Diversità

UniCV CENTRO UNIVERSITÁRIO CIDADE VERDE

MINERAÇÃO E ANÁLISE DE DADOS ACERCA DA PRODUÇÃO ACADÊMICA NO PERÍODO DE 1987 ATÉ 2018 DOS CURSOS DE PÓS-GRADUAÇÃO NO BRASIL

Thiago Gama
Leonardo Catharin

RESUMO

O presente trabalho apresenta uma abordagem de mineração de dados educacionais para analisar os avanços dos programas de pós-graduação cadastrados na Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), uma fundação vinculada ao Ministério da Educação do Brasil. Para a análise destes avanços aplicou-se métodos estatísticos em um conjunto de dados no intervalo de trinta e dois anos, de 1987 até 2018, obtidos nos dados abertos por meio do Portal Brasileiro de Dados Abertos. Como resultado deste estudo, obteve-se a visualização gráfica de curva de crescimento da produção científica sob diferentes aspectos, sendo eles: região do Brasil, área do conhecimento das produções acadêmicas, instituição de ensino superior, nível da pós-graduação e data da publicação.

Palavras-chave: Ciência de dados. Produção acadêmica. Panorama histórico. Cursos de pós-graduação. Brasil.

ABSTRACT

The present work presents an educational data mining approach to analyze the advances of graduate programs registered at the Coordination for the Improvement of Higher Education Personnel (CAPES), a foundation linked to the Ministry of Education of Brazil. For the analysis of these advances, statistical methods were applied to a set of data in the interval of thirty-two years, from 1987 to 2018, obtained from open data through the Brazilian Open Data Portal. As a result of this study, graphic visualization of the growth curve of scientific production was obtained under different aspects, namely: region of Brazil, area of knowledge of academic productions, higher education institution, postgraduate level and date of publication.

Keywords: Data Science. Academic production. Historical overview. Postgraduate courses. Brazil.

INTRODUÇÃO

A comunidade acadêmica brasileira sempre buscou informações sobre sua própria realidade e os avanços acontecidos em cada região, para assim demonstrar à sociedade os avanços alcançados por suas produções e a importância de investimentos na educação e na ciência. O problema observado na área de estudo encontra-se no fato de não existir nenhum trabalho científico que minere a vasta quantidade de dados relativos à produção de teses e dissertações nos cursos de pós-graduação no Brasil em sua totalidade de anos cobertos pelo órgão governamental de fomento da produção científica nacional desde 1987.

O presente trabalho tem como intuito preencher esta lacuna de estudo identificada, a fim de proporcionar a geração de dados estatísticos para as mais diversas áreas de

conhecimento categorizadas pela CAPES. Habilitando assim a avaliação da evolução, tanto temática quanto quantitativa, dos trabalhos desenvolvidos por pesquisadores brasileiros em suas respectivas instituições de ensino superior (IES), apresentados nos programas de pós-graduação existentes no Brasil, na qual estes alunos estão associados, mapeados nos conjuntos de dados utilizados como entrada para o pretendido estudo.

Este estudo apresenta métodos de mineração de dados para obter informações sobre a produção científica no Brasil e dar apoio à qualidade de ensino nos cursos de pós-graduação. Neste trabalho, objetiva-se realizar uma análise do crescimento das áreas de conhecimento da educação superior por região do Brasil. Buscou-se dados de cursos de pós-graduação de nível mestrado e doutorado de 1987 até 2018 com o histórico de todos os trabalhos de dissertação e teses, respectivamente, vinculados à CAPES, e também avaliar a evolução da quantidade de trabalhos acadêmicos produzidos nos cursos de pós-graduação do Brasil, para aplicar este estudo e se ter uma ideia dos avanços quantitativos relativos a cada área de conhecimento existente, conforme a classificação apresentada na tabela de áreas conhecimento da CAPES (CAPES, 2014).

A justificativa deste trabalho consiste na consolidação de uma base de dados estatísticos baseada nos dados abertos fornecidos pelo Portal Brasileiro de Dados Abertos (2020), mais especificamente no conjunto de dados abertos da CAPES, no intuito de criar uma referência para todas as áreas do conhecimento tabuladas pela CAPES que objetivem expressar os valores estatísticos ligados à produção acadêmica de uma área do conhecimento em particular.

Os dados utilizados nesta pesquisa foram obtidos em virtude da aprovação da Lei de Acesso à Informação (LAI) (BRASIL, 2011) e foram extraídos do conjunto de dados abertos da CAPES (2020), que conta com mais de 7 mil conjuntos de dados abertos que são analisáveis por qualquer cidadão.

ESTUDOS ANTERIORES

O referencial teórico deste trabalho foi feito por meio do levantamento de conceitos e definições de obras como artigos que abordam temas como produção científica e mineração de dados. A produção científica também foi restrita nos níveis de mestrado e doutorado e suas variantes como pós-graduação profissional.

Os fundamentos sobre produção científica podem ser vistos a partir da literatura referenciada a Oliveira, Rodrigues e Henriques (2004), que demonstra que os meios de divulgação da produção científica possibilitam a mineração e análise de dados e análises estatísticas quanto a sua dimensão de informação. Enquanto a produção científica teve sua evolução analisada por Moulin (2020) e o efeito da pandemia do COVID-19 e a evasão dos cursos de pós-graduação sob a curva da produção, conforme diz Rossoni (2020) e Ambiel (2020). Essa evolução é observada em diferentes aspectos em relação às áreas de pesquisa, uma vez que cada área tem ênfase em sua produção, assim como este trabalho analisa a produção científica sob a ótica da ciência de dados, assim como Moulin (2020) dá ênfase à área da administração.

A mineração e análise de dados reúne uma coleção de dados pertencentes a um grande grupo ou perfil de valores que satisfaçam a pesquisa feita sobre determinado universo de dados através de métodos de transformação em padrões que levam ao conhecimento. Métodos e técnicas de limpeza de bases de dados para otimizar o processo de pesquisa e mineração dos dados são abordados por Oliveira, Rodrigues e Henriques (2004). Enquanto os conceitos de mineração de dados necessários para ser feita a escolha da melhor estratégia de obtenção de dados de uma base são apresentados por Camilo e Silva (2009).

A mineração de dados educacionais, de acordo com Baker (2011), é uma área que pode ser explorada para otimizar a pesquisa de dados e a qualidade do ensino. Nas análises estatísticas encontradas na literatura são mostrados os tipos de inferência sobre as medidas relacionadas aos dados obtidos em uma pesquisa.

A análise de dados é feita através de ferramentas como o RStudio e R Project, este último desenvolvido por Ripley et al. (2001), para inferências acerca dos dados coletados, modelando em perspectivas a respeito dos dados mais relevantes e o objetivo da análise e trazendo um panorama do objeto de estudo, que no caso deste trabalho é mapear o perfil da produção científica nos cursos de pós-graduação do Brasil.

