Revista Tecnologia e Sociedade
ISSN: 1984-3526

https://periodicos.utfpr.edu.br/rts

Estudo sobre o uso de big data na Produção Estatística

Elizabeth Belo Hypolito
Escola Nacional de Ciências Estatísticas (ENCE/IBGE), Rio de Janeiro, RJ.
elizabeth.hypolito@ibge.gov.br

Andrea Diniz da Silva
Escola Nacional de Ciências Estatísticas (ENCE/IBGE), Rio de Janeiro, RJ.
andrea.silva@ibge.gov.br

Antonia Xavier
Escola Nacional de Ciências Estatísticas (ENCE/IBGE), Rio de Janeiro, RJ.
antonia2301@outlook.com

Átila Kopplin Chiquito
Escola Nacional de Ciências Estatísticas (ENCE/IBGE), Rio de Janeiro, RJ.
atilakopplin@gmail.com

Lucas Uchoa Moreira Gomes
Escola Nacional de Ciências Estatísticas (ENCE/IBGE), Rio de Janeiro, RJ.
lucasgomes.uchoa@gmail.com

Isis Gonçalves Peixoto
Escola Nacional de Ciências Estatísticas (ENCE/IBGE), Rio de Janeiro, RJ.
isis.peixoto@gmail.com

Beatriz Menezes Marques de Oliveira
Escola Nacional de Ciências Estatísticas (ENCE/IBGE), Rio de Janeiro, RJ.
biameny@gmail.com

Antonio Etevaldo Teixeira Junior
Senac, Departamento Nacional, Rio de Janeiro, RJ.
antonio.junior@senac.br

Alvaro de Moraes Frota
Instituto Brasileiro de Geografia e Estatística (IBGE)
alvrofr@ibge.gov.br

RESUMO

A necessidade de estatísticas públicas de alta qualidade é cada vez mais premente em nossa sociedade, enquanto os recursos destinados à sua produção são cada vez mais escassos. Nesse contexto, pesquisadores e institutos de estatística vêm investindo cada vez mais em fontes de dados alternativas, incluindo big data. O presente trabalho tem como objetivo revelar a produção acadêmica em torno do tema a partir de uma amostra probabilística de artigos, selecionada de um cadastro extraído do Google Scholar. Os resultados permitem estimar que, de 2013 a 2020, 1211 artigos gratuitos, escritos em português, espanhol ou inglês, que estudavam o uso de big data na produção de estatísticas para tópicos selecionados foram disponibilizados nessa plataforma. Desses, 554 abordavam o tópico saúde. A observação das características gerais dos artigos permitiu a construção de uma base de metadados capaz de contribuir para a discussão dos limites e as potencialidades de big data na produção de estatísticas.

PALAVRAS-CHAVE: Big data. Estatísticas públicas. Estatísticas oficiais. Pesquisa bibliográfica.

INTRODUÇÃO

A demanda por mais e melhores estatísticas cresce em direção oposta à disponibilidade de recursos destinados à produção de estatísticas públicas, em especial os oficiais. Para melhorar a oferta de tais estatísticas e atender demandas nacionais e internacionais, institutos de estatística de vários países têm investido cada vez mais no uso de fontes alternativas de dados como é o caso de big data. A Plataforma Global das Nações Unidas registrou pelo menos 111 experiências nacionais no uso de raspagem da web, dados de telefone celular, redes sociais e imagens de satélite, entre outras fontes, para produzir um amplo número de estatísticas, incluindo preços, migração, agricultura, mobilidade e indicadores para ao menos os Objetivos de desenvolvimento Sustentável (ODS) 1, 2, 5, 6, 11, 14 e 15[1] (MacFeely, 2019).

Embora seja promissor, big data vem acompanhado de limitações que devem ser observadas. Seu uso traz desafios que incluem a necessidade de buscar estratégias para a coleta de dados que sejam representativos da população de interesse, de identificar métodos estatísticos confiáveis e precisos para estimar as quantidades de interesse e ferramentas para análise de dados que possibilitem reduzir vieses decorrente de sua forma de obtenção. Além disso, é essencial a disponibilidade de profissionais qualificados, infraestrutura computacional e legislação que facilite o acesso aos dados.

Para apoiar institutos nacionais de estatística no uso de big data, as Nações Unidas criaram quatro hubs regionais: Brasil, China, Emirados Árabes Unidos e Ruanda. O Hub Regional da ONU para Big Data no Brasil, que atende a América Latina e o Caribe, vem desenvolvendo diversas ações para promover a capacitação e fomentar o interesse de jovens estatísticos sobre o uso de big data, apoiar o compartilhamento de experiências e conhecimentos, fortalecer laços e promover a cooperação na Região.

Dentre suas atividades, encontra-se a pesquisa bibliográfica ora apresentada. Tal pesquisa busca contribuir para o avanço da discussão sobre a temática, classificando os documentos do tipo artigo segundo sua abordagem e tema, e ofertando uma base de metadados que pode ser utilizada para discutir os limites e as potencialidades de big data.

