mapeamento sistemático e os respectivos trabalhos correlatos selecionados. Os trabalhos foram organizados por área correlata, tipo de estudo realizado e uma descrição do propósito do artigo. As áreas correlatas conforme apresentado na Fig. 2 foram classificadas em: IA Ética, Impacto da Pesquisa, Expansão, Mineração de Dados e Aprendizado de Máquina. Quanto ao tipo de estudo, os trabalhos correlatos foram classificados em: Análise, Modelo, Processo, Programas e Método. A Tabela II apresenta a lista de trabalhos correlatos por autoria.

III. MODELAGEM

Este capítulo descreve a principal contribuição desta pesquisa, o **ADA-PG** - um META modelo de **A**nálise **D**e **A**ssimetrias para a **P**ós-**G**raduação. Este modelo tem por objetivo analisar assimetrias na pós-graduação, como a distribuição de programas e alunos, bem como sua taxa de crescimento, entre outros indicadores, visando conhecer as assimetrias existentes na Educação no país.

O META-Modelo **ADA-PG** apresentado na Fig. 3 possui 4 modelos integrantes: (i) inteligência artificial ética, para endereçar inconvenientes e ameaças que o uso de novas tecnologias pode (potencialmente) introduzir, definindo princípios para a utilização de IA; (ii) análise de dados abertos, para realizar a preparação dos dados sobre as assimetrias nos programas considerando os dados abertos a partir de 2011; (iii) expansão da pós-graduação e impacto, com o intuito de medir o resultado dos programas de expansão, e analisar a evolução dos programas por mesorregiões e também os tópicos de pesquisa a partir da análise de produções científicas; e, (iv) aprendizado de máquina, para aplicação de inteligência artificial para análise dos dados abertos CAPES.

A. O modelo integrante para IA ética

O primeiro passo em um modelo prático é identificar o *framework*. No modelo proposto, um dos melhores *framework* é o IEEE Ethically Aligned Design, porque a iniciativa identifica necessidades e cria consenso para padrões, certificações e códigos de conduta relacionados à implementação ética de IA. A parte 1 deste modelo integrante refere-se a definição das métricas e três foram selecionadas inspiradas no *framework* do IEEE: (i) transparência, é um princípio parte da diretriz 5 no documento IEEE e define que a base de uma decisão particular do Sistema Autônomo e Inteligente (SA/I) deve ser sempre detectável; (ii) responsabilidade, faz parte da diretriz 6 no documento IEEE e define que o SA/I deve ser criado e operado para fornecer uma justificativa inequívoca para todas as decisões tomadas; e, (iii) explicabilidade, baseada na seção 2 do IEEE Normas de Implementação em Sistemas Autônomos e Inteligentes [31], propiciará que uma comunidade entenda, preveja e modifique o SA/I, porque as normas embutidas em SA/I são continuamente atualizadas e refinadas.

A parte 2 deste modelo integrante refere-se a definição de mecanismos para uso da IA. Os mecanismos propostos são uma combinação de itens essenciais ao se aplicar a IA na educação combinando estudos do Instituto de IA Ética na Educação da Universidade de Buckingham [10] e a Fundação Nacional para a Ciência, Tecnologia e Artes do Reino Unido (NESTA - National Endowment for Science, Technology and the Arts) [32].

Sete mecanismos foram definidos: (i) requalificação de pessoal, é um mecanismo que aborda o treinamento para os educadores e usuários de um sistema de IA, para entendimento sobre a tecnologia, e os resultados que serão observados; (ii) entender o contexto, visa a simplicidade da ferramenta e entendimento sobre o seu uso, e evita que os usuários da linha de frente que tem pouco tempo, tenham um desgaste mental ou resistam à ferramenta devido à complexidade; (iii) demonstrar benefícios requer esclarecer sobre as fontes de dados utilizadas, demonstrar o valor da ferramenta, e coletar *feedbacks* para construir confiança pois os usuários relutam em usar ferramentas que eles não entendem; (iv) demonstrar o valor da ferramenta, requer explicar o que são ferramentas de análise preditiva e como elas oferecem melhor suporte à tomada de decisão; (v) gerir a mudança ao introduzir uma nova tecnologia, é um mecanismo para encorajar adaptação dos usuários, e a habilidade no uso da ferramenta e evitar que pareça mais fácil *'deixar isso para os especialistas'*; (vi) permitir *feedback* em tempo real, é um mecanismo que permite coletar dados sobre quando os usuários discordam das informações fornecidas pela ferramenta e utilizar cenários de testes que ajudem aos usuários sobre como navegar casos complexos; (vii) encorajar o uso da ferramenta, é sobre o equilíbrio delicado entre desencorajar complacência do praticante, mas não a ponto de os praticantes ficarem nervosos para usar a ferramenta.

B. O modelo integrante para análise de dados abertos

Neste modelo integrante o primeiro passo é determinar os tipos de dados que serão coletados. Para este estudo foram

TABELA I: QUESTÕES MAPEADAS

ID	Questões	Justificativas
QM1	Quais são os principais canais de publicação para avaliação da pós-graduação ?	Identificar onde podem ser encontradas pesquisas sobre avaliação da pós-graduação e a publicação de estudos futuros.
QM2	Quais são os principais canais de publicação sobre o aprendizado de máquina (*machine learning* - ML) e sua aplicação na educação ?	Identificar onde podem ser encontradas pesquisas sobre os tipos de ML, aplicação na educação bem como estudos futuros.
QM3	Como é a pós-graduação no Brasil ?	Descobrir o processo atual e futuro sobre a pós-graduação reportados na literatura existente.
QM4	O que são as assimetrias na pós-graduação ?	Descobrir pesquisas sobre assimetrias e quais são os estudos futuros.
QM5	A inteligência artificial pode ser justa na educação?	Para descobrir os documentos mais importantes, transformadores e relevantes sobre a IA ética na literatura existente sobre educação.
QM6	Quais são as técnicas de ML aplicadas nos trabalhos selecionados?	Descobrir as técnicas de ML mais comuns aplicadas nos trabalhos selecionados no mapeamento sistemático.

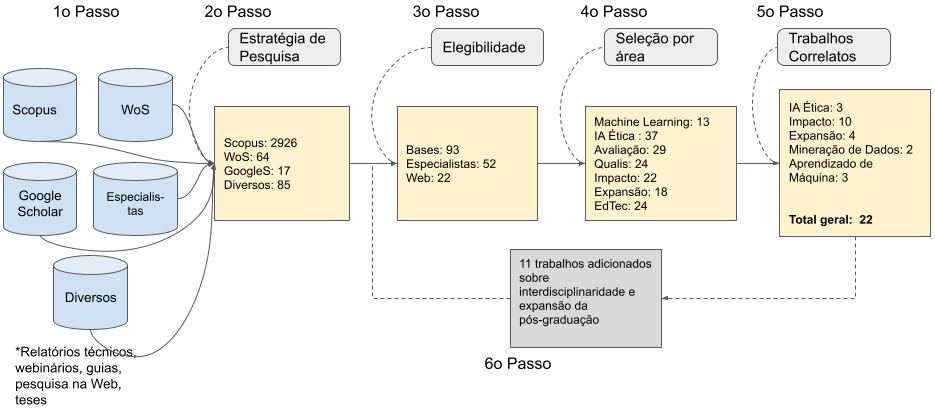


Fig. 2: Fluxograma com o processo de mapeamento sistemático e os respectivos trabalhos correlatos selecionados.