MATERIAL E MÉTODOS

A seguir é apresentado a metodologia utilizada para coleta, mineração e análise dos dados referentes às publicações acadêmicas realizadas pelos discentes matriculados nos cursos de pós-graduação do Brasil.

Conjunto de dados

O conjunto de dados extraído do Portal Brasileiro de Dados Abertos é composto por 32 anos de geração de dados sobre as dissertações e teses defendidas nos programas de pós-graduação vinculados à CAPES em âmbito nacional.

Este conjunto de dados reúne 1.142.090 trabalhos defendidos de pós-graduação. A base de dados é uma ferramenta de busca e consulta, com resumos, área do conhecimento, sigla da Instituição de Ensino Superior (IES), número de páginas, relativos a teses e dissertações defendidas desde 1987 até o ano de 2018. As informações são fornecidas diretamente à CAPES pelos programas de pós-graduação, que se responsabilizam pela autenticidade dos dados.

O conjunto de dados coletados é composto de arquivos no formato *Comma Separated Values* (CSV), organizado por ano e compostos de 42 a 57 variáveis e tamanhos de arquivos que variam de 5 MB até 261 MB. Nesta base de dados encontram-se informações sobre as principais variáveis presentes nos trabalhos acadêmicos defendidos em suas respectivas instituições de ensino.

A utilização do conjunto de dados e técnicas empregadas para análise de dados provê as condições à análise do desenvolvimento acadêmico por nível de pós-graduação, região e área do conhecimento que mais formaram pós-graduados que publicaram seus estudos no período sob análise, 32 anos.

Foram empregadas técnicas e modelos de predição de dados para cálculo de uma previsão de como estará a produção científica nos próximos anos, relativo às defesas realizadas embasado nos dados apurados dos anos anteriores. Inicialmente foi feita uma pesquisa sobre os conceitos necessários para o desenvolvimento do trabalho, destacando as teorias fundamentais para sustentação desta pesquisa acerca dos tópicos pontuados nas palavras-chave deste estudo.

O trabalho enquadra-se na área de estudo de inteligência computacional, mais precisamente no ramo da ciência de dados, conforme definido em Amaral (2016). Nesta área existem cinco linhas de estudo, conforme pode ser observado na listagem abaixo, sendo a última vertente listada, métodos estatísticos, a única que será abordada e devidamente aprofundada no decorrer deste trabalho.

1. **Lógica difusa:** a lógica difusa (comumente conhecida como lógica *fuzzy*) é uma forma de lógica na qual o valor da verdade das variáveis pode ser qualquer número real entre 0 e 1 (ZADEH, 2019). Ela é empregada para lidar com o conceito de verdade parcial, onde o valor da verdade pode variar entre completamente verdadeiro e completamente falso (NOVÁK; PERFILIEVA; MOCKOR, 2012);

2. **Redes neurais artificiais:** são sistemas computacionais inspirados nas redes neurais biológicas que constituem os cérebros dos animais (HARDESTY, 2017);

3. **Computação evolucionária:** família de solucionadores de problemas de tentativa e erro baseados na população com um caráter metaheurístico ou de otimização estocástica (EIBEN, 2003).

4. **Teoria da aprendizagem computacional:** é um subcampo da inteligência artificial dedicado ao estudo do projeto e análise de algoritmos de aprendizado de máquina (KEARNS; VAZIRANI, 1994); e

5. **Métodos estatísticos:** são procedimentos científicos que dizem respeito à coleta, organização, análise, interpretação e apresentação de dados (ROMEIJN, 2014).

O estudo proposto foi elaborado por meio do emprego das tecnologias R Project de Ripley et al. (2001), uma linguagem consolidada e voltada especificamente para análises estatísticas, além de Python que, segundo Rossum e Drake Jr. (2014), é uma linguagem de programação dinâmica, que tem sido bastante utilizada atualmente e bem provida de bibliotecas e recursos que auxiliam na solução de complexos problemas estatísticas e computacionais.

Tratamento e limpeza dos dados

De acordo com Oliveira, Rodrigues e Henriques (2004), o principal objetivo da fase de tratamento e limpeza dos dados é a remoção de desconformidades, erros e anomalias

presentes nos dados visando assim viabilizar a análise correta e precisa dos dados, garantindo assim a qualidade e veracidade dos resultados alcançados.

A limpeza dos dados envolve uma verificação manual ou automática da consistência das informações que serão processadas. A correção de possíveis erros de digitação ou de formatação, assim como o preenchimento ou a eliminação de valores nulos ou redundantes, o que tornará a análise confiável e com menos valores que desviem do padrão de dados adotado, ou *outliers*. Esta fase corrige e normaliza a base de dados utilizada, eliminando consultas desnecessárias que seriam executadas pelas funções estatísticas, podendo desse modo interferir negativamente na velocidade de processamento computacional e geração dos gráficos apresentados neste trabalho. Também não houve tratamento dos dados para pesquisas que foram desenvolvidas em mais de uma área do conhecimento.

RESULTADOS

Nesta seção estão apresentados os principais achados desta pesquisa, respondendo aos objetivos específicos deste trabalho. O tratamento e limpeza de dados da base de dados é feito através das tabelas com dados coletados do período de 1987 até 2018, disponíveis em planilhas em formatos CSV. As variáveis, e consequentemente colunas, sem utilização na análise foram descartadas para otimizar a inferência dos dados, tornando o conjunto de 42 a 57 variáveis em um conjunto de 7 variáveis, como é mostrado no Quadro 1 que exibe os nomes e tipos dessas variáveis.

Quadro 1. Variáveis e tipos de dados

Variável	Tipo da Variável
AnoBase	Numérico
Regiao	Texto
SiglaIes	Texto
GrandeAreaDescricao	Texto
AreasConhecimento	Texto
Nivel	Texto
Uf	Texto

Fonte: Elaborado pelos autores (2022).

No Quadro 1, observa-se que as variáveis selecionadas são referentes ao ano da publicação (AnoBase), região (Regiao), sigla da instituição de ensino superior (SiglaIes), nome do programa (NomePrograma), grande área do conhecimento (GrandeAreaDescricao), área de conhecimento (AreasConhecimento), nível do curso de pós-graduação (Nivel), data da defesa (DataDefesa) e unidade federativa (Uf).

A limpeza dos arquivos foi realizada na relação de arquivos com os dados coletados e representada em uma lista na forma de vetor. Após a criação da lista, o vetor é percorrido pelo número de arquivos registrados na lista e armazena os dados do arquivo selecionado, de acordo com o caminho fornecido, em uma tabela.