Para o cumprimento de tal objetivo, uma pesquisa inicial foi realizada por meio do mecanismo de busca Google Scholar, permitindo a construção de um cadastro de documentos técnicos e acadêmicos sobre big data. A análise desse cadastro forneceu características gerais da produção bibliográfica entre 2004 e 2020. Posteriormente, uma amostra foi então selecionada com base no cadastro, possibilitando a classificação de artigos acadêmicos segundo o tópico e a abordagem adotada. Os dados da amostra propiciaram um estudo mais aprofundado sobre a produção do período 2013 a 2020.

MÉTODOS

Cadastro de referência

O cadastro foi elaborado a partir da lista de documentos fornecida pelo indexador Google Scholar, o qual foi escolhido por apresentar maior cobertura da produção acadêmica. Segundo um prognóstico de novembro de 2018, esse indexador conta com cobertura de cerca de 80% a 90% dos periódicos científicos publicados em inglês e tem mais de 389 milhões de documentos indexados, o que o torna o maior indexador e base de dados acadêmicos (GUSENBAUER, 2018). Com 98% de toda produção científica mundial sendo produzida em inglês (Ramírez-Castañeda, 2015), o Google Scholar se tornou o mais popular indexador e base de dados para a produção acadêmica entre os pesquisadores, gerando até mesmo polêmicas quanto a tal monopólio.

Para corroborar com a escolha do indexador Google Scholar, primeiramente, seus resultados foram comparados com os dos indexadores SciELO e DOAJ, para a chave de busca *big data AND "official statistics"* e verificou-se que o Google Scholar apresentou maior número de resultados. Depois, um teste de sobreposição dos resultados foi realizado utilizando os seguintes indexadores e bancos de dados: WorldWideScience, SciELO, Science Direct, Periódicos Capes, Biblioteca Brasileira de teses e dissertações, Scholarpedia, ERIC, Science Research, Science.gov, DOAJ, REDALYC, JSTOR, Sistema de Bibliotecas da UNICAMP, Aminer, BASE, The Lens, CORE, Semantic Scholar e Microsoft Academic (desativado ao final de 2021). Foram selecionados 10 resultados nas primeiras 20 páginas de cada indexador, os quais foram recuperados usando apenas a chave de busca big data. Somente o indexador Science.gov apresentou 2 resultados não indexados pelo Google Scholar. Para os demais, toda a produção constava também do indexador.

A pesquisa foi feita na língua inglesa por ser o idioma mais utilizado nos documentos indexados pelo Google Scholar. Foram pesquisados os títulos dos documentos, buscando-se os termos big data e um tópico considerado de relevância, os quais foram elencados por MacFeely (2019) com base no inventário do UN big data sobre o uso desse tipo de dados para a produção de estatística. As chaves de busca utilizadas são apresentadas no Quadro 1.

A pesquisa foi realizada com uso da técnica web scraping. Para implementar a busca, foi utilizado o programa Publish Or Perish (HARZING, 2007). Ao todo, foram obtidos 5.042 documentos, incluindo artigos, teses, dissertações, relatórios, apresentações, entre outros, sendo o documento mais antigo do ano de 2004 e o mais recente do ano de 2020. Para cada documento do cadastro, foram obtidas informações como título, autores, ano, editor, cidade de publicação, URL do artigo, DOI, tipo de documento disponível (PDF, DOC ou HTML), entre outras.

Quadro 1: Chaves de busca utilizadas no Google Scholar para a elaboração do cadastro

Chaves de busca
big data AND "agriculture", *big data AND "corruption",* *big data AND "crime",* *big data AND "disaster risk reduction",* *big data AND "disease",* *big data AND "energy",* *big data AND "environment",* *big data AND "geographical",* *big data AND "health",* *big data AND "inequality",* *big data AND "labour market",* *big data AND "land use",* *big data AND "migration",* *big data AND "mobility",* *big data AND "population",* *big data AND "poverty",* *big data AND "prices",* *big data AND "spatial",* *big data AND "tourism" e* *big data AND "transport".*

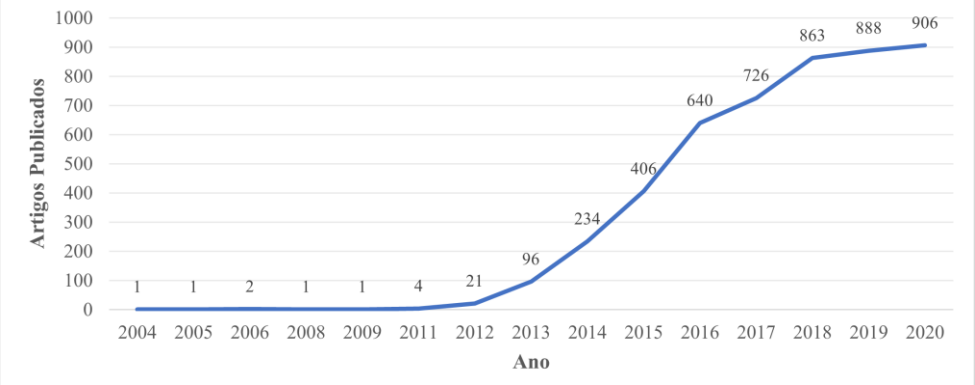
Fonte: Elaboração das/os autoras/es.