TABELA II: CRONOLOGIA DOS TRABALHOS CORRELATOS SELECIONADOS

Autores	Propósito	Tipo	Área Correlata
Acemoglu e Paschoal (2020) [9]	Tipos de IA e implicações éticas	Análise	IA Ética
Universidade de Buckingham (2020) [10], Mittelstadt (2019) [11]	Modelos para IA Ética	Modelo	IA Ética
CAPES (2020) [12], FAPESP (2020) [13], WRIGHT et al. (2020) [14], Harzing (2019) [15], THOMAZ et al. (2011) [16]	Citações para medir relevância	Análise	Impacto da Pesquisa
Viggiani et al (2020) [17], Barradas (2017) [18], SOMA et al. (2016) [19]	Classificação de Periódicos	Processo	Impacto da Pesquisa
Plos (2018) [20]	Pré-print	Processo	Impacto da Pesquisa
GRUDNIEWICZ et al. (2019) [21]	Periódicos predadores	Análise	Impacto da Pesquisa
PAULA et al. (2020) [22], GONZAGA et al. (2020) [23], MOURA et al. (2019) [24], DALMARCO et al. (2018) [25]	Análise de desempenho	Reuni, CsF, ProEB	Expansão
Romero, C (2020) [26], Luo, Qi (2008) [27]	Mineração de Dados na Educação	Processo	Mineração de Dados
Bringsjord et al. (2020) [28], RUSSEL et al.(2013) [29], WOLFGANG (2011) [30]	Tipos de Machine Learning	Método	Aprendizado de Máquina

considerados inicialmente os dados do portal de dados abertos do governo brasileiro e aos portais das bases comerciais sobre produções científicas. A Fig. 4 apresenta as principais etapas deste modelo integrante. O segundo passo é projetar o banco de dados. Os dados brutos sobre a pós-graduação vêm em uma variedade de formatos (como exportações MySQL, extrações JSON e arquivos CSV) dependendo das fontes e representam diferentes aspectos da pós-graduação e suas assimetrias. Esses dados serão descompactados em diversas tabelas organizadas de acordo com sete domínios de análise: georreferenciamento, mesorregiões, programas, discentes, áreas de conhecimento, expansão da pós-graduação e impacto da pesquisa. O terceiro passo é a harmonização dos dados, ou seja, transformar os dados originais em tamanho e forma gerenciáveis, a fim de executar análises subsequentes, como por exemplo criar novas tabelas para mapeamento a partir de atributos existentes. Algumas junções de dados envolvem mapear a mesorregião e crescimento da população a partir da correlação entre os dados de cidade, localidade e georreferenciamento, e, a agregação dos dados também pode requerer a criação de novas tabelas para a contagem de discentes de mestrado e doutorado, contagem de programas de mestrado e doutorado, distribuição

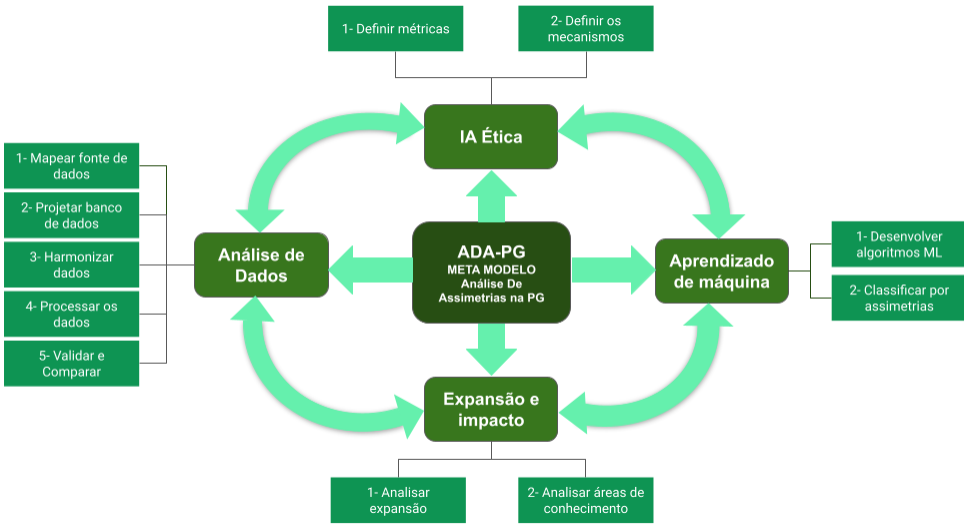


Fig. 3: O META modelo proposto para a análise de assimetrias na Pós-Graduação.

de discentes e programas por áreas, estados, mesorregião, municípios, entre outros. O quarto passo envolve a utilização de técnicas de processamento de dados e aprendizado de máquina para classificação dos programas, sua distribuição e crescimento. Com relação à produção científica, será utilizado mineração de texto e processamento de linguagem natural para determinar os vocabulários utilizados em um campo científico e portanto posteriormente apoiar a identificação da multidisciplinaridade e impacto dos programas.

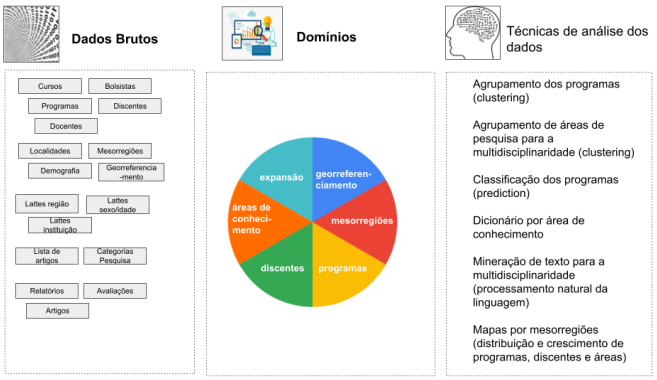


Fig. 4: Visão do processamento de dados para assimetrias na pós-graduação

O quinto passo para a validação do modelo envolve utilizar a matriz de confusão, que é uma tabela que indica os erros e acertos do modelo, comparando com o resultado esperado. A Tabela III demonstra um exemplo de uma matriz de confusão.