Em seguida, a partir da tabela é feita a remoção das variáveis desnecessárias, salvando-a em um novo arquivo para futuras consultas. No script observa-se que primariamente são importados os pacotes ggplot2 e Tidyverse de Wickham (2017). Ao se utilizar a biblioteca Tidyverse, há uma melhor reutilização das estruturas de dados existentes, além de possibilitar o manuseio de dados com programação funcional.

A variável global de armazenamento do caminho do diretório para uma fonte de dados sincronizada em nuvem e a prevenção de uma eventual perda de dados, e armazena-se o diretório de trabalho na variável “diretorio”. O ano de início da análise é armazenado na variável global “anoInicio”, sendo o ano da coleta mais antiga e disponível no banco de dados do governo, que é 1987.

A variável global “diretorio” armazena o caminho do diretório para uma fonte de dados sincronizada em nuvem e a prevenção de uma eventual perda de dados, e armazena-se o diretório de trabalho na variável “diretorio”. O ano de início da análise é armazenado na variável global “anoInicio”, sendo o ano da coleta mais antiga e disponível no banco de dados do governo, que é 1987.

O script então define a função “LimpDatas”, que tem em sua assinatura três parâmetros: a data atual da lista de planilhas coletadas “NomeDataAtual”, a nova data “NomeDataNovo” e o ano em questão “ano”.

Durante a execução da função “LimpDatas”, verifica-se se o ano em questão é maior que 1989. Caso confirmado, a data atual é substituída pelo ano atualmente lido, e o valor das variáveis “AreasConhecimento” e “AreasConhecimentoCodigo” são atribuídos pelo conteúdo das variáveis “AreaConhecimento” e “AreaConheciementoCodigo”.

A variável “NomeDataNovo” tem atribuído em seu valor o resultado do subconjunto das variáveis da data atual, acompanhada, respectivamente, dos nomes das variáveis a seguir: unidade federativa (Uf), região (Regiao), ano (AnoBase), sigla da instituição de ensino superior (SiglaIes), nome do programa de pós-graduação (NomePrograma), nome da grande área do programa (GrandeAreaDescricao), nome da área de conhecimento (AreasConhecimento), nível (Nivel) e a data de defesa da publicação (DataDefesa).

Em seguida é executada uma estrutura de repetição definida, em que a variável contadora “i” recebe, inicialmente o valor unitário, incrementando seu valor até equivaler ao comprimento do nível do valor da nova data.

No primeiro laço de repetição é verificado se o valor do vetor das datas de defesa, com a posição apontada pela variável contadora do laço, é equivalente a “Mestrado” ou “Doutorado”. Caso satisfeito, o valor da nova data tem o valor atribuído com o valor da data de defesa da nova data na posição de contagem atual.

Ainda na função de limpeza de datas, a variável da nova data, para mestrado, “NomeDataNovo1” tem atribuída em seu valor o subconjunto do valor da nova data e “NomeDataNovo” quando o atributo “Nivel” for equivalente a “Mestrado”. Já a variável da nova data para doutorado, “NomeDataNovo2” tem atribuída em seu valor o subconjunto do valor da nova data, “NomeDataNovo” quando o atributo “Nivel” for equivalente a “Doutorado”. Assim é atribuído na variável do valor da nova data, a combinação das linhas de mestrado e doutorado “NomeDataNovo1” e “NomeDataNovo2”, e é retornado como valor de saída da função.

Na seção global do script, há uma estrutura de repetição, onde a variável contadora “w” percorre duas posições de 21 e 22. Se o valor do contador for menor que 18 ou maior que 26, então é assinalado a atribuição “Ano” na variável “anoInicio” e o separador o valor nulo. Faz-se a leitura do arquivo CSV, com separadores “;” e caracteres de comentários “#”.

Nesta mesma estrutura verifica-se se o contador é maior que 26. Caso verdade, imprime-se na tela a confirmação que passou do valor e é atribuído o valor de “Ano” na variável “anoInicio” e renomeado o nome das variáveis para os valores respectivos apresentados no Quadro 2. Caso contrário, os valores das variáveis da planilha varrida são mantidos e atribuídos às variáveis. Após isso, imprime-se o ano e seu valor com o separador.

Assinala-se a chamada da função de Limpeza de Datas “LimpDatas”, em que os argumentos são do Ano, Ano de Base, e variável “anoInicio”.

Quadro 2. Atualização do nome das variáveis dos conjuntos de dados adotados

Nome antigo das variáveis	Nome atual das variáveis
AN_BASE	AnoBase
NM_REGIAO	Regiao
SG_ENTIDADE_ENSINO	SiglaIes
NM_PROGRAMA	NomePrograma
NM_GRANDE_AREA_CONHECIMENTO	GrandeAreaDescricao
NM_AREA_CONHECIMENTO	AreasConhecimento
NM_GRAU_ACADEMICO	Nivel
DT_TITULACAO	DataDefesa
SG_UF_IES	Uf
NM_ENTIDADE_ENSINO	NomeIes
CD_GRANDE_AREA_CONHECIMENTO	GrandeAreaCodigo
CD_AREA_CONHECIMENTO	AreasConhecimentoCodigo

Fonte: Elaborado pelos autores (2022).

Caso o contador seja 1, a variável “testi” tem em seu valor a obtenção do resultado da atribuição da variável “AnoIncio” em "Ano". Assim, remove-se da lista a variável atribuída em “testi”. Caso contrário a variável "testi" tem em seu valor a combinação da variável de seu próprio valor com a obtenção do resultado da atribuição da variável “AnoIncio” em “Ano”. Assim, é removida a variável atribuída em “testi”.

Assim a estrutura de repetição finaliza com o incremento unitário na variável “anoInicio”. Observa-se um trecho do código para pré-análise das variáveis e identificar a melhor representação destas nos gráficos.

Este código permite, através da função de obtenção de incidências únicas, chegar a valores em que o ano é definido por um intervalo de 32 anos. A região é definida por cinco variáveis: Norte, Nordeste, Centro-Oeste, Sudeste e Sul. As instituições de ensino superior representadas pelos dados são, no total, 316 IES. O número calculado de programas de pós-graduação chega a um valor de 1753 cursos. As variáveis adotadas neste estudo estão apresentadas no Quadro 3.