Características gerais do cadastro de referência

Com base no cadastro de 5.042 documentos, extraídos do Google Scholar para o período de 2004 a 2020, foi possível identificar que, nos últimos anos, houve um crescimento na quantidade de documentos sobre big data que, com base nos termos do Quadro 1, podem estar relacionados à produção estatística, em especial a partir de 2013. Até 2011, a quantidade não passou de 4 por ano, mas em 2020 foram obtidos 906 documentos (Gráfico 1).

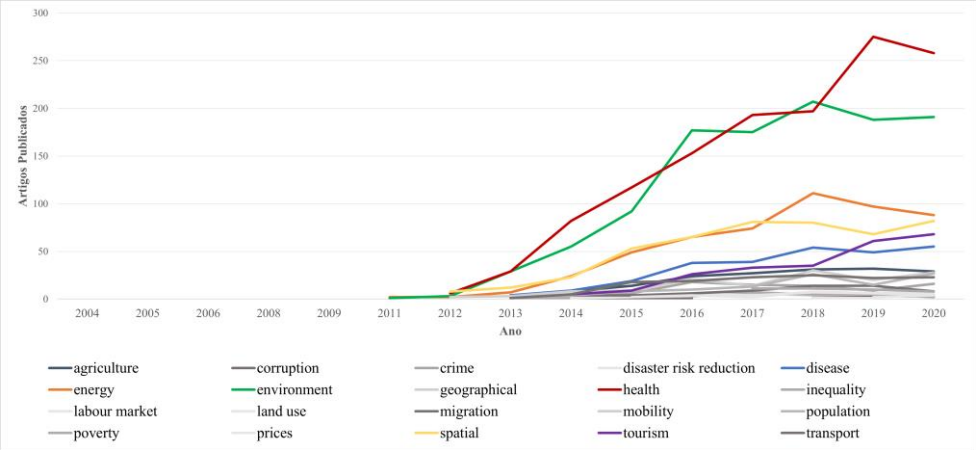
As chaves de busca que geraram maior número de resultados foram big data AND "health" e big data AND "environment". Os documentos contendo o termo saúde (health) no título mostraram aumento consistente até o ano de 2019, sofrendo redução durante os dois últimos anos, o que surpreende, já que foram anos em que as questões relacionadas com saúde ganharam mais visibilidade por causa da pandemia da Covid-19. Já os documentos contendo o termo ambiente (environment) tiveram trajetória menos linear a partir de 2016 e apresentaram crescimento no período pandêmico. A quantidade de documentos com o termo turismo (tourism) cresceu desde 2018. A chave de busca big data AND "spatial" foi a quarta com mais resultados em 2020 (Gráfico 2).

Gráfico 1: Número de documentos disponíveis no Google Scholar sobre uso de big data, para tópicos selecionados, 2004 a 2020.



Fonte: Google Scholar, 2021.

Gráfico 2: Número de documentos sobre uso de big data, por tópico selecionado, 2004 - 2020



Fonte: Google Scholar, 2021.

Ademais, os documentos com os termos saúde (health) e ambiente (environment) representavam, aproximadamente, metade do total encontrado (50,8%). Energia (energy), que se encontrava em terceiro lugar tinha menos da metade dos resultados de seu antecessor no ranking, ambiente (environment), mesmo tendo um crescimento significativo em 2020. Termos como desigualdade (inequality), geografia (geographical), corrupção (corruption), mercado de trabalho (labour market) e redução de risco de desastre (disaster risk reduction) ocupavam as últimas posições em material encontrado, somando 1,5% dos resultados (Tabela 1).

Tabela 1: Número de documentos sobre uso de big data, por tópico selecionado, 2004 a 2020

Tópico utilizado na busca	Número de documentos	
	Valor Absoluto	Valor Relativo
Health	1383	27,4%
Environment	1179	23,4%
Energy	543	10,8%
Spatial	501	9,9%
Disease	279	5,5%
Tourism	256	5,1%
Agriculture	182	3,6%
Mobility	142	2,8%
Transport	140	2,8%
Population	108	2,1%
Crime	87	1,7%
Migration	63	1,2%
Poverty	40	0,8%
Land use	34	0,7%
Prices	29	0,6%
Inequality	25	0,5%
Geographical	26	0,5%
Corruption	12	0,2%
Labour market	8	0,2%
Disaster risk reduction	5	0,1%
Total	5042	100,0%

Fonte: Google Scholar, 2021.

Aspectos metodológicos do levantamento amostral

Considerando que também era objetivo do estudo classificar artigos sobre o uso de big data para a produção estatística e, dado o grande volume de documentos no cadastro anteriormente descrito (5.042 documentos), optou-se por utilizar técnicas de amostragem.