TABELA III: EXEMPLO DE MATRIZ DE CONFUSÃO.

Tipos Real/Detectado	Sim (Detectado)	Não (Detectado)
Sim (Real)	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Não (Real)	Falso Positivo (FP)	Verdadeiro Negativo (VN)

No modelo integrante para a IA ética foram apresentadas as métricas que representam os valores fundamentais do modelo e, como estas métricas se relacionam com a diretriz IEEE. Aqui no modelo integrante para análise de dados abertos serão aplicadas métricas complementares referentes à recuperação da informação. Dentre elas, a precisão (*precision*), a sensibilidade (*recall*) e a exatidão (*accuracy*). Matematicamente, a taxa de precisão está apresentada na Eq. III-B (1), a sensibilidade está apresentada na Eq. III-B (2), e a exatidão está apresentada na Eq. III-B (3) [33].

$$Precisão: \quad P = \frac{VP}{(VP+FP)} \times 100\% \tag{1}$$

$$Sensibilidade: \quad S = \frac{VP}{(VP+FN)} \times 100\% \tag{2}$$

$$Exatidão: \quad E = \frac{VP+VN}{(VP+FP+VN+FN)} \times 100\% \tag{3}$$

Onde "VP", "FP", "FN" e "VN" indicam "verdadeiro positivo", "falso positivo", "falso negativo" e "verdadeiro negativo" na recuperação das informações, respectivamente.

C. O modelo integrante para expansão e impacto da pós-graduação

Para a análise da expansão da pós-graduação aplicando o processo apresentado na Fig. 5a foram selecionadas 3 fontes de dados: GeoCapes Indicador de distribuição de discentes da pós-graduação no Brasil; IBGE Estimativas da População; IBGE Divisão Territorial Brasileira. Os dados brutos sobre a pós-graduação vêm em uma variedade de formatos (como exportações MySQL, extrações JSON e arquivos CSV) dependendo das fontes. E esses dados foram descompactados em diversas tabelas organizadas de acordo com sete domínios de análise: georreferenciamento, mesorregiões, programas, discentes, áreas de conhecimento e expansão da pós-graduação.

Novas tabelas foram criadas a partir de dados existentes como o mapeamento da mesorregião e crescimento da população a partir da correlação entre os dados de cidade, localidade, georreferenciamento e mesorregião. Os dados foram

também agregados por discentes de mestrado e doutorado, programas de mestrado e doutorado, programas por áreas de avaliação, entre outros. Foi realizada a validação dos mapas comparando-se os dados entre 2018 e 2019. Observou-se uma pequena variação no crescimento de titulados e novas mesorregiões atendidas pela pós-graduação. Embora, uma possível causa para a expansão dos programas acadêmicos seja provavelmente o Reuni, e para os programas profissionalizantes provavelmente ocorra um contribuição do ProEB, as universidades privadas também podem ter aumentado as vagas.

A segunda parte deste modelo integrante refere-se a analisar as áreas de conhecimento devido ao seu impacto no desenvolvimento da pós-graduação. A CAPES busca melhorar sua árvore do conhecimento para corresponder à realidade das transformações do conhecimento do mundo e do País, daí a necessidade de se identificar lacunas em áreas novas, emergentes, interdisciplinares e multidisciplinares. Entretanto, isto deve levar em conta a necessidade de se preservar algumas áreas tradicionais que não devem ser desativadas e também ter a preocupação com o desenvolvimento de áreas de ponta em que o Brasil se encontra particularmente atrasado [34].

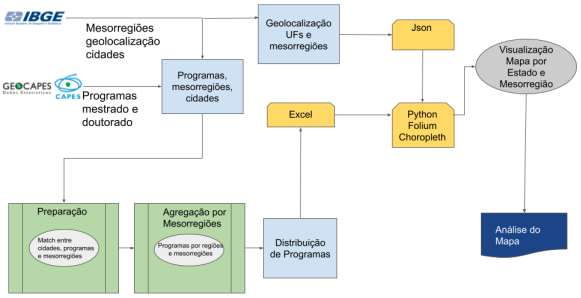
Adicionalmente, a aplicação do processamento de linguagem natural ajudará a agregar as publicações por conceitos científicos (por exemplo, “inteligência artificial") ou até mesmo aplicações diferentes do mesmo conceito (por exemplo, “inteligência artificial ética" e “inteligência artificial geral"). Essas análises podem exigir substancialmente mais dados e recursos computacionais, mas poderiam ajudar a diferenciar entre ideias modistas e verdadeiramente revolucionárias ou revelar ideias promissoras, joias escondidas e pouco apreciadas [35].

Neste estudo, o algoritmo utilizado para processamento de linguagem natural é o LDA (abreviação de Latent Dirichlet Allocation), um modelo de aprendizado de máquina não supervisionado que recebe documentos como entrada e encontra tópicos como saída [36]. Dito isto, o método proposto na Fig. 5b identificará os tópicos de pesquisa utilizando dados sobre os artigos produzidos pelos programas e utilizando o algoritmo LDA para identificar a partir dos artigos as áreas investigadas em uma publicação através da identificação de tópicos a partir do texto destes artigos científicos.

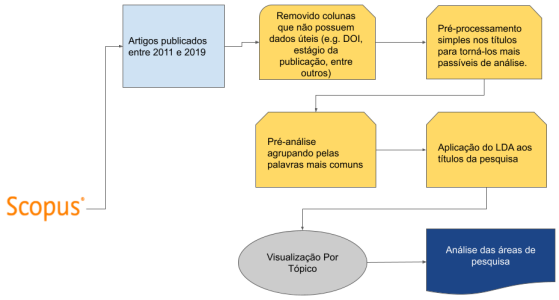
D. O modelo integrante para aprendizado de máquina

Os algoritmos de aprendizado de máquina serão utilizados neste modelo integrante, conforme apresentado na Fig. 4. O processamento de dados quantitativos envolve a utilização de algoritmos supervisionados e/ou não supervisionados, para a identificação das assimetrias pré-definidas bem como para as assimetrias ainda não identificadas e poderão ser conhecidas após a aplicação do algoritmo.

O processamento de dados qualitativos envolve a utilização de aprendizado de máquina para avaliar a abrangência da pesquisa no Brasil e suas fronteiras. Neste caso serão utilizados algoritmos não supervisionados para o agrupamento das áreas de conhecimento, e algoritmos supervisionados em conjunto com processamento de linguagem natural para identificar as áreas de conhecimento a partir do resumo dos artigos.



(a) Modelo de mapeamento da expansão



(b) Método de mapeamento dos tópicos de pesquisa

Fig. 5: O modelo integrante para expansão e impacto da pós-graduação.