Quadro 3. Variáveis empregadas na mineração e análise de dados

Grande área do conhecimento	Nível de pós-graduação	Região
Ciências humanas	Doutorado acadêmico	Norte
Ciências biológicas	Doutorado profissional	Nordeste

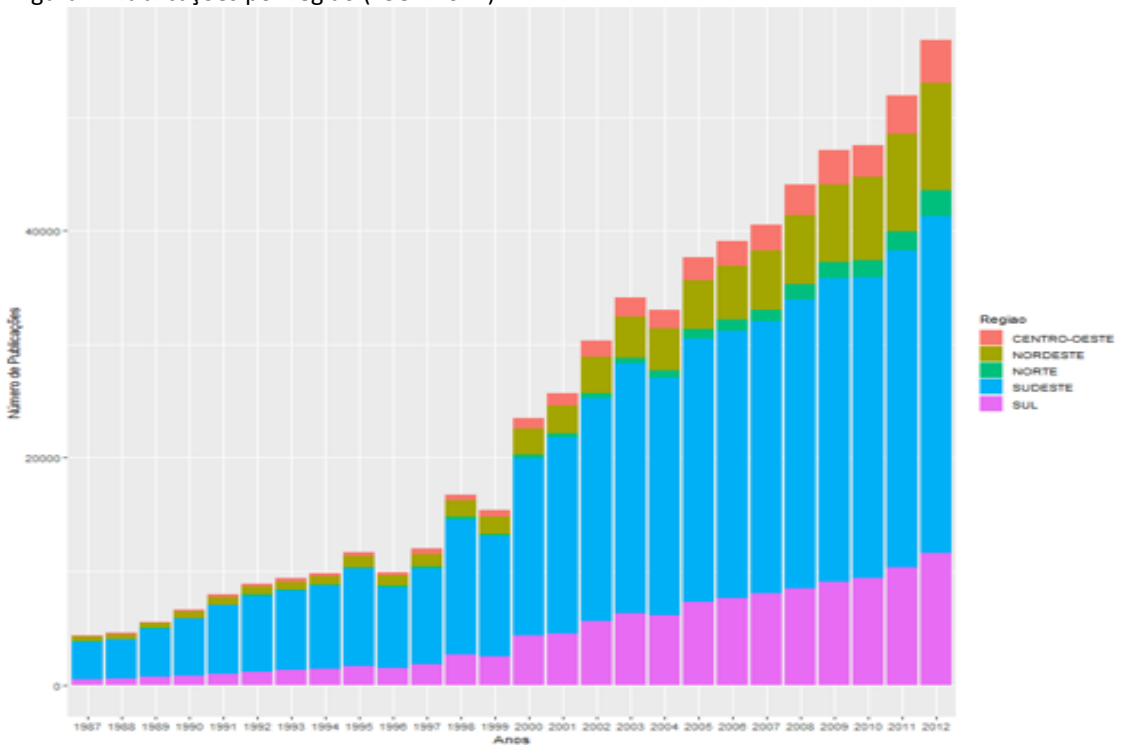
Ciências agrárias	Mestrado acadêmico	Centro-Oeste
Engenharias	Mestrado profissional	Sudeste
Ciências exatas e da terra		Sul
Ciências da saúde		
Ciências sociais aplicadas		
Linguística, letras e artes		
Multidisciplinar		
Grande área não informada		

Fonte: Elaborada pelos autores (2022).

As áreas de conhecimento consolidam 546 grupos. Os níveis do curso de pós-graduação armazenados no banco de dados são os de mestrado e doutorado. As datas de defesa chegam a um valor de 1841 dias de defesa, e o banco de dados cobre todas as 27 unidades federativas (26 estados e o Distrito Federal) do Brasil.

O script de leitura dos dados e geração dos gráficos, e seu armazenamento para futuras análises. Após o tratamento dos dados são gerados oito gráficos: Figura 1, Figura 2, itens a) e b) da Figura 3, Figura 4, Figura 5 e itens a) e b) da Figura 6. O primeiro é referente aos dados da variável “test”, seu mapeamento é feito conforme as premissas “AnoBase” e a quantidade de publicações por região, como mostrado na Figura 1.

Figura 1. Publicações por região (1987-2012)



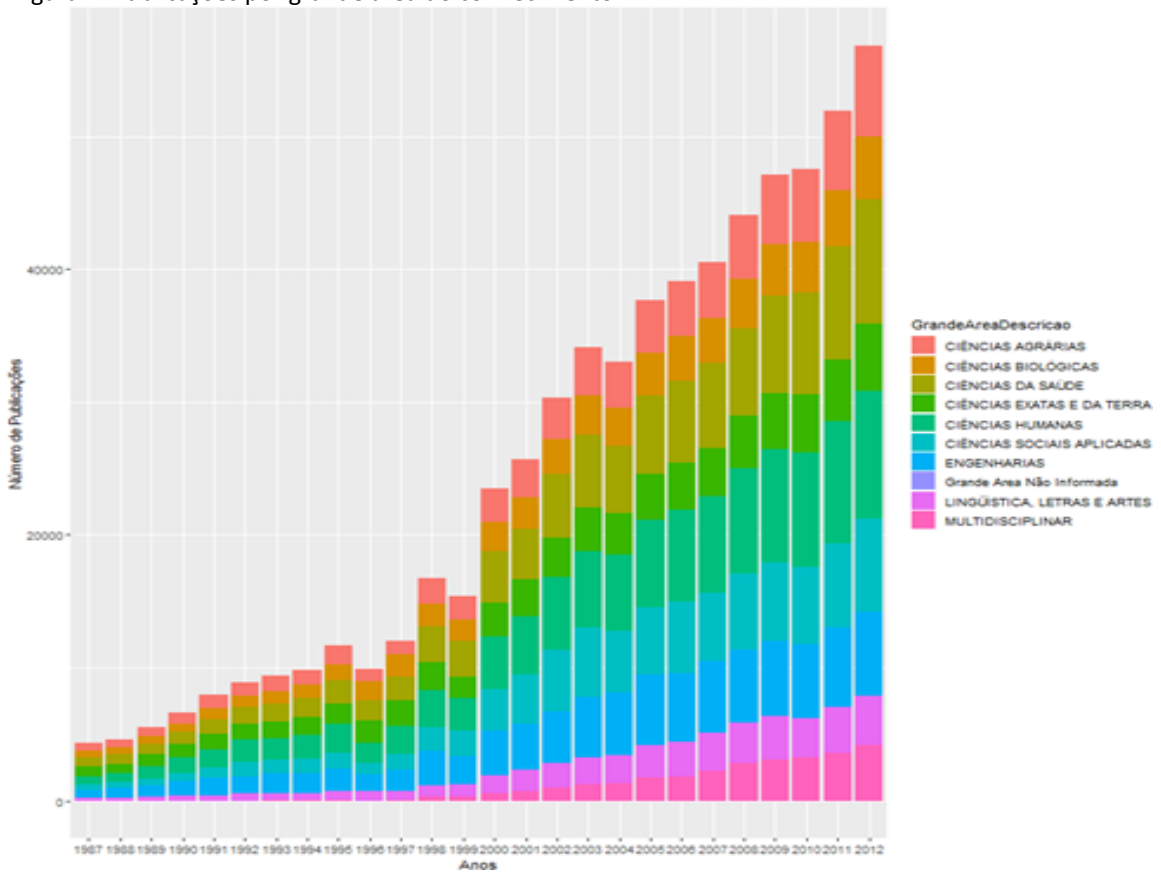
Fonte: Elaborada pelos autores (2022).