População alvo

A população alvo foi composta pelos documentos do tipo artigo, publicados a partir de 2013, em inglês, português ou espanhol, com acesso gratuito e que, de fato, abordavam a produção estatística. A opção de analisar apenas artigos se justifica pelo fato de serem os mais utilizados e citados na literatura acadêmica. Como ano inicial foi escolhido aquele para o qual a produção sobre o tema já se aproximava de uma centena. Por fim, a escolha por artigos gratuitos teve o objetivo de retratar um acervo democrático, que pudesse ser acessado por qualquer pessoa com acesso à Internet, reduzindo assim a seletividade por condições materiais ou financeiras.

Planejamento amostral

Primeiramente, foram excluídos do cadastro de referência textos com ano de publicação anterior a 2013. Após essa exclusão o cadastro passou a conter 4.603 documentos. Ademais, foram criados 2 domínios:

- Domínio A: 918 documentos com PDF, DOC ou HTML disponível, ou seja, supostamente com gratuidade de acesso; e
- Domínio B: 3.685 documentos sem informação sobre disponibilidade.

Após a definição dos domínios, foi realizado um procedimento de estratificação. Em ambos os domínios os estratos foram formados a partir dos termos de busca, apresentados no Quadro 1. No domínio B, também foi usado o ano de publicação do documento. Essa estratificação resultou em 12 estratos no domínio A e 25 estratos no domínio B.

O tamanho da amostra foi arbitrado em 1.000 artigos, quantidade que forneceria precisão satisfatória para a realização do estudo. Devido à incerteza da quantidade de artigos nos 2 domínios, a alocação da amostra não foi proporcional à quantidade de artigos de cada domínio. Apesar do domínio A ser praticamente quatro vezes menor que o domínio B, 40% da amostra foi alocada para ele (400 artigos), enquanto os 600 artigos restantes, foram alocados no domínio B.

É importante destacar que o cadastro não possuía todas as informações necessárias para identificar de antemão os documentos pertencentes à população-alvo. Dessa forma, a verificação da elegibilidade com base no tipo de documento (artigo ou outro tipo), na língua de origem (inglês, português ou espanhol) e, no caso do domínio B, na gratuidade do documento, foi feita no momento da coleta. Para tal, foram adotados os seguintes protocolos de amostragem inversa (HALDANE, 1945):

- Geração de números aleatórios para cada documento;
- Ordenação dos documentos dentro de cada estrato de acordo com os números aleatórios.
- Percorrimento de cada estrato, ou seja, verificação de elegibilidade de cada documento, até se atingir o respectivo tamanho amostral.
- Registro de todas as tentativas, tanto as bem-sucedidas quanto as malsucedidas.

Durante essa operação, verificou-se que a maioria dos documentos dos estratos relativos à chave de busca *big data AND "enviroment"* do domínio B não estavam em consonância com o objetivo do estudo, já que se referiam a ambientes computacionais e outros tipos de ambientes que não estão diretamente relacionados à produção de estatística. Por isso, esses estratos foram descartados do processo de amostragem.

Como mencionado anteriormente, na amostragem inversa, cada estrato é percorrido até a obtenção do tamanho amostral planejado para ele. Assim, nem todos os elementos do cadastro foram verificados e classificados como pertencentes ou não à população alvo. Logo, foi necessário estimar o tamanho da população alvo, levando em conta a proporção de documentos cuja elegibilidade foi confirmada entre os documentos pesquisados. Dessa forma, a população alvo foi estimada em 1.211 artigos.

Pesos amostrais

O tamanho final da amostra foi de 819 artigos. Para realizar a expansão da amostra com vista à obtenção de valores populacionais, os pesos amostrais foram obtidos mediante a divisão do tamanho da população alvo estimada no estrato pelo total de artigos gratuitos nesse mesmo estrato. Os pesos variaram no intervalo de 1,00 a 2,18, com média de 1,48.

Principais variáveis

Além das informações provenientes do cadastro, tais como título, autor e ano de publicação dos artigos, e aquelas observadas durante atapa de verificação de elegibilidade, como língua de publicação, cada um dos 819 artigos da amostra foi analisado para que outras informações de interesse fossem acrescentadas à base de metadados. Essas novas variáveis foram o número de páginas, as palavras-chaves (*keywords*), o tipo de material, a abordagem e o tema.

O número de páginas e as palavras-chaves definidas pelos autores foram coletadas com base em observações diretas dos artigos. Já a classificação do tipo de material em quatro categorias: a) artigo em anais/eventos, b) artigo em revista (journal), c) artigo em livro e d) artigo em outros canais, exigiu, em alguns casos, uma análise mais apurada do material e das características de seu editor.

Em relação à abordagem e ao tema, foi necessário primeiramente definir as categorias que seriam de interesse para o estudo. Posteriormente, a classificação do artigo foi feita com base na leitura de seu resumo ou, se necessário, do texto completo.

Para a abordagem, isto é, enfoques e elementos principais destacados pelo artigo, foram definidas as categorias a saber: metodológica, conceitual, aplicação com dados reais, aplicação com dados sintéticos, levantamento de vantagens e desvantagens, revisão bibliográfica, pesquisa bibliográfica, boas práticas e qualidade de dados. É importante destacar que as opções não eram mutuamente excludentes, ou seja, um artigo podia ser classificado em uma ou mais categorias de abordagem.