Dentre as técnicas para análise dos dados foram identificadas: (i) agrupamento dos programas (*clustering*); (ii) agrupamento de áreas de pesquisa (*clustering*); (iii) classificação dos programas (*prediction*); dicionário por área de conhecimento; (iv) mineração de texto para a multidisciplinaridade (processamento de linguagem natural); e, (v) mapas por mesorregiões (distribuição e crescimento de programas, discentes e áreas).

Na fase seguinte deste projeto serão aplicados os algoritmos de aprendizado de máquina e de processamento de linguagem natural. A análise dos diferentes tipos de assimetrias precisa ser realizada cuidadosamente antes de tirar conclusões e divulgar informações a terceiros, pois a análise e interpretação dos dados possui natureza subjetiva, portanto será necessário organizar os temas que se revelam e os padrões identificados.

IV. Visualização

Esta seção descreve os critérios utilizados para a visualização de dados, dentre eles, os tipos de programas, as mesorregiões, as escalas dos mapas e os tópicos de pesquisa.

A. Tipos de Programas

Os cursos de pós-graduação são organizados em Lato Sensu e Stricto Sensu. Lato Sensu é destinado a desenvolver o conhecimento que pode ser aplicado no dia a dia profissional. Não busca conhecer em detalhes as teorias por trás do conhecimento e enfoca nas técnicas, modelos e metodologias aplicadas. Constituem-se de especializações ou mestrados profissionalizantes. Já os programas Stricto Sensu abordam o conhecimento de uma forma mais exploratória com foco na

sua contribuição científica e acadêmica [37] [38]. Constituem-se dos programas de mestrado e doutorado acadêmicos. Os programas de mestrado Stricto Sensu tem duração de entre dois e dois anos e meio, enquanto no doutorado a duração é de quatro anos. Em ambos os profissionais passam por uma banca avaliadora, e apresentam uma dissertação ou tese, respectivamente. Ao final, se aprovado é concedido um diploma de grau acadêmico [37] [38]. À época em 2019, havia um total de 2535 programas de doutorado acadêmico, 3732 programas de mestrado acadêmico e 1256 programas de mestrado profissionalizante com matriculados somadas todas as IES (alguns programas em rede ou associação são oferecidos em mais de uma instituição) [39].

B. As mesorregiões

De acordo com o Instituto Brasileiro de Geografia e Estatística (IBGE) [40], a Divisão Regional do Brasil consiste no agrupamento de Estados e Municípios em regiões com a finalidade de atualizar o conhecimento regional do País e viabilizar a definição de uma base territorial para fins de levantamento e divulgação de dados estatísticos. Adicionalmente a compreensão da organização do território nacional visa assistir o governo federal, bem como Estados e Municípios, na implantação e gestão de políticas públicas e investimentos.

A divisão em macrorregiões foi elaborada em 1970, e resultaram nas seguintes denominações: Região Norte, Região Nordeste, Região Sudeste, Região Sul e Região Centro-Oeste, que permanecem em vigor até o momento atual. Foram criadas também as mesorregiões que congregam diversos municípios de uma área geográfica de um estado brasileiro com similaridades econômicas e sociais, dividindo-se posteriormente em microrregiões compostas de municípios limítrofes com organização espacial em comum e específica [41].

Para este estudo foram utilizados os dados da divisão territorial do Brasil a partir dos dados IBGE de 2019 [42]. À época em 2019, havia um total de 137 mesorregiões, dividindo o território das unidades federativas brasileiras e agrupando-os em 5.570 municípios. No federalismo brasileiro, o Distrito Federal não é composto por nenhum município, diferente dos estados, e as trinta e uma regiões administrativas em que está dividido não têm correspondência com municípios. Optou-se por mapas por mesorregião pois propiciam de forma simples a visualização quantitativa da expansão da pós-graduação, bem como o ritmo como a expansão ocorreu. Adicionalmente a utilização da mesorregião propicia captar contrastes e diferenciações por região e poderá auxiliar nas avaliações e orientar políticas de fomento.

C. Escalas dos Mapas

Para observar a distribuição de titulados foram agregados os dados de titulados por mesorregião, realizando a soma de discentes titulados nas cidades da mesorregião que possuem programas de pós-graduação. Para observar o ritmo de crescimento de titulados no país nos programas de doutorado e mestrado acadêmico e nos programas de mestrado profissional, foi realizado o cálculo demonstrado a seguir utilizando dados de 2011 e 2019, 9 anos, conforme apresentado na Eq. IV-C. Este trabalho focou na análise de titulados pois trata-se de um indicador utilizado no Plano Nacional de Educação do Ministério da Educação (PNE) visando avaliar a ampliação da formação no pais. Conforme descrito pelo Ministério da Educação, o Plano Nacional de Educação (PNE) determina diretrizes, metas e estratégias para a política educacional no período de 2014 a 2024 [43].

$$taxa\_de\_crescimento\_ao\_ano(\%) = \left(\frac{discentes\_por\_habitantes\_em\_2019}{discentes\_por\_habitantes\_em\_2011}\right)^{\frac{1}{9}} \quad (4)$$

O cálculo foi realizado considerando o aumento da população na mesorregião. Para o cálculo do número de habitantes foram somados os habitantes de todas as cidades que compõem uma mesorregião. Os mapas preparados nesta análise possuem escalas específicas. Por exemplo, para o mapa quantitativo de matriculados, devido ao número de matriculados ser maior em relação ao de titulados, a escala de matriculados é de 1 a 16000 discentes, sendo que a maioria das mesorregiões possuem até 1000 matriculados e depois temos a escala aumentando a cada 1500 até 7000, representando as demais regiões.

Para o mapa quantitativo de titulados, para uma melhor visualização da interiorização da pós-graduação, optou-se por uma escala de 1 a 100 para representatividade das regiões menores e depois temos a escala aumentando a cada 200 até 1200, representando as variações nas quantidades de titulados por mesorregião. As regiões metropolitanas foram representadas em duas cores, a cor azul para representar as mesorregiões com até 3000 titulados, e a cor verde claro para representar as mesorregiões com até 6000 titulados.

Para o mapa de crescimento de titulados, para uma melhor visualização de como foi o crescimento e a interiorização da pós-graduação, optou-se por uma escala de –10% até 50%. As regiões com –10% a 10% representam uma estagnação em relação a crescimento de titulados, pois para o cálculo foi utilizado o número de titulados por habitantes na mesorregião. Portanto em alguns casos, embora o número de titulados tenha aumentado, houve também o aumento da população, o que poderá ser uma das causas para a representação da região como em estagnação. As regiões acima de 50% representam uma expansão que poderia ser considerada próxima a criação de um novo programa. As regiões com programas novos foram representadas em uma cor específica, verde, pois não é possível calcular o crescimento uma vez que estes programas não existiam nestas mesorregiões em 2011. Em todos os mapas as mesorregiões sem programas estão representadas em cinza.