Este gráfico mostra que todas as publicações vêm aumentando no decorrer dos anos em todas as regiões, sendo que a maior quantidade está na região Sudeste, desde o começo da série. Mas no decorrer do tempo outras regiões vieram tendo uma expressiva elevação de publicações, principalmente nas regiões Sul e Nordeste, em ordem de publicações, e com maior elevação na região Centro-Oeste.

O segundo gráfico adota como *front-end* de dados o ano de 1987 e mapeando o nível das publicações por grande área do conhecimento, conforme é mostrado na Figura 2. Este gráfico mostra que todas as publicações vêm aumentando no decorrer dos anos em todas as grandes áreas do conhecimento. Observa-se que existem registros sem a informação correta da grande área da publicação, o que leva a inconsistências na leitura.

Ainda nesse gráfico, observa-se que a produção acadêmica nas áreas ciências humanas e ciências sociais aplicadas se destacam nos anos de 1987 e 1990. Após isso, observou-se que as áreas multidisciplinar e grande área não informada tiveram maior aumento da sua participação no crescimento das publicações.

Figura 2. Publicações por grande área do conhecimento

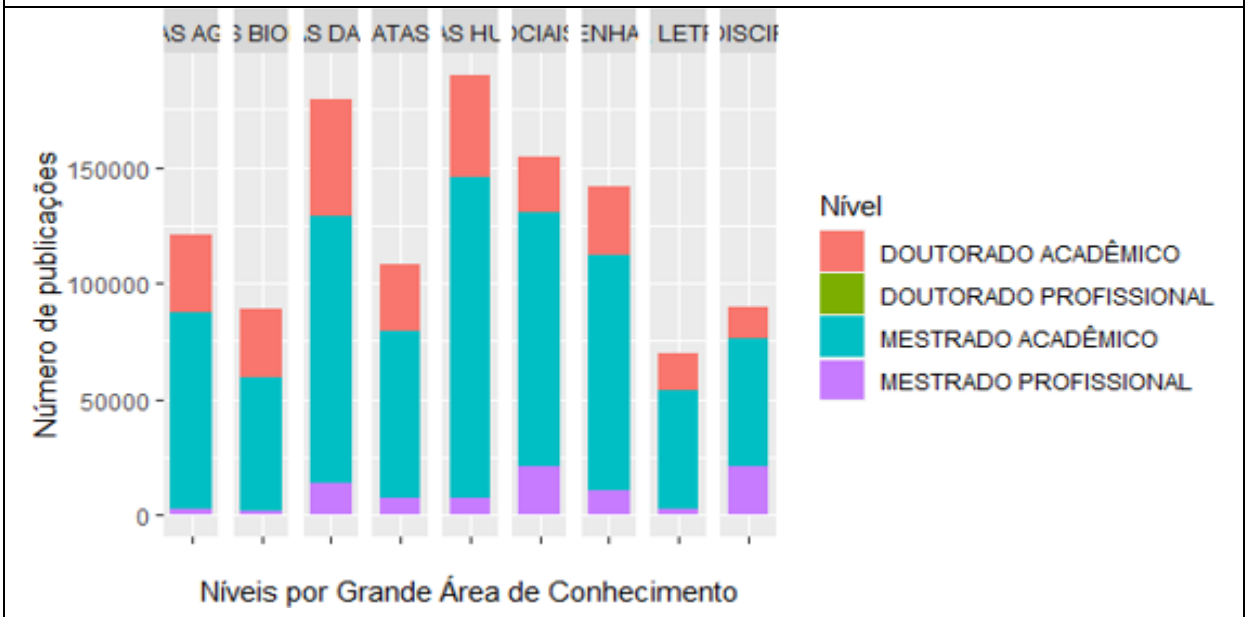


Fonte: Elaborada pelos autores (2022).

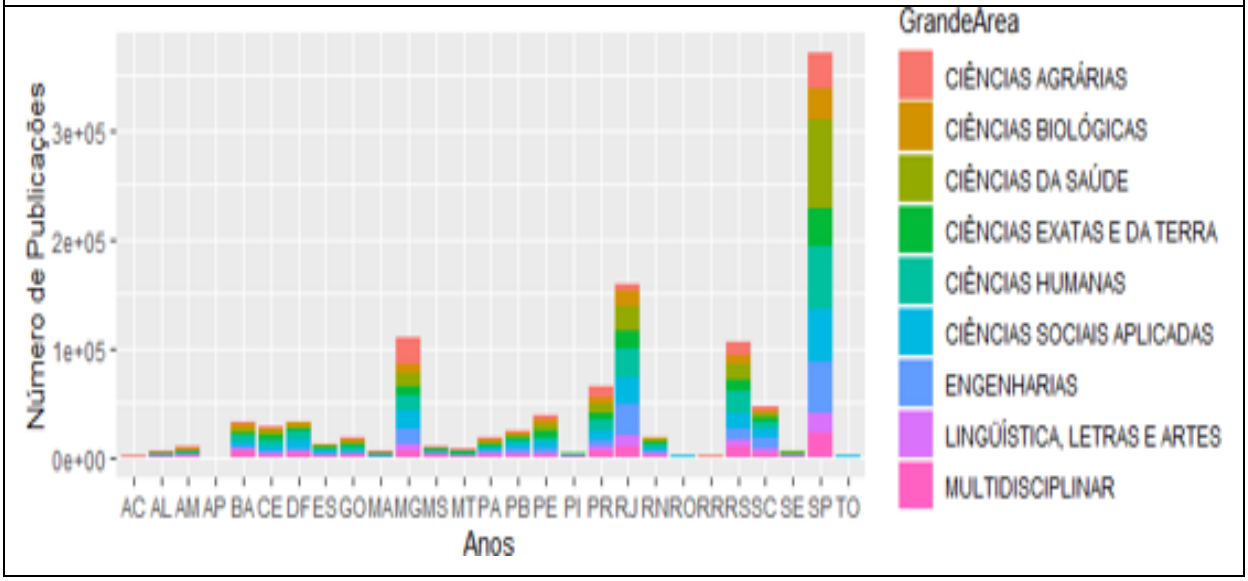
Na análise do item a) e do item b) da figura abaixo, as publicações foram separadas de acordo com o nível de pós-graduação e cada grande área do conhecimento, como pode ser visto na Figura 3. Nesses gráficos, observa-se que a maioria das publicações são de nível de mestrado e que, a menor diferença entre publicações de mestrado e doutorado se encontra na grande área do conhecimento de ciências biológicas. Já a maior diferença está nas grandes áreas de ciências sociais aplicadas, linguística, letras e artes, e multidisciplinar.