Na abordagem metodológica destaca-se o desenvolvimento de métodos específicos ou a descrição mais detalhada de um método existente, o qual tenha sido utilizado. A abordagem conceitual se baseia na definição de conceitos apresentada pelos artigos. Artigos com alguma aplicação com dados de big data foram classificados na abordagem “aplicação com dados reais” ou “aplicação com dados sintéticos” a depender do tipo de dado utilizado. Em levantamento de vantagens e desvantagens foram classificados os artigos que realizassem alguma discussão a respeito dos prós e contras, seja do uso de big data seja do método ou aplicação apresentada pelo próprio artigo.

As abordagens de revisão bibliográfica, pesquisa bibliográfica, boas práticas e qualidade de dados são mais auto evidentes. As duas primeiras tratam de artigos com revisão de literatura a respeito da temática de big data. O terceiro caso faz referência aos artigos que levantam alguma questão relacionada aos Princípios Fundamentais das Estatísticas Oficiais (UN, 2014), como por exemplo a questão da confidencialidade. Já a abordagem “qualidade de dados” se refere aos artigos que fazem uma discussão a respeito das propriedades das informações trabalhadas.

Em relação aos temas ou tópicos, as categorias foram pré-definidas como: agricultura, covid-19, crime e corrupção, desigualdade, educação, energia, gênero, meio ambiente, mercado de trabalho e rendimento, migração, mobilidade, ods, outros tópicos demográficos, outros tópicos econômicos, pobreza, publicidade, redução de risco de desastres, saúde e doença, segurança alimentar e fome, transporte, turismo, uso de terra, vendas e, por fim, sem tópico. Cada artigo podia ser classificado em tantas categorias quanto fossem seus temas.

CARACTERÍSTICAS GERAIS DOS ARTIGOS PUBLICADOS A PARTIR DE 2013, EM INGLÊS, PORTUGUÊS OU ESPANHOL, COM ACESSO GRATUITO

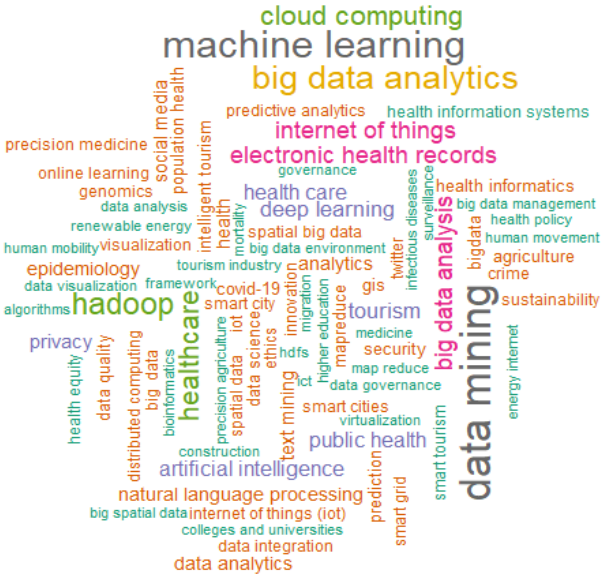
Com base no levantamento amostral realizado, estimou-se que, os artigos, em sua maioria, foram publicados em revistas acadêmicas (69,23%) ou em anais de eventos ou congressos (22,1%). Os demais foram publicados em livros ou outros canais como repositórios de universidade ou de institutos nacionais de estatística. O número mediano de páginas foi igual a 8 e, no tocante à linguagem, 99,4% dos artigos foram publicados em inglês.

Em relação às palavras-chaves, podemos dizer que elas são um bom indicador temático do material, uma vez que elas são escolhidas pelos próprios autores para representar os eixos centrais de seus artigos. Na Figura 1 é possível observar uma nuvem de palavras onde o tamanho dos termos são proporcionais à frequência dos mesmos.

Para construir a nuvem de palavras, termos com grafias semelhantes e mesma semântica foram agregados em um único termo. Ademais, o termo big data foi omitido uma vez que, considerando a forma como as chaves de busca foram definidas, todos os documentos da população alvo têm esse termo no título.

É possível observar que termos computacionais como *Data mining, Internet of things, Cloud computing*, dentre outros, se destacam. No entanto, palavras-chave de temas associados com estatísticas públicas como *Healh care, Tourism* e *Sustantability* também se sobressaem.

Figura 1: Nuvem de Palavras-chaves



Fonte: Elaboração das/os autoras/es.

Para a análise dos tipos de abordagem por artigo, as categorias usadas na classificação foram reagrupadas em: metodológica, conceitual, aplicação com dados reais ou sintéticos, levantamento de vantagens e desvantagens, revisão de literatura (revisão bibliográfica ou pesquisa bibliográfica), boas práticas e qualidade de dados.