D. Os Tópicos de Pesquisa

Com relação a aplicação de um algoritmo de processamento de linguagem natural, neste trabalho será utilizado o algoritmo LDA (abreviação de Latent Dirichlet Allocation), que é um modelo de aprendizado de máquina não supervisionado que recebe documentos como entrada e encontra tópicos como saída. O LDA reconcilia a ambiguidade ao representar o artigo como uma distribuição sobre tópicos, e cada tópico como uma distribuição sobre palavras. Portanto, neste trabalho será observado os grupos de palavras em um determinado tópico

(clustering). É recomendado para trabalhos futuros mapear estes tópicos em um conjunto de rótulos que poderiam ser aplicados em um modelo de aprendizado de máquina supervisionado. Porém é importante considerar que rotular estes tópicos requer considerar o significado humano e pertinente à pesquisa científica em diferentes áreas. Assim as métricas apresentadas anteriormente referentes a matriz de confusão são recomendadas para trabalhos futuros que possam utilizar os tópicos mapeados em rótulos e de forma supervisionada.

Existem 3 parâmetros principais do modelo LDA: (i) o número de tópicos; (ii) o número de palavras por tópico; e, (iii) o número de tópicos por documento. Para a análise dos artigos, há interesse apenas nos dados de texto associados ao artigo, bem como no ano em que o artigo foi publicado e foi necessário remover todas as colunas que não contêm informações de texto úteis, por exemplo, DOI, afiliações entre outros. Em seguida foi verificado o número de publicações por ano, para entender a extensão da 'revolução' na área. Depois, foi realizado um pre-processamento do texto. O objetivo agora foi analisar os títulos dos diferentes artigos para identificar tendências de pesquisa em uma área, por exemplo, a engenharia. Foi realizado também um pré-processamento simples nos títulos para torná-los mais passíveis de análise, por exemplo remover qualquer pontuação no título e em seguida, converter os títulos em letras minúsculas. Pode-se fazer uma nuvem de palavras dos títulos dos trabalhos de pesquisa, isso dará uma representação visual das palavras mais comuns. A visualização é fundamental para entender os tópicos preliminares e permite verificar se será necessário um pré-processamento adicional.

Como próximo passo, foi realizado a aplicação do algoritmo LDA. O LDA é capaz de realizar a detecção de tópicos em grandes conjuntos de documentos, determinando quais são os principais 'tópicos' em um grande conjunto de textos não rotulados. Porém o LDA não funciona diretamente em dados de texto. Primeiro, é necessário converter os documentos em uma representação vetorial simples. Esta representação será então utilizada pelo LDA para determinar os tópicos. Cada entrada de um 'vetor de documento' corresponderá ao número de vezes que uma palavra ocorreu no documento. Em resumo, a lista de títulos é convertida em uma lista de vetores, todos com comprimento igual ao vocabulário. Pode-se então plotar as 10 palavras mais comuns com base no resultado dessa operação (a lista de vetores de documentos). Na seção V será demonstrado o resultado de aplicação do algoritmos nos dados dos artigos publicados pela escola de Engenharia do Instituto Tecnológico de Aeronáutica. É importante observar que este mesmo fluxo de processamento de texto, poderá ser utilizado para processar um novo conjunto de documentos de outras escolas, como a medicina, a física, entre outras, pois o fluxo construído pode ser facilmente aplicado para um novo conjunto de dados de texto.

V. Reporte de Estimativas

Com a expansão da pós-graduação houve um aumento significativo de alunos matriculados e titulados por mesorregiões brasileiras em seus programas de doutorado acadêmico, mestrado acadêmico e, mestrado profissional conforme apresentado nos mapas gerados nesta análise. A seguir teremos os exemplos de leitura de mapas para a distribuição de discentes e o ritmo de crescimento de titulados. Serão apresentados os resultados da aplicação do algoritmo de aprendizado de máquina LDA para a analisar tópicos de pesquisa a partir de artigos científicos.

A. Distribuição de discentes

A Fig. 6 apresenta a distribuição de discentes titulados por mesorregião e por tipo de programa nos anos de 2011 e 2019. A Fig. 6a e a Fig. 6d, apresenta a distribuição de alunos de doutorado titulados e sua evolução entre 2011 e 2019 respectivamente. Como um exemplo de leitura dos mapas, há uma expansão no número de doutores titulados em todas as regiões com uma expansão maior nas mesorregiões no interior do Sul e Sudeste. Em um outro exemplo de leitura dos mapas, é verificado que este crescimento também ocorre no mestrado acadêmico, porém com um cobertura ainda maior em todo o território nacional no número de titulados. Com uma expansão expressiva no interior. Apenas poucas mesorregiões não possuem titulados no mestrado acadêmico conforme apresentado na Fig. 6b para os dados de 2011 e Fig. 6e para os dados de 2019.

Com relação ao mestrado profissional em uma leitura compreensiva dos mapas, observa-se que em 2011, apenas algumas mesorregiões localizadas no Sudeste e no Sul do país possuíam titulados no mestrado profissional. Em 2019 observa-se também uma expansão significativa em todas as regiões do país, com exceção do Norte do país, onde algumas mesorregiões não possuem titulados no mestrado profissional conforme apresentado na Fig. 6c para os dados de 2011 e Fig. 6f para os dados de 2019.

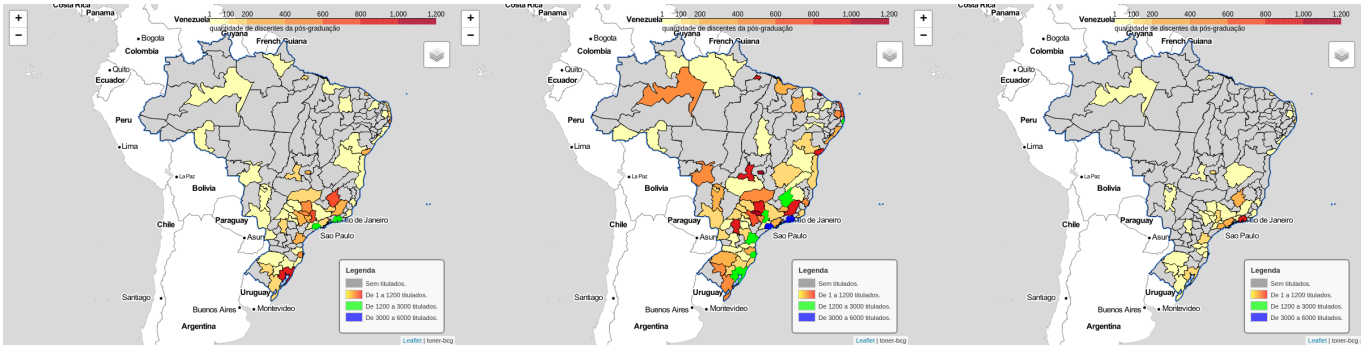
B. Ritmo de crescimento de titulados

Para observar o ritmo de crescimento de titulados no país nos programas de doutorado e mestrado acadêmico e nos programas de mestrado profissional, foi realizado o cálculo demonstrado na equação IV-C e mapeado os resultados apresentados na figura 7.