Figura 3. Publicações por nível de pós-graduação e grande área do conhecimento (1987-2018)

a) Gráfico expandido



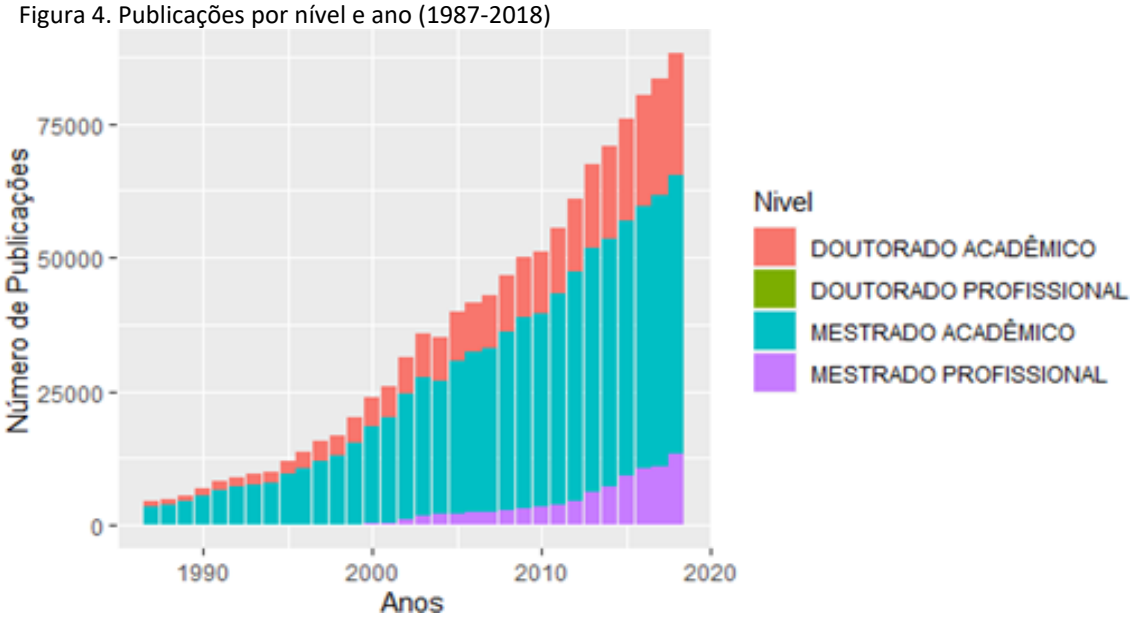
b) Gráfico compacto



Fonte: Elaborada pelos autores (2022).

Na análise da figura abaixo, as publicações foram separadas de acordo com o nível de pós-graduação por ano, como é apresentado na Figura 4.

Figura 4. Publicações por nível e ano (1987-2018)

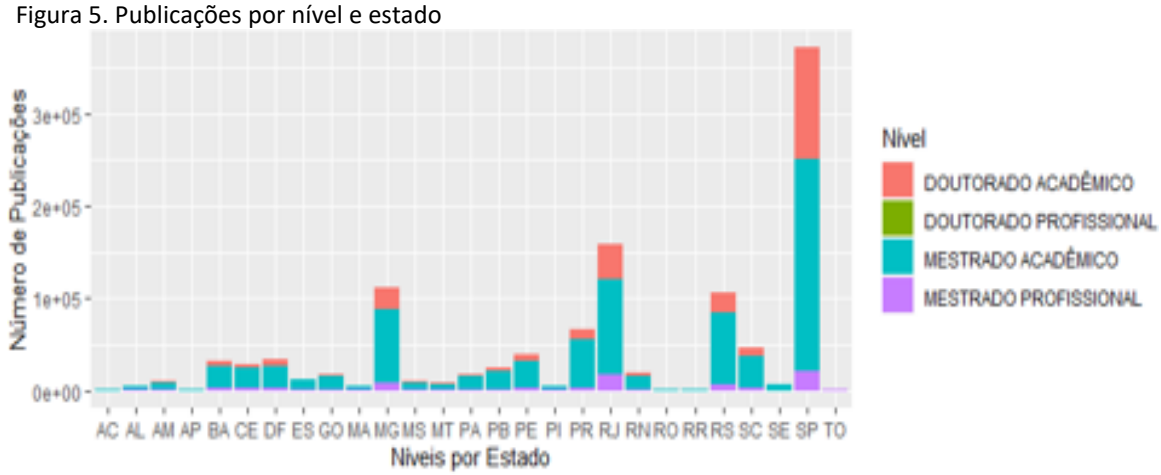


Fonte: Elaborada pelos autores (2022).

No gráfico mostrado na Figura 4, observa-se que desde 1987 até 2018 o número de publicações teve uma elevação na produtividade, principalmente nos trabalhos de doutorado. Observar o crescimento em três momentos de baixa na produtividade, tanto de mestrado como de doutorado. As baixas foram nos anos de 1996, 1999 e 2004. No quinto gráfico em análise, as publicações foram separadas de acordo com o nível de pós-graduação, em quantidades absolutas. Nesse gráfico, nota-se que, em termos absolutos, a quantidade de publicações do nível de mestrado é mais que o dobro das publicações de doutorado.

Na análise do gráfico a seguir, as publicações foram separadas de acordo com o nível de pós-graduação para cada estado, como pode ser visto na Figura 5.