A abordagem metodológica apresentou maior frequência relativa (58%), seguido da conceitual (44,5%), aplicação com dados reais ou sintéticos (39,5%), revisão de literatura (19,5%), boas práticas (10,4%) e qualidade de dados (5,5%). A maior incidência da abordagem metodológica indica um esforço da recente literatura em tornar os métodos do uso e tratamento de big data mais conhecido. Apesar da abordagem conceitual ainda apresentar uma alta incidência, sua frequência é menor em relação à metodológica, o que aponta que o conceito de big data, apesar de recente, já encontra solidez no ramo acadêmico (Gráfico 3).

Por outro lado, discussões sobre boas práticas e qualidade de dados ainda são minoritárias na academia. No entanto, a aprovação da Lei Geral de Proteção de Dados Pessoais (LGPD) em 2018 e o crescente interesse dos Institutos Nacionais de Estatística sobre o uso de big data para produção de estatísticas oficiais indicam um potencial de crescimento de artigos que tratem dessas abordagens nos próximos anos.

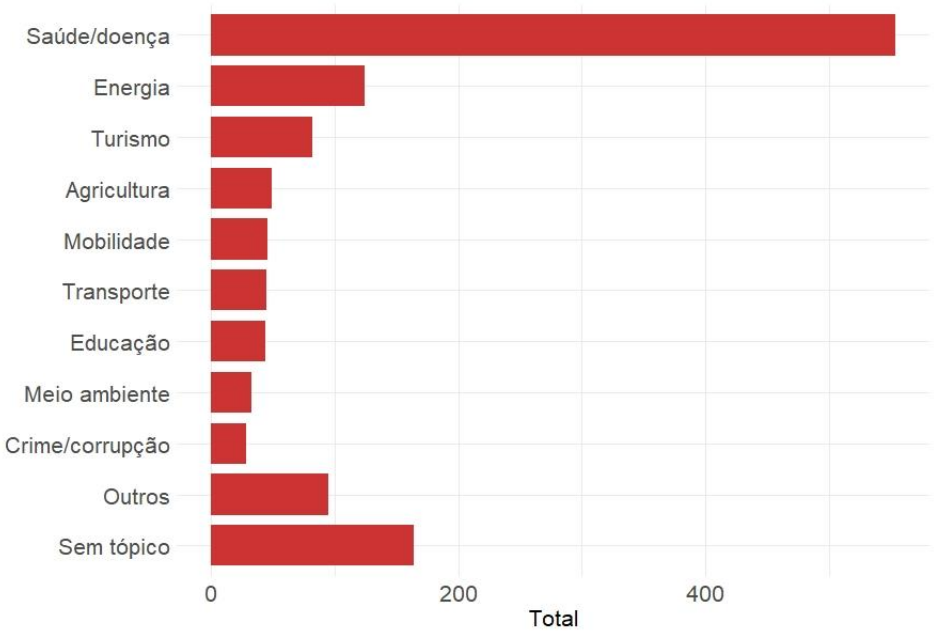
Gráfico 3: Frequência relativa de artigos sobre uso de big data, por tipo de abordagem, 2013 – 2020

Fonte: Elaboração das/os autoras/es.

Sabendo-se que as boas práticas levam em conta questões relativas à geração de estatísticas com elevada qualidade, tem-se que 26,9% dos artigos que abordam qualidade dos dados, abordam também boas práticas. Por outro lado, 14% do total de artigos que abordam boas práticas também abordam qualidade de dados. Nesse sentido, é possível inferir que 86% dos artigos cuja abordagem é boas práticas, acionam outras dimensões que não as relativas diretamente à qualidade dos dados.

Em relação ao tema ou tópico dos artigos, verificou-se que 95,4% deles diziam respeito a apenas um tópico. Em valores absolutos, o tópico saúde e doença, que foi agregado com COVID-19, obteve destaque com 554 artigos, seguidos de energia com 124, turismo com 82, agricultura com 49, mobilidade com 46, transporte com 45, educação com 44, meio ambiente com 33 e crime/corrupção com 28. Artigos sem tópico específico somaram 164 e os outros tópicos somaram 95 artigos (Gráfico 4). Em termos percentuais, o tópico saúde e doença representou cerca de 45,7%, artigos sem tópico representaram quase 13,5% do total e outros tópicos atingiu 7,8% dos artigos.

Gráfico 4 – Total de artigos sobre uso de big data em estatística por tema, 2013 a 2020



Fonte: Elaboração das/os autoras/es.

CONSIDERAÇÕES FINAIS

O presente trabalho apresenta os resultados de uma pesquisa bibliográfica realizada em duas etapas. Na primeira delas, criou-se um cadastro com a produção científica disponível no Google Scholar, cujo título continha o termo big data e um tópico considerado de relevância para a a produção de estatísticas, definido com base em MacFeely (2019). Utilizando a técnica de web scraping implementada no programa Publish Or Perish (HARZING, 2007), foram listados 5.042 documentos (artigos, teses de doutorado, dissertações de mestrado, entre outros) no período de referência de 2004 a 2020. Esse cadastro possibilitou a análise da evolução temporal das publicações sobre big data, mostrando que, enquanto no período de 2004 a 2011 a produção não superou 4 documentos por ano, em 2020 ela foi igual a 906.