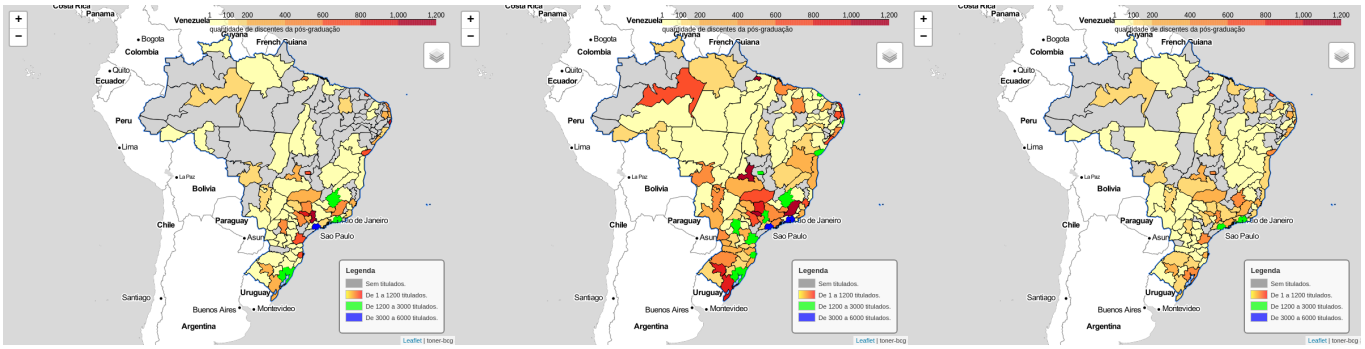
As mesorregiões apresentadas em verde não possuíam programas em 2011. Apresentadas em tonalidades de azul estão as mesorregiões que tiveram expansão no número de titulados, e em tonalidades azul claro as regiões que tiveram uma pequena variação no número de titulados. Em vermelho as regiões que não tiveram titulados. Em roxo as regiões que tiveram taxa de crescimento acima de 50%.Observa-se, como um exemplo de leitura do mapa na Fig. 7a, que com relação ao doutorado, Sul e Sudeste do país passam a ter titulados na grande maioria de suas mesorregiões, devido ao aumento de vagas e também à criação de programas. Regiões do Centro-Oeste, Nordeste e Norte, também tiveram a criação de programas de doutorado iniciando uma expansão da pós-graduação para o interior destas regiões. Outro exemplo de leitura dos mapas é o crescimento do mestrado acadêmico apresentado na Fig. 7b, onde observa-se uma expansão significativa devido a criação de novos programas e na maioria das mesorregiões das regiões

Fig. 6: Distribuição de discentes titulados por mesorregião em 2011 e 2019

(a) Doutorado Acadêmico em 2011 (b) Mestrado Acadêmico em 2011 (c) Mestrado Profissional em 2011



(d) Doutorado Acadêmico em 2019 (e) Mestrado Acadêmico em 2019 (f) Mestrado Profissional em 2019



Norte, Nordeste e Centro-Oeste. As regiões Sul e Sudeste mantém o número de titulados com uma pequena variação em algumas mesorregiões, e com um aumento significativo em regiões do interior, e como resultado temos titulados em quase todas as suas mesorregiões. Com relação ao mestrado profissional observa-se em um exemplo de leitura do mapa apresentado na Fig. 7c que houve uma expansão significativa devido a criação de programas em mesorregiões no interior do país, porém há também um aumento significativo de mestres profissionais nas regiões Sul e Sudeste. A expansão propicia a disponibilização de mestres profissionais nas principais regiões do país.

C. Tópicos de Pesquisa

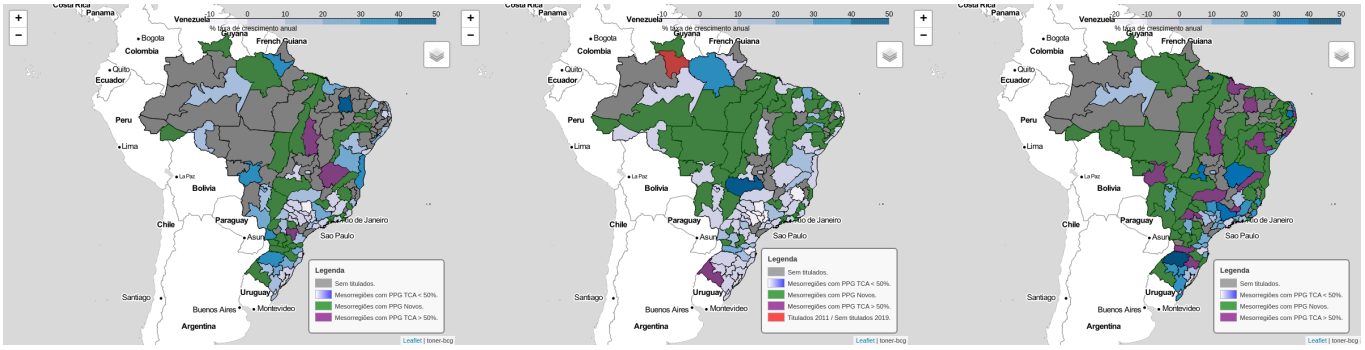
Para mapeamento das áreas de conhecimento, foi aplicado o processamento de linguagem natural para identificação de tópicos de pesquisa, utilizando como uma amostra os artigos publicados na Scopus entre 2011 e 2019 pelo Instituto Tecnológico de Aeronáutica (ITA), apenas para a área de engenharia. Foram analisados o crescimento no número de publicações entre 2011 e 2019, as principais revistas, os principais termos de pesquisa, e para concluir os principais tópicos de pesquisa, agrupados nas 10 principais palavras encontradas nos títulos das publicações.

Ao analisar o número de publicações é possível identificar um grande aumento no número de publicações pelo ITA na Scopus, de pouco mais de 100 artigos em 2011 para mais de 300 artigos em 2019, conforme apresentado na Figura 8. Na Tabela IV são apresentadas as dez principais revistas em que foram publicados os artigos do ITA entre 2011 e 2019. Antes do processamento dos tópicos de pesquisa a partir dos artigos publicados, foram verificadas as principais palavras mencionadas nos artigos conforme apresentado no gráfico com as 10 principais palavras apresentado na Figura 9. E finalmente foi aplicado o algoritmo LDA ao título dos artigos publicados pelo ITA na Scopus no período de 2011 e 2019, sendo identificados 10 tópicos, como os mais relevantes para a pesquisa na instituição, listados a seguir.