Figura 5. Publicações por nível e estado

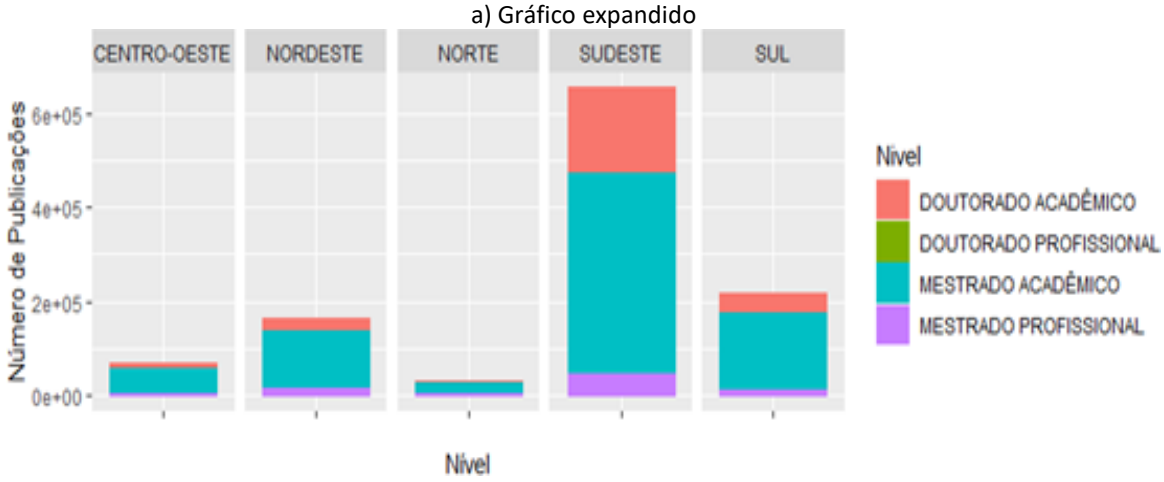


Fonte: Elaborada pelos autores (2022).

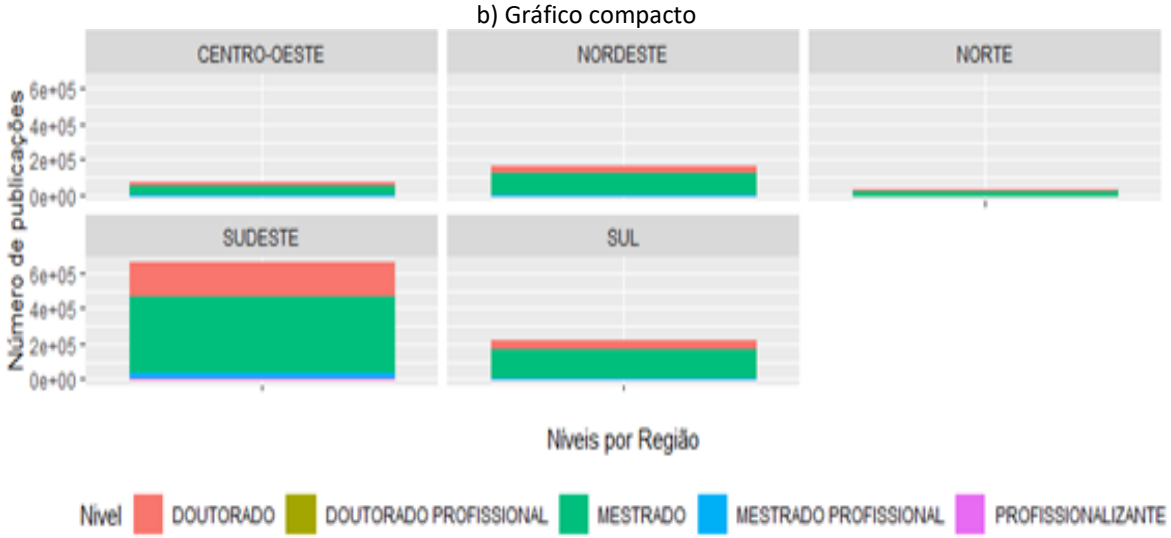
Na Figura 5, observa-se que a distribuição das publicações pelos estados é bem irregular. Há também uma irregularidade nas proporções nas publicações de mestrado e doutorado entre as unidades federativas.

O estado que tem a mais elevada concentração das publicações, tanto em cursos de mestrado como de doutorado, é o estado de São Paulo. Neste mesmo estado, observa-se que a diferença entre mestrado e doutorado é menor que a média. Já os estados com números menos expressivos que a média são: Sergipe, Alagoas, Mato Grosso do Sul, Mato Grosso, Piauí, Maranhão, Rondônia, Acre, Tocantins, Roraima e Amapá. No sétimo gráfico em análise, as publicações foram separadas de acordo com o nível de pós-graduação por região, como pode ser observado no item a) da Figura 6.

Figura 6. Publicações por nível de pós-graduação e região (1987-2018)
a) Gráfico expandido



b) Gráfico compacto



Fonte: Elaborada pelos autores (2022).

No gráfico do item a) da figura apresentada acima, observou-se que a região com mais publicações foi a região Sudeste, com diferença entre o número dos trabalhos de mestrado e doutorado aproximadamente ao valor da média. Já a região com menos publicações foi a região Norte. Enquanto a região do Brasil com uma diferença maior que a média entre as publicações de mestrado e doutorado foi a região Sul.

Na análise abaixo, as publicações foram separadas de acordo com o nível de pós-graduação por região, representado de forma compacta, como visto no item b) da Figura 6. Nesse gráfico, observou-se um outro ângulo da distribuição das publicações por região, comparado ao visualizado no item a) da Figura 6.

Esses resultados se devem ao fato da quantidade de número de cursos por região, que tem relação proporcional calculada a partir da quantidade de estudantes, instituições de pesquisa e ensino, população em geral e investimento público na pesquisa científica.

CONCLUSÕES E TRABALHOS FUTUROS

Como resultados, durante a coleta dos dados, observou-se que as variáveis dos anos iguais e posteriores a 2013 não seguem um padrão nos nomes de suas colunas e, em algumas planilhas, estão disponíveis novas informações, comprometendo assim a realização de uma análise histórica mais abrangente do conjunto dados empregado neste estudo. Observa-se que ao longo dos anos, a tendência dos números de artigos é de crescimento em todas as regiões do Brasil.

Em relação aos resultados das análises das variáveis das: siglas de instituição de ensino superior, nome do programa, área de conhecimento e data de defesa, mostra-se que o número de valores para as variáveis são representáveis por meio de estatísticas descritivas (como média, mediana, desvio padrão, máximo, mínimo, entre outras) em uma relação com as variáveis trabalhadas.

Em relação às análises dos gráficos gerados, todos satisfizeram as previsões calculadas e mostraram que algumas regiões têm mais tendência de crescimento que outras, da mesma forma acontece com grandes áreas do conhecimento que têm definição não informada, ou seja, áreas não catalogadas ou mais específicas estão apresentando um maior crescimento em relação às demais grandes áreas do conhecimento exibidas no Quadro 3.

Algumas variáveis como, por exemplo, a quantidade de páginas da publicação não é apresentada em alguns *datasets* e, por isso, esses dados são ignorados na análise realizada neste estudo. Uma solução encontrada para ser aplicada trabalhos futuros, em relação a coleta dos dados dos últimos anos é ter colunas que apresentem o mesmo tipo de informação, e selecionadas para a análise dos dados, sejam renomeadas e os nomes das colunas das planilhas mais antigas, que são as colunas adotadas e com maior segurança, uma vez que é mais rápido recuperar informações das extremidades do banco de dados.