Na segunda etapa, para uma análise mais detalhada da produção e, consequentemente, a construção de um banco de metadados sobre o tema, foi realizado um levantamento amostral probabilístico, tendo como cadastro de seleção, a lista de documentos gerada da primeira etapa da pesquisa. A população alvo foi definida como artigos publicados entre 2013 e 2020, em português, espanhol ou inglês, que abordavam a produção de estatística e com acesso livre.

Ao todo, 819 artigos foram investigados. Com base em sua leitura parcial ou total, foram coletadas informações como o tipo de publicação do artigo, a língua utilizada no texto, as palavras-chaves e o número de páginas. Além disso, os artigos foram classificados quanto às abordagens e aos tópicos/temas utilizados.

Com base neste levantamento, estimou-se que 1211 dos documentos disponíveis no Google Scholar, contendo os termos big data e um dos tópicos definidos pelos pesquisadores, e publicados entre 2013 e 2020, são artigos de livre acesso, escritos em português, espanhol ou inglês, que abordam a produção

estatística. Desses, 99,4% foram escritos em inglês. Além disso, 69,1% foram publicados em revistas acadêmicas. A metade deles (50,0%) possui até 8 páginas.

A análise das palavras chaves mostrou que termos computacionais como *Data Mining, Machining learning, Internet of things* e *Cloud computing* foram bastante frequentes. Entre os termos mais associados a estatísticas públicas, *Healh care* foi o que apresentou o maior destaque. Em relação à abordagem, 58,0% dos artigos discutiram metodologia. Quanto ao tema, 45,7% artigos trataram de saúde.

As informações coletadas no levantamento, assim como os pesos amostrais, são disponibilizadas ao público mediante solicitação via correio eletrônico. Essa base de dados e metadados permite que pesquisadores, estudantes e demais usuários de big data possam realizar uma busca de artigos mais direcionada que a propiciada pelo Google Scholar, uma vez que contém outras informações, além daquelas fornecidas pelo indexador. Por exemplo, usuários interessados em boas práticas no uso de big data ou na sua utilização em estudos de mobilidade podem facilmente localizar artigos, eliminando aqueles considerados muito pequenos ou cujo tipo de publicação que não seja de seu interesse.

Uma limitação da base de metadados construída é que ela reflete o acervo do Google Scholar em 2021. Após esse período, algumas mudanças podem ter ocorrido como, por exemplo, artigos que não eram de livre acesso podem ter passado a ser ou vice-e-versa, o endereço eletrônico dos artigos pode ter sofrido modificações, entre outras variações. Ademais, considerando o rápido crescimento no número de artigos e avanços e novidades sobre o tema big data, a atualização da base deve ser realizada com intervalos de tempo não muito elevados.

Estudos futuros podem incluir, além da atualização da referência temporal, uma revisão do termo *environment*, utilizado, juntamente com big data, como termo de busca no título dos documentos disponíveis no indexador. A adequação de outros termos como *environmental* pode ser testada no intuito de captar um volume maior de documentos sobre o tema meio ambiente.

Por fim, cabe destacar que o tema big data ainda é muito pouco utilizado na produção de estatísticas públicas. Dessa forma, o presente estudo é um importante passo para a difusão e consolidação dos conhecimentos sobre o uso dessa fonte de dados, assim como para a geração de novas ideias e avanços na sua implementação.

Study on the use of big data in Statistical Production

ABSTRACT

The need for high quality public statistics is increasingly pressing in our society, while the resources for their production are increasingly scarce. Against this backdrop, national institutes of statistics are increasingly investing in alternative data sources, including big data. The present work aims to reveal the academic production around the theme from a probabilistic sample of articles, selected from a framework extracted from Google Scholar. The results allow estimating that, from 2013 to 2020, 1211 free articles, written in Portuguese, Spanish or English, that studied the use of big data in the production of statistics for selected topics were made available on this platform. Of these, 554 addressed the topic of health. Observing the general characteristics of the articles allowed the construction of a metadata base capable of contributing to the discussion of the limits and potential of big data in the production of statistics.

KEYWORDS: Big data. Public statistics. Official statistics. Bibliographic research.

NOTAS

[1] ODS 1: Acabar com a pobreza em todas as suas formas, em todos os lugares. ODS 2: Acabar com a fome, alcançar a segurança alimentar e melhoria da nutrição e promover a agricultura sustentável. ODS 5: Alcançar a igualdade de gênero e empoderar todas as mulheres e meninas. ODS 6: Assegurar a disponibilidade e gestão sustentável da água e o saneamento para todos. ODS 11: Tornar as cidades e os assentamentos humanos inclusivos, seguros, resilientes e sustentáveis. ODS 14: Conservar e usar sustentavelmente os oceanos, os mares e os recursos marinhos para o desenvolvimento sustentável. ODS 15: Proteger, recuperar e promover o uso sustentável dos ecossistemas terrestres, gerir de forma sustentável as florestas, combater a desertificação, deter e reverter a degradação da terra e deter a perda de biodiversidade.

AGRADECIMENTOS

Autoras e autores agradecem a todos que contribuíram com essa pesquisa. Em especial, o quarto e o quinto autores agradecem ao IBGE pela bolsa de Iniciação Científica concedida, assim como a sexta e a sétima autoras agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES.