- *Topic #0: control based design using model flight analysis management implementation development*
- *Topic #1: using composite study analysis jet control turbulent design optimization plasma*
- *Topic #2: structures model systems coherent shape control based wave edge review*
- *Topic #3: using based aircraft control design networks systems analysis prediction composite*
- *Topic #4: films study high earth development wind characterization properties 3d influence*
- *Topic #5: analysis based microstrip using cylindrical model systems antennas buckling technique*
- *Topic #6: using model time aircraft method transfer based simulation heat systems*
- *Topic #7: control analysis aircraft flexible stability dynamics experimental based using measurements*
- *Topic #8: state using asynchronous model non hybrid design modeling approach mode*
- *Topic #9: gas performance turbine flows approach based evaluation pressure high design*

Fig. 7: Crescimento de discentes titulados por mesorregião em 2011 e 2019

(a) Doutorado Acadêmico (b) Mestrado Acadêmico (c) Mestrado Profissional



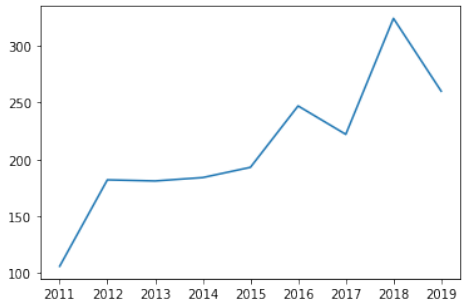


Fig. 8: Publicações do ITA na Scopus entre 2011 e 2019

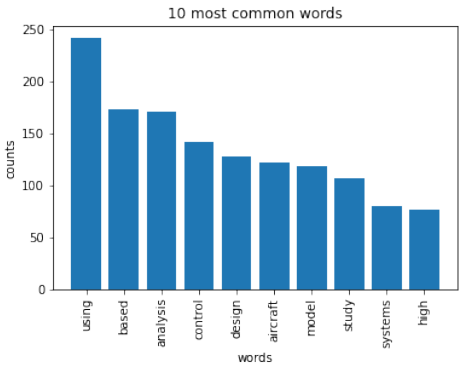


Fig. 9: Principais Termos das Publicações do ITA na Scopus entre 2011 e 2019

TABELA IV: As Dez Principais Revistas em que o ITA Publicou Artigos entre 2011 e 2019.

Nome da Revista	Quantidade de Artigos
10th Annual International Systems Conference, SysCon 2016 - Proceedings	3
11th AIAA Aviation Technology, Integration,and Operations (ATIO) Conference, including the AIAA Balloon Systems Conference and 19th AIAA Lighter-Than-Air Technology Conference	1
11th Annual IEEE International Systems Conference, SysCon 2017 - Proceedings	2
11th International Symposium on Turbulence and Shear Flow Phenomena, TSFP 2019	1
11th World Congress on Computational Mechanics, WCCM 2014, 5th European Conference on Computational Mechanics, ECCM 2014 and 6th European Conference on Computational Fluid Dynamics, ECFD 2014	1
WSEAS Transactions on Fluid Mechanics	7
Wear	1
Welding International	1
Wireless Networks	1
XXI IMEKO World Congress "Measurement in Research and Industry"	3

VI. Conclusões

Este trabalho visou contribuir com conhecimentos e método para analisar a situação da pós-graduação no país, entre 2011 e 2019, e o impacto dos principais programas para expansão da pós-graduação desde os anos 2000, como Reuni, CsF e ProEB. Adicionalmente, este trabalho atingiu o objetivo através da utilização de um modelo proposto na seção 2, Métodos, proporcionando resultados preliminares para a visualização da distribuição de programas bem como sua taxa de crescimento com informações baseadas em dados abertos de diversas fontes governamentais confiáveis como a CAPES e o IBGE. Diferentemente de outros métodos, o modelo proposto utiliza-se de uma visão detalhada por mesorregiões, assim proporcionando informações mais atualizadas e com maior completude sobre os programas de pós-graduação em território nacional, constituindo-se na primeira contribuição adicional. Esta característica também permite a identificar expansões e criação de novos programas em novas mesorregiões. Através dos exemplos de leitura dos mapas, observa-se que de maneira geral há uma expansão na região Sudeste e Sul e na ordem seguinte, Nordeste, Centro-Oeste e Norte. Adicionalmente, observa-se uma expansão significativa nas regiões Nordeste, Centro-Oeste e Norte, e uma possível causa é o programa Reuni. A expansão do mestrado profissional segue a mesma tendência dos programas de doutorado e mestrado acadêmico, com um possível contribuição do programas profissionalizantes da educação básica do ProEB. Além das instituições de ensino público, as instituições de ensino privado também poderão ter tido uma expansão.

O Modelo pôde ser aplicado com sucesso conforme demonstrado nos resultados apresentados, a partir de dados abertos de órgãos governamentais, devido ao modelo se basear em dados

sobre a pós-graduação constituído de informações amplas, de livre acesso e de compartilhamentos entre entidades. Tais informações formam uma base aumentada, complementando as ferramentas e os trabalhos correlatos, e portanto, possuindo maior capacidade e generalidade. Portanto, podendo ser aplicado em outros contextos de análise sobre a pós-graduação. Durante os testes práticos desta pesquisa, notou-se a possibilidade de refinamento de algumas investigações adicionais nas áreas de desenvolvimento da pós-graduação, como por exemplo incluir novas fontes de dados para incluir novos indicadores como a relação entre o crescimento da graduação e da pós-graduação e incorporar mudanças geográficas ao longo do tempo.

O estudo demonstra também como aplicar um modelo para visualizar os tópicos envolvidos nas pesquisas científicas publicadas pelas universidades brasileiras entre 2011 e 2019. Como exemplo foi utilizado uma amostra de dados das publicações em engenharia do Instituto Tecnológico de Aeronáutica, e demonstrado como os principais tópicos de pesquisa poderiam ser identificados utilizando processamento de linguagem natural.

AGRADECIMENTOS

Os autores agradecem a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Instituto Tecnológico da Aeronáutica (ITA) pelos suportes financeiros para o desenvolvimento do presente trabalho. E ao Grupo Boticário pelo apoio geral para a conclusão deste estudo.

REFERENCES

[1] CAPES, "Autoavaliação de programas de pós-graduação," in Relatórios Técnicos DAV e Grupos de Trabalho, Fundação CAPES, 2019.
[2] CAPES, "Plano Nacional de Pós-Graduação- PNPG 2011-2020 - volume i," 2010. Acessado por ultimo em 06 de Maio de 2021.
[3] CAPES, "Programa de apoio a planos de reestruturação e expansão das universidades federais - Reuni," 2009. Acessado por ultimo em 04 de Fevereiro de 2021.
[4] CAPES, "O programa ciência sem fronteiras," 2011. Acessado por ultimo em 04 de Fevereiro de 2021.
[5] CAPES, "Capes abriu mais de duas mil vagas no proeb em 2020," 2020. Acessado por ultimo em 15 de Fevereiro de 2021.
[6] NYT, "British grading debacle shows pitfalls of automating government," 2020.
[7] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in 12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12, pp. 1–10, 2008.
[8] M. Karampela, M. Isomursu, T. Porat, C. Maramis, N. Mountford, G. Giunti, I. Chouvarda, and F. Lehocki, "The extent and coverage of current knowledge of connected health: Systematic mapping study," Journal of Medical Internet Research, vol. 21, no. 9, p. e14394, 2019.
[9] D. Acemoglu and P. Restrepo, "The wrong kind of ai? artificial intelligence and the future of labour demand," Cambridge Journal of Regions, Economy and Society, vol. 13, no. 1, pp. 25–35, 2020.
[10] The University of Buckingham, "Developing a shared vision of ethical ai in education: An invitation to participate," 2020.
[11] B. Mittelstadt, "Principles alone cannot guarantee ethical ai," Nature Machine Intelligence, pp. 1–7, 2019.
[12] CAPES, "Impacto e a relevância econômica e social na pós-graduação," 2020. Acessado por ultimo em 14 de Janeiro de 2021.
[13] FAPESP, "Revista fapesp, o medo da indiferença," 2020. Acessado por ultimo em 14 de Janeiro de 2021.
[14] M. Wright, D. J. Ketchen Jr, and T. Clark, How to get published in the best management journals. Edward Elgar Publishing, 2020.
[15] A.-W. Harzing, The publish or perish book. Tarma Software Research Pty Limited, 2010.
[16] P. G. Thomaz, R. S. Assad, and L. F. P. Moreira, "Uso do fator de impacto e do índice h para avaliar pesquisadores e publicações," Arquivos Brasileiros de Cardiologia, vol. 96, no. 2, pp. 90–93, 2011.
[17] E. Viggiani and L. Calabró, "Does faculty disciplinary background play a role in the publication pattern of an interdisciplinary research area? the case of science education in brazil," Scientometrics, vol. 125, no. 2, pp. 893–908, 2020.
[18] R. d. C. B. Barata, "Dez coisas que você deveria saber sobre o qualis," Boletim Técnico do PPEC, vol. 2, no. 1, pp. 17p–17p, 2017.
[19] N. Yoshihiro Soma, A. Donizeti Alves, and H. Hideki Yanasse, "O qualis periódicos e sua utilização nas avaliações.," RBPG. Revista Brasileira de Pós-Graduação, vol. 13, no. 30, 2016.
[20] Plos, "Power to the pre-print," 2018. Acessado por ultimo em 14 de Janeiro de 2021.
[21] A. Grudniewicz, D. Moher, K. D. Cobey, G. L. Bryson, S. Cukier, K. Allen, C. Ardern, L. Balcom, T. Barros, M. Berger, et al., "Predatory journals: no definition, no defence," 2019.
[22] C. H. d. Paula and F. M. d. Almeida, "O programa Reuni e o desempenho das ifes brasileiras," Ensaio: Avaliação e Políticas Públicas em Educação, no. AHEAD, 2020.
[23] G. R. Gonzaga, D. C. d. Paiva, and M. L. Eichler, "Desafios e perspectivas atuais na formação do professor de química: expectativas sobre o mestrado profissional em química em rede nacional (profqui)," Química Nova, vol. 43, no. 4, pp. 493–505, 2020.
[24] A. M. B. Filardi, "Desenvolvimento do Reuni: crítica à sua implantação e sua relação econômica," Linhas críticas, vol. 20, no. 43, pp. 563–582, 2014.
[25] G. Dalmarco, W. Hulsink, and G. V. Blois, "Creating entrepreneurial universities in an emerging economy: Evidence from brazil," Technological Forecasting and Social Change, vol. 135, pp. 99–111, 2018.
[26] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 3, p. e1355, 2020.
[27] Q. Luo, "Advancing knowledge discovery and data mining," in First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008), pp. 3–5, IEEE, 2008.
[28] S. Bringsjord and N. S. Govindarajulu, "Artificial Intelligence," in The Stanford Encyclopedia of Philosophy (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, summer 2020 ed., 2020.
[29] S. Russel, P. Norvig, et al., Artificial intelligence: a modern approach. Pearson Education Limited, 2013.
[30] E. Wolfgang, "Introduction to artificial intelligence," Translated by Nathanael Black With illustrations by Florian Mast, Springer-Verlag London Limited, 2011.
[31] IEEE, "Ieee ethically aligned design," 2020. Acessado por ultimo em 18 de Novembro de 2020.
[32] NESTA, "Decision-making in the age of the algorithm," 2019. Acessado por ultimo em 07 de Janeiro de 2021.
[33] F. Ali, D. Kwak, P. Khan, S. R. Islam, K. H. Kim, and K. S. Kwak, "Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling," Transportation Research Part C: Emerging Technologies, vol. 77, pp. 33–48, 2017.
[34] CAPES, "Pós-graduação: Enfrentando novos desafios documentos e resultados do seminário realizado pela capes," Boletim Informativo Vol. 9, No 2 e Vol. 9, No 3, 2001.
[35] J. W. Weis and J. M. Jacobson, "Learning on knowledge graph dynamics provides an early warning of impactful research," Nature Biotechnology, pp. 1–8, 2021.
[36] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, vol. 3, no. Jan, pp. 993–1022, 2003.
[37] PUCRS, "O que é pós-graduação. conceitos e tipos de especializações," 2021. Acessado por ultimo em 29 de Abril de 2021.
[38] A. Almeida Júnior, N. Sucupira, C. Salgado, J. Barreto Filho, M. R. Silva, D. Trigueiro, A. A. Lima, A. Teixeira, V. Chagas, and R. Maciel, "Parecer cfe nº 977/65, aprovado em 3 dez. 1965," Revista Brasileira de Educação, no. 30, pp. 162–173, 2005.
[39] CAPES, "Dados e estatísticas GeoCapes," 2003. Acessado por ultimo em 04 de Fevereiro de 2021.
[40] IBGE, "Instituto Brasileiro de Geografia e Estatística," 2021. Acessado por ultimo em 29 de Abril de 2021.
[41] IBGE, "Divisão regional do Brasil em mesorregiões e microrregiões geográficas," 2021. Acessado por ultimo em 29 de Abril de 2021.
[42] IBGE, "Divisão territorial brasileira 2019," 2021. Acessado por ultimo em 29 de Abril de 2021.
[43] MEC, "Plano Nacional de Educação," 2014. Acessado por ultimo em 04 de Fevereiro de 2021.