Esse trabalho foi desenvolvido porque viu-se a necessidade de visualizar a evolução histórica da produção de trabalhos acadêmicos produzidos pelas universidades e institutos de pesquisa nacionais. Como resultado deste estudo, espera-se que a partir dos dados, códigos-fonte e gráficos apresentados ou disponibilizados haja um maior entendimento acerca dessa área de estudo para que futuros pesquisadores e cientistas possam usufruir dessas informações em suas teses e projetos pessoais.

Tendo em vista que os resultados alcançados se limitaram a mostrar a quantidade de trabalhos e, desse modo, não especificando quais foram as temáticas, áreas do conhecimento abordadas e metodologias desses trabalhos. Propõe-se, como trabalho futuro, demonstrar quais foram as palavras-chave mais utilizadas para, nesse sentido, evidenciar quais foram as contribuições dos estudos analisados, classificando-os em por meio de suas respectivas áreas de concentração.

Os dados coletados podem contribuir para pesquisadores da área da educação, mais especificamente na subárea dos cursos de pós-graduação, a fim de aferir conclusões acerca de quais áreas do conhecimento e locais (estados e/ou regiões) necessitam de maior apoio e

investimento e, com isso, desbloquear novas oportunidades de pesquisa acadêmica que podem ser convertidas em: *blueprints*[1], patentes, fórmulas e novos produtos a serem comercializados nacionalmente e internacionalmente.

Disponibilizou-se publicamente toda a documentação e o código-fonte implementado neste trabalho na plataforma de repositórios de software GitHub na URL https://github.com/thiagoddcqg/ciencia-de-dados-producao-academica-brasil, para que pesquisadores, desenvolvedores e cientistas de dados interessados possam usufruir, melhorar ou adaptar os artefatos de software produzidos neste estudo em suas respectivas pesquisas e projetos pessoais.

REFERÊNCIAS

AMARAL, Fernando. **Introdução à ciência de dados: mineração de dados e big data**. Alta Books Editora, 2016.

AMBIEL, Rodolfo Augusto Matteo et al. Motivos de evasão na pós-graduação no Brasil: um instrumento de medida. **Interação em Psicologia**, v. 24, n. 1, 2020.

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineração de dados educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 02, p. 03, 2011.

BOAVENTURA, Michel et al. Caracterização temporal das redes de colaboração científica nas universidades brasileiras: anos 2000-2013. In: **Anais do III Brazilian Workshop on Social Network Analysis and Mining**. SBC, 2014. p. 9-20.

BRASIL. Lei n.12.527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do artigo 5º, no inciso II do 3º do art. 37 e no 2º do art.216 da Constituição Federal; altera a Lei n.8.112, de 11 de dezembro de 1990; revoga a lei n.11.111, de 5 de maio de 2005, e dispositivos da Lei n. 8.159, de 8 de janeiro de 1991; e dá outras providências. Diário Oficial da União, Brasília, 19 nov. 2011.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, v. 1, n. 1, p. 1-29, 2009.

CAPES. [1987 a 2012] Catálogo de Teses e Dissertações - Brasil. **Dados Abertos CAPES**, 2019. Disponível em: <https://dadosabertos.capes.gov.br/dataset/1987-a-2012-catalogo-de-teses-e-dissertacoes-brasil>. Acesso em: 19 maio 2022.

CAPES. [2013 a 2016] Catálogo de Teses e Dissertações - Brasil. **Dados Abertos CAPES**, 2020. Disponível em: <https://dadosabertos.capes.gov.br/dataset/catalogo-de-teses-e-dissertacoes-de-2013-a-2016>. Acesso em: 20 maio 2022.

[1] Segundo Liebing (1999), *blueprint* é a reprodução de um desenho técnico ou desenho de engenharia usando um processo de impressão de contato em folhas sensíveis à luz.

CAPES. [2017 a 2018] Catálogo de Teses e Dissertações - Brasil. **Dados Abertos CAPES**, 2019. Disponível em: <https://dadosabertos.capes.gov.br/dataset/2018-catalogo-de-teses-e-dissertacoes-da-capes>. Acesso em: 3 dez. 2021.

CAPES. Página inicial. **Dados Abertos CAPES**, 2020. Disponível em: <https://dadosabertos.capes.gov.br/>. Acesso em: 19 maio 2022.

CAPES. Tabela Áreas Conhecimento. **CAPES**, 2014. Disponível em: <http://www.capes.gov.br/images/stories/download/avaliacao/TabelaAreasConhecimento_042009.pdf>. Acesso em: 20 maio 2022.

DIAS, Thiago Magela Rodrigues et al. Obtenção de dados científicos a partir de repositórios de dados curriculares. **Cadernos BAD**, n. 1, p. 326-333, 2018.

EIBEN, Agoston E. et al. **Introduction to evolutionary computing**. Berlin: springer, 2003.

HARDESTY, Larry. Explained: neural networks. **MIT News**, v. 14, 2017.

KEARNS, Michael J.; VAZIRANI, Umesh. **An introduction to computational learning theory**. MIT Press, 1994.

LIEBING, Ralph W. **Architectural working drawings**. John Wiley & Sons, 1999.

MOULIN, Gabriela et al. Produção Científica e Sociedade: a Fronteira entre o Passado e o Futuro. **Gestão e Sociedade**, v. 14, n. 37, p. 3439-3460, 2020.

NOVÁK, Vilém; PERFILIEVA, Irina; MOCKOR, Jiri. **Mathematical principles of fuzzy logic**. Springer Science & Business Media, 2012.

OLIVEIRA, Paulo; RODRIGUES, Fátima; HENRIQUES, P. Limpeza de dados-uma visão geral. **Data Gadgets**, p. 39-51, 2004.

RIPLEY, Brian D. et al. The R project in statistical computing. **MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network**, v. 1, n. 1, p. 23-25, 2001.

ROMEIJN, Jan-Willem. Philosophy of statistics. 2014.

ROSSONI, Luciano. Covid-19, Organizações, Trabalho em Casa e Produção Científica. **Revista Eletrônica de Ciência Administrativa**, v. 19, n. 2, p. 158-168, 2020.

VAN ROSSUM, Guido; DRAKE JR, Fred L. The python language reference. **Python software foundation**, 2014.

WICKHAM, Hadley. The Tidyverse, R Package Ver. 1. 1. 1. **Open-source statistical software package**, 2017.

ZADEH, Lotfi. Some Thoughts About Appealing Directions for the Future of Fuzzy Theory and Technologies Along the Path Traced. In: **Fuzzy Logic and Applications: 12th International Workshop, WILF 2018, Genoa, Italy, September 6-7, 2018, Revised Selected Papers**. Springer, 2019. p. 240.