REFERÊNCIAS

Aminer. ArnetMiner, 2022. Disponível em: <https://www.aminer.org>. Acesso em: 05/07/2022.

BASE - Bielefeld Academic Search Engine, 2022. Disponível em: <https://www.base-search.net>. Acesso em: 05/07/2022.

BDTD - Biblioteca Digital Brasileira de Teses e Dissertações, 2022. Disponível em: <https://bdtd.ibict.br/vufind/>. Acesso em: 05/07/2022.

CORE, 2022. Disponível em: <https://core.ac.uk>. Acesso em: 05/07/2022.

DOAJ - Directory of Open Access Journals, 2022. Disponível em: <https://doaj.org>. Acesso em: 05/07/2022.

ERIC - Education Resources Information Center, 2022. Disponível em: <https://eric.ed.gov>. Acesso em: 05/07/2022.

Google Scholar, 2022. Disponível em: <https://scholar.google.com.br>. Acesso em: 05/07/2022.

GUSENBAUER, MICHAEL. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. **Scientometrics.** Vol. 118, 177–214. 2019. Disponível em: <https://doi.org/10.1007/s11192-018-2958-5>. Acesso em: 23/06/2022.

HALDANE, J. B. S. On a method of estimating frequencies. Biometrika. Vol 33, 222–225. 1945.

HARZING, A.W. Publish or Perish, 2007. Disponível em: <https://harzing.com/resources/publish-or-perish>. Acesso em: 23/06/2022.

JSTOR - Journal Storage, 2022. Disponível em: <https://www.jstor.org>. Acesso em: 05/07/2022.

MacFeely, Steve. The Big (data) Bang: Opportunities and Challenges for Compiling SDG Indicators. **Global Policy**. Vol 10, 121-133. 2019. Disponível em: <https://onlinelibrary.wiley.com/doi/epdf/10.1111/1758-5899.12595>. Acesso em: 23/06/2022.

Microsoft Academic, 2021. Disponível em: <https://www.microsoft.com/en-us/research/project/academic>. Acesso em: 05/07/2022. Desativado ao final de 2021

Periódicos da CAPES, 2020. Disponível em: <https://www-periodicos-capes-gov-br.ezl.periodicos.capes.gov.br>. Acesso em: 05/07/2022.

RAMÍREZ-CASTEÑEDA, V. Disadvantages in preparing and publishing scientific papers caused by the dominance of the English language in science: The case of Colombian researchers in biological sciences. PLOS ONE, 15 (9). 2020. Disponível em: <https://doi.org/10.1371/journal.pone.0238372>. Acesso em: 23/06/2022.

Redalyc - La Red de Revistas Científicas de América Latina y el Caribe, España y Portugal, 2022. Disponível em: <https://www.redalyc.org>. Acesso em: 05/07/2022.

SBU - Sistema de Bibliotecas da UNICAMP, 2022. Disponível em: <http://www.sbu.unicamp.br/sbu>. Acesso em: 05/07/2022.

Scholarpedia, 2016. Disponível em: <http://www.scholarpedia.org/article/Main_Page>. Acesso em: 05/07/2022.

SciELO - Scientific Electronic Library Online, 2021. Disponível em: <https://scielo.org>. Acesso em: 05/07/2022.

Science Research, 2022. Disponível em: <https://www.scienceresearch.com/scienceresearch/mobile/en/search.html>. Acesso em: 05/07/2022.

Science.gov, 2022. Disponível em: <https://www.science.gov>. Acesso em: 05/07/2022.

ScienceDirect, 2022. Disponível em: <https://www.sciencedirect.com>. Acesso em: 05/07/2022.

Semantic Scholar, 2022. Disponível em: <https://www.semanticscholar.org>. Acesso em: 05/07/2022.

The Lens, 2022. Disponível em: <https://www.lens.org>. Acesso em: 05/07/2022.

United Nations. **Fundamental Principles of Official Statistics**. Resolution adopted by the General Assembly on 29 January 2014. Sixty-eighth session. 2014. Disponível em: <https://ilostat.ilo.org/about/data-collection-and-production/fundamental-principles-of-official-statistics/#:~:text=Fundamental%20Principles%20of%20Official%20Statistics%201%201.%20Relevance%2C,...%208%208.%20National%20coordination%20...%20Mais%20itens>. Acesso em: 05/07/2022.

WorldWideScience, 2021. Disponível em: <https://worldwidescience.org>. Acesso em: 05/07/2022.

Recebido: 28/11/2022
Aprovado: 28/11/2023
DOI: 10.3895/rts.v20n59.16167

Como citar:

BELO HYPOLITO, Elizabeth; DINIZ DA SILVA, Andrea; XAVIER, Antonia et al. Estudo sobre o uso de big data na produção estatística.
Tecnol. Soc., Curitiba, v. 20, n. 59, p.160-177, jan./abr., 2024. Disponível em:

https://periodicos.utfpr.edu.br/rts/article/view/16167
Acesso em: XXX.

Correspondência: