Uma descrição dos procedimentos tecnológicos para a identificação da mobilidade acadêmica brasileira

Higor Alexandre Duarte Mascarenhas
CEFET-MG
Divinópolis, Brasil
https://orcid.org/0000-0001-6057-3888

Thiago Magela Rodrigues Dias
CEFET-MG
Divinópolis, Brasil
https://orcid.org/ 4687858846001290

***Abstract*— In recent years, a fact that stands out in other locations on the planet is the displacement of individuals to other locations at some point in their lives. There are several causes that motivate displacement, among them, one of the main reasons for training at a high level of academic education. Given this scenario, this work aims to carry out an analysis of how Brazilian academic mobility takes place, through data extracted from the Lattes Platform. After extracting the data, the extracted data were filtered and then processed, removing irrelevant and incomplete terms and, finally, improving the data with information on the geographic location of each analyzed institution. As a result, it was possible to obtain a picture of how the process of scientific training in Brazil takes place, making it possible to measure the migratory flow of individuals and trends in training processes.**

***Resumo* — Nos últimos anos, um fato que se destaca em demais localidades do planeta é o deslocamento de indivíduos para outras localidades em algum momento de suas vidas. São várias as causas que motivam o deslocamento, entre elas, um dos principais motivos de capacitação em alto nível da formação acadêmica. Diante desse cenário, este trabalho tem como objetivo realizar uma análise de como ocorre a mobilidade acadêmica brasileira, por meio de dados extraídos da Plataforma Lattes. A partir da extração dos dados, ocorreu uma filtragem dos dados, e após foi efetuado um tratamento dos dados extraídos, retirando termos irrelevantes e incompletos e, finalmente, aprimorando os dados com informações sobre a localização geográfica de cada instituição analisada. Como resultado, foi possível obter um retrato de como ocorre o processo de formação científica brasileira, possibilitando mensurar o fluxo migratório dos indivíduos e as tendências nos processos de formação.**

***Palavras-chave*—: Plataforma Lattes; Êxodo Científico Brasileiro; Análise de Dados.**

I. INTRODUÇÃO

A emigração de brasileiros para outros países e para outros estados tem aumentado de forma significativa, de modo que, no Brasil, estudos demonstram que em algumas cidades possuem taxa de 10 a 30% de migrantes que não vivem no seu estado de origem [1]. Em muitos casos, brasileiros saem em busca de emprego, ou estudos, visando sempre qualidade de vida.

Dentre os principais motivos para migração está a necessidade de capacitação em alto nível de formação. Das principais causas por optar pela mobilidade de indivíduos no território brasileiro, refere-se à qualidade de ensino superior em outros estados, a busca de novas oportunidades e mais experiências em suas áreas [2]. Outro refúgio para tais estudantes condiz na ida para outros países, buscando assim, intercâmbio cultural e melhor investimento em bolsas de pesquisa. A saída do estudante para outros países não é interessante somente ao discente, mas também às instituições de origem, pois, o estudante retorna na maioria das vezes mais produtivo, com rede de contato mais extensa, maior vivência, e podendo futuramente compartilhar suas experiências com outros estudantes da instituição de origem.

De acordo com [3] a cada dia tem se tornado mais difícil produzir pesquisa científica no Brasil, devido a cortes de investimentos destinados a bolsas. Um dos principais motivos para a emigração de pesquisadores brasileiros para outros países pode ser apontado pela falta de apoio do governo. Logo, com esse cenário, pesquisadores brasileiros saem do país, dificultando assim o retorno pela falta de oportunidades. Grande parte dos cientistas brasileiros que voltam para o Brasil não conseguem emprego na sua área de formação, fazendo assim que não progridam nas suas carreiras.

Um programa que facilitou e auxiliou bastante o ingresso de estudante às instituições com sedes em outros países foi o Ciência Sem Fronteiras, por se referir a um programa que amparou estudantes, oferecendo bolsas de estudos. Em 2015, o governo pretendia alcançar 101.000 bolsas de estudos para pesquisadores, graduandos, doutorandos, alunos ingressados no pós-doutorado, incentivando os discentes a se capacitarem em instituições de reconhecida relevância [4]. Recentemente, o Programa perdeu bastante influência no ingresso de estudantes para outros países, por motivos de cortes de investimento.

Diante deste cenário, este trabalho apresentará um estudo sobre o êxodo de estudantes brasileiros que partiram do seu estado/cidade de nascimento para outros estados/cidades e/ou aqueles que foram para outros países em busca de capacitação. Para obtenção dos dados dos estudantes brasileiros analisados neste estudo será utilizado o *framework LattesDataXplorer* [5], ferramenta responsável por extrair e tratar currículos de indivíduos cadastrados na Plataforma Lattes. Atualmente, o repositório de currículos da Plataforma Lattes, que registra

informações acadêmicas/científicas e profissionais, possui 7.200.000 currículos cadastrados. Um conjunto de componentes desenvolvidos para os propósitos deste estudo foram elaborados e incorporados ao *framework*, viabilizando dessa forma uma visão ampla e inédita sobre o êxodo científico brasileiro.

II. METODOLOGIA

A. *Aquisição dos dados*

Como principal fonte de dados foi utilizado o repositório curricular da Plataforma Lattes. A justificativa da escolha da Plataforma se dá por: (1) registrar a trajetória e a contribuição de cada estudante, técnico e pesquisador brasileiros cadastrados [6]; (2) os dados estarem disponíveis na internet [7]; (3) representa a experiência do CNPq na integração de base de dados de currículos e de instituições da área de Ciência e tecnologia [8]; (4) por se tratar de uma importante fonte de dados de alta qualidade com o intuito de medir e avaliar o desempenho acadêmico nacional [9].

Apesar da reconhecida relevância dos currículos cadastrados na Plataforma Lattes para análise e entendimento sobre a evolução da ciência brasileira conforme descrito anteriormente, o acesso ao repositório de dados passa a ser um fator limitante para análises que considerem todos os indivíduos independentemente de suas áreas de atuação, ou nível de formação acadêmica. Apesar da viabilidade de acesso individual a cada um dos currículos ser possível através de interface de consulta dos currículos da Plataforma Lattes, a análise de grandes grupos de indivíduos passa a ser um fator limitante para análises abrangentes. Logo, no contexto deste trabalho, para extração de todo o conjunto de currículos a serem analisados, foi utilizado o *LattesDataXplorer* [5] para extração e tratamento dos dados.

A extração dos dados foi realizada em maio de 2019 totalizando 308.317 currículos de indivíduos com doutorado concluído. O *framework* utilizado é responsável por conter uma coleção de componentes que visam realizar a coleta e tratamento dos dados, conforme ilustra a Figura 1.

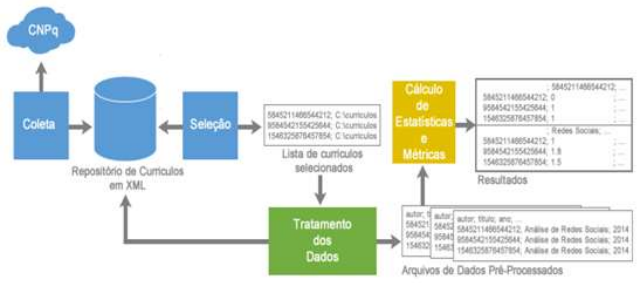

Fig. 1. Visão geral do *LattesDataXplorer*. Fonte: Dias (2016).

Utilizando o *framework* é possível fazer um refinamento da busca de currículos específicos, baseado em parâmetros como nome, titulação, idioma, nacionalidade, grande área e área de atuação, dentre outros. Sendo gerado assim listas de currículos que atendem aos parâmetros informados.

Todo o procedimento de extração e tratamento de dados realizado pelo *LattesDataXplorer* se inicia a partir da aquisição dos códigos de currículos da Plataforma Lattes, baseado no refinamento executado na busca, com o propósito de no futuro estes códigos sejam armazenados localmente (Figura 2). A lista resultante da consulta refinada, possui todos os códigos de identificação de todos os currículos cadastrados, possibilitando ter o acesso individual em cada um destes existentes na Plataforma Lattes.

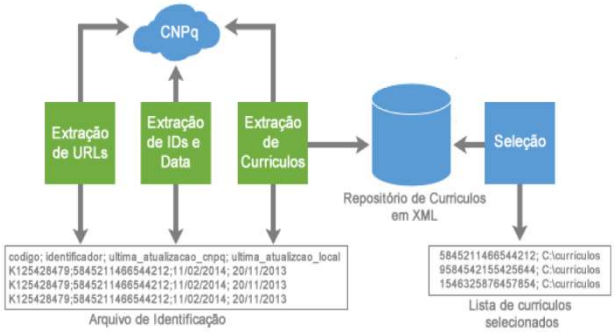

Fig. 2. Componente de coletor do *LattesDataXplorer*. Fonte: Dias (2016).

Todos currículos possuem uma URL (*Uniform Resource Locator*) a fim de permitir o acesso individual a cada um destes. A URL possui o endereço de busca da plataforma concatenada com o código do currículo do indivíduo. O componente de extração de Ids e Data tem a responsabilidade de acessar cada um dos currículos cujos códigos foram salvos, extraindo seu identificador único e a data da última atualização do mesmo. Para esse fim, somente o cabeçalho de cada currículo em que estão presentes estas informações necessitam ser carregados. Tais dados são armazenados em um arquivo de identificação contendo código, identificador, data da última atualização no CNPq e data da atualização do currículo armazenado localmente.

Para a atualização do repositório de currículos é necessário o arquivo de identificação, por se tratar da base para extração do mesmo. Durante a atualização o primeiro componente do processo de extração é executado, resultando na extração de todos os códigos cadastrados na plataforma. Os códigos já registrados no arquivo de identificação são ignorados, e no final do arquivo são adicionados aqueles novos códigos, representantes dos novos currículos ainda não extraídos.

Com o uso dos códigos, são acessados os cabeçalhos de cada um dos currículos sendo extraídos códigos identificadores e as datas de atualização junto à Plataforma Lattes, tanto para currículos já extraídos, como para os novos

currículos, ocorrendo a atualização do arquivo de identificação a cada nova extração. O acesso ao cabeçalho torna mais ágil a extração de dados, por não haver a necessidade de esperar todo o currículo ser gerado.

Por fim, acontece a extração de currículos, sendo o extrator o responsável por verificar se possuem currículos cuja Data de Atualização Local é diferente da Data de Atualização junto ao CNPq; caso seja divergente, o currículo da Plataforma é extraído e substitui o currículo local, modificando a data de atualização. Quando não, o currículo permanece sem alterações. Por fim são extraídos novos currículos cadastrados, para serem inseridos ao final do arquivo base. Inicialmente esses novos currículos não possuem data de atualização por terem sidos extraídos pela primeira vez, sendo assim, é inserido sua data de atualização local.

Finalmente, todos os currículos são armazenados em XML, não necessitando de um novo repositório de dados feito por bancos de dados relacionais, buscando um menor custo computacional. Vale ressaltar que com todos os currículos armazenados localmente, torna-se mais fácil e flexível a manipulação de dados coletados a partir da Plataforma Lattes.

B. Componentes desenvolvidos para tratamento dos dados

O *LattesDataXplorer* foi utilizado especificamente para a coleta e seleção dos dados curriculares da Plataforma Lattes, no qual obteve-se o todo o Repositório de Currículos em formato XML.

A "Seleção" do conjunto de dados a ser analisado utiliza a linguagem de consulta XPath (*XML Path Language*) para pesquisa e posterior geração dos subgrupos a serem analisados. A linguagem XPath possibilita a construção de expressões que vão processar e percorrer um documento XML de forma similar ao uso de expressões regulares. Portanto, possibilita o agrupamento de um conjunto de currículos com parâmetros desejados. Assim sendo, em busca pelos parâmetros em cada um dos currículos, independentemente ou não de qual seção ele(s) seja(m) encontrado(s), tais currículos são selecionados e formam um grupo para análises. A partir de então os dados dos currículos são organizados em uma lista de currículos que foram selecionados. A lista armazena os identificadores de cada currículo e o caminho que ele está armazenado localmente, sendo assim, será possível analisar somente os currículos selecionados.

Diante do exposto foram coletados somente currículos de indivíduos com doutorado concluído, por se tratar do grupo com o maior nível de formação acadêmica; por se tratar de currículos que são frequentemente atualizados e grande parte dos parâmetros necessários para o presente trabalho estarem registrados em seus currículos.

A fim de mapear o êxodo de indivíduos brasileiros cadastrados na Plataforma Lattes, foi efetuada a mineração de dados para filtrar os dados relevantes para esta pesquisa, logo após os dados serem filtrados ocorreu um tratamento com o intuito de enriquecê-los para as análises a serem realizadas. A Figura 3 apresenta um aspecto geral do conjunto de componentes que foram desenvolvidos objetivando obter as análises desejadas.

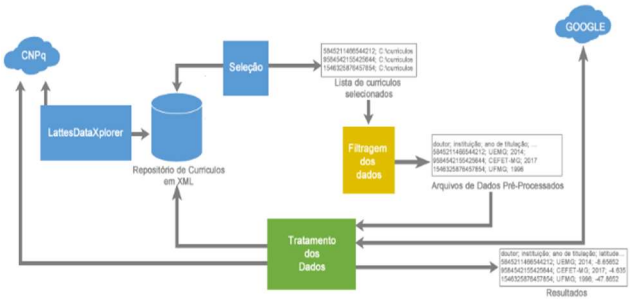

Fig. 3. Aspecto geral do conjunto de componentes utilizados. Fonte: Autores.

Após a Seleção do conjunto a ser analisado, ocorre o módulo “Filtragem dos dados” (Figura 4). Esta fase é responsável por analisar os arquivos XML com o intuito de obter informações relevantes à pesquisa, armazenando-as em um estrato de dados formatados (Arquivos de dados pré-processados). As informações dos currículos presentes no arquivo possuem: identificador do currículo; estado e cidade de nascimento; grande área; área; código, identificador, nome e CEP do vínculo atual de atuação do indivíduo, além do código de identificação, e nome da instituição, início e fim de cada nível de formação acadêmica concluída, juntamente com a instituição em que foi realizada a formação considerando desde a graduação até o doutorado.

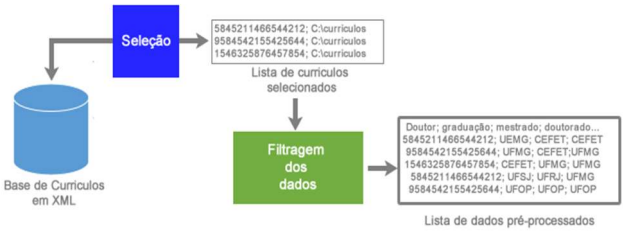

Fig. 4. Filtragem dos dados. Fonte: Autores.

Posteriormente é executado o Módulo de "Tratamento dos Dados" (Figura 5) que tem o intuito de processar os dados dos doutores coletados, tratá-los e caracterizá-los, resultando em outros arquivos, a fim de facilitar as análises dos dados. Nesse processo são realizadas quatro etapas: Obtenção de CEP da instituição; Busca pela localização geográfica; Limpeza e agrupamento de dados e Normalização dos dados.

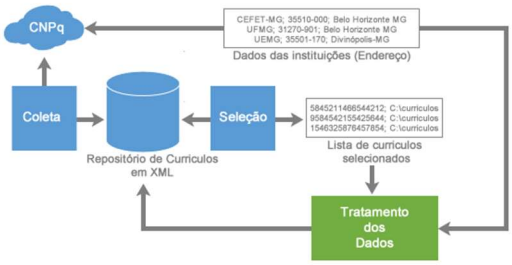

Fig. 5. Tratamento dos dados. Fonte: Autores.

- A primeira etapa efetuada, é a "**Obtenção de CEP da instituição**" em que a partir do código da instituição recuperado no currículo do indivíduo, este é consultado no diretório de instituições da Plataforma Lattes no intuito de obter os dados da instituição e dessa forma, recuperar da seção de endereço o CEP da instituição.
- A etapa de "**Busca pela localização geográfica**" é uma tarefa a ser realizada com a finalidade de obter coordenadas geográficas de uma instituição. Em que ao acessar a API (*Application Programming Interface*) de geolocalização do Google, será enviado o CEP da instituição, para posteriormente ter como retorno a localização geográfica da instituição.
- Na etapa de "**Limpeza e agrupamento de dados**", ocorre a realização de exclusão de possíveis termos irrelevantes para a pesquisa, com o intuito de diminuir o volume de dados a serem processados e analisados. Como exemplo: a remoção de *stopWorlds* nos nomes das cidades; a normalização para extrair palavras acentuadas, e substituí-las pelo seu equivalente sem acentuação.
- Já a etapa de "**Normalização dos dados**" tem o intuito de reduzir a redundância de informações, descartando atributos com a ausência de dados, como por exemplo dados de geolocalização com ausência de algarismos.

Posteriormente é gerado o arquivo de "Resultados" representando uma sumarização de todos os dados obtidos nos currículos de doutores brasileiros, não necessitando de consultas dos arquivos XML dos currículos extraídos, possuindo todos os dados específicos para a realização das análises desta pesquisa.

Logo, após toda a execução das etapas descritas anteriormente, e com os arquivos formatados de tal forma a facilitar as análises, diversas métricas são aplicadas para compreensão sobre como tem ocorrido a mobilidade dos doutores brasileiros ao longo do seu processo de formação acadêmica.

III. RESULTADOS

Inicialmente foi possível realizar uma caracterização do conjunto analisado. Somente foram incluídos na análise as formações cujo status no currículo estejam como "concluído", resultando em um total de 308.317 indivíduos.

No intuito de avaliar o potencial de análise dos dados extraídos, a Figura 6 apresenta a quantidade de currículos que possuem atributos cadastrados, tais como: cidade de nascimento, instituição de atuação e instituições de formação.

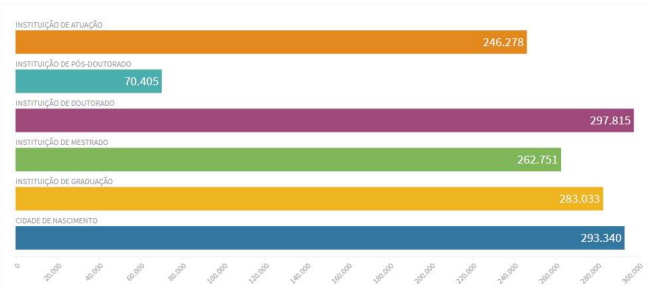

Fig. 6. Quantitativo de currículos que possuem informações que são objetos de análises. Fonte: Autores.

A quantidade de cadastros que apresentam a cidade de nascimento é de 293.340 (95%), por se tratar de um campo obrigatório ao ser realizado o cadastro na Plataforma Lattes. Aqueles indivíduos que não possuem cadastro de cidade de nascimento julgam-se que são currículos mais antigos, em que não era obrigatório o cadastramento da cidade. Dos demais campos demonstrados no gráfico, aquele que se identifica superior em dados quantitativos é a *instituição de conclusão do doutorado* totalizando uma quantidade de 297.815 (96%) cadastros, por se tratar do grupo selecionado para a realização do estudo. A *instituição de pós-doutorado* é aquela que possui a menor quantidade de cadastros, com um montante de 70.405 (22%), tal fato se justifica, pois, a seleção dos grupos foi direcionada aos indivíduos com doutorado concluído; por este motivo, nem sempre um indivíduo que possui o título de doutorado possui o pós-doutorado. Foi possível apresentar na Tabela 1 o cálculo das distâncias médias em quilômetros entre um nível de formação até outro.

TABELA I
Distância média em Km entre os níveis de formação. Fonte: Autores.

Distância(Km)	Graduação	Mestrado	Doutorado
Nascimento	291,58	548,59	1.000,75
Graduação	-	432,38	901,26

Mestrado	-	-	619,00

É possível observar o resultado da distância média de todas as etapas de capacitação dos doutores brasileiros, durante sua formação acadêmica. Pode ser observado que a distância média entre as etapas tem uma variação significativa. Inicialmente, analisando a distância média do local de nascimento para graduação, percebe-se que esta é a menor distância média calculada. Um dos fatores que influenciam tal fenômeno é que grande parte das cidades brasileiras possuem instituições que proporcionam ao estudante cursos de graduação, e aquelas que não possuem na maioria das vezes, ficam próximas a outras cidades que ofertam os cursos neste nível de capacitação de interesse dos estudantes. Já as maiores distâncias estão entre o local de nascimento e de capacitação a nível de doutorado, seguindo da graduação/doutorado em que o deslocamento é maior que os outros níveis de formação. Destaca-se que o valor médio da díade graduação-doutorado é influenciado por um quantitativo de indivíduos que realizam seus doutorados no exterior, cujas distâncias são mais representativas.

Foi possível também obter uma visualização do percurso intraestadual e interestadual percorrido pelos doutores brasileiros em seus processos de formação acadêmica (Figura 7).

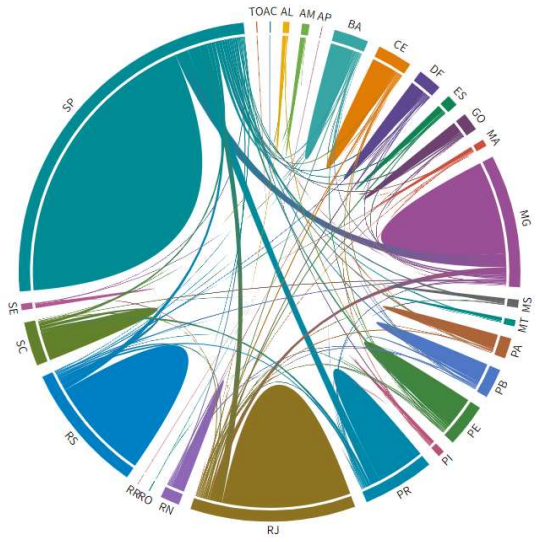

Fig. 7. Fluxos a nível estadual percorridos pelos doutores. Fonte: Autores.

É possível observar o quanto estados como São Paulo, Rio de Janeiro, Minas Gerais e Rio Grande do Sul se destacam por possuir mais caminhos percorridos dentro de seus próprios territórios em detrimento de outros estados brasileiros. Um dos principais fatos a que se pode justificar tal movimento é que são os quatro estados que possuem maior quantidade de universidades públicas no país, concentrando ainda a maioria dos Programas de doutorado.

Destaca-se o quanto os estados com maiores quantidades interestaduais se interagem, principalmente na emigração de Rio de Janeiro, Minas Gerais e Paraná para a imigração no estado de São Paulo que representa o estado com maior número de doutores atuando.

Ao observar os movimentos interestaduais, também se realça aqueles indivíduos que saem do estado Rio Grande do Sul para São Paulo e reciprocamente aqueles que saem de São Paulo para o estado do Paraná, caracterizando uma ampla rede de colaboradores no processo de formação acadêmica.

Ressalta-se também que todos os estados possuem vínculo com os outros estados do país, apesar de alguns em menores quantidades, como o Acre, Roraima, Rondônia e Amapá. Sendo esses os estados menos representativos, já que também se destacam por possuir quantidades menos representativas de indivíduos que nasceram neles.

Foi possível ainda identificar aqueles doutores que fizeram o doutorado no exterior e que retornaram para o Brasil, destacando as cidades que os indivíduos mais optaram por retornar (Figura 8).

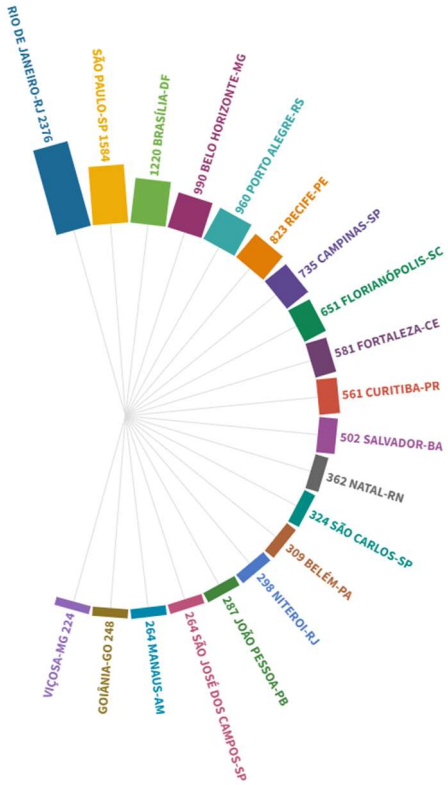

Fig. 8. Cidades de atuação profissional de quem realizou o doutorado no exterior. Fonte: Autores.

Percebe-se o quão salientam as capitais brasileiras neste gráfico, sendo representadas pelas localizações com maiores quantidades de imigrações para o Brasil na atuação profissional. Destaca-se que a cidade de Rio de Janeiro é mais influente que a cidade de São Paulo se tratando destas condições.

Observa-se algumas cidades que não correspondem a capitais brasileiras presentes na Figura 8, como a cidade de Campinas (SP), São Carlos (SP), Niterói (RJ), São José Dos Campos (SP) e Viçosa (MG), todas estas da região Sudeste do Brasil possuem grande representatividade.

Importante destacar que a maioria das cidades apresentadas possuem grande representatividade de instituições de ensino e pesquisa contemplando grandes universidades públicas do Brasil.

Logo, percebe-se que no Brasil a localização geográfica possui forte influência sobre o processo migratório para capacitação. Verifica-se que os estudantes tendem a percorrer menores distâncias em seu processo de formação quando estão em certas regiões, como sudeste e sul do país, em detrimento de outras, como por exemplo da região norte cuja oferta de cursos de capacitação principalmente a nível de pós-graduação é menor.

IV. Conclusão

A partir dos resultados obtidos foi possível verificar a viabilidade em adotar os currículos cadastrados na Plataforma Lattes como fonte de dados para análises sobre como ocorre o Êxodo Científico brasileiro.

A escolha do grupo de doutores se caracteriza como uma parcela significativa de todo o conjunto de dados cadastrados na Plataforma Lattes, tendo em vista que são os indivíduos com maior nível de formação acadêmica concluída. Percebeu-se também que em geral, tais currículos são recentemente atualizados e a maioria possui endereço profissional cadastrado.

Ficou nítido como a região sudeste concentra a grande maioria dos doutores brasileiros, fato este influenciado diretamente pela concentração das principais universidades públicas do país, e de uma maior oferta de oportunidades para atuação profissional.

Identificou-se a distância média percorrida pelos indivíduos ao longo de sua formação acadêmica, além de identificar a mediana das distâncias entre os níveis de formação, sendo observado que em média as distâncias percorridas foram pequenas, e que muitas das vezes os indivíduos optaram em se capacitar na mesma instituição, principalmente na díade mestrado-doutorado.

Percebe-se que as capitais brasileiras são influentes no processo de capacitação de indivíduos brasileiros, uma vez que possuem maiores quantidades de universidades federais em que estão a maioria dos programas de pós-graduação e em grande parte possuem melhores ofertas de emprego comparando as cidades do interior.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) e do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG).

Referências

[1] Almeida GCR. Fluxos migratórios: a distribuição da população de cada estado pelo país [Internet]. Nexo Jornal. 2017 [citado 19 de julho de 2021]. Disponível em: https://www.nexojornal.com.br/grafico/2017/12/01/Fluxos-migrat%C3%B3rios-a-distribui%C3%A7%C3%A3o-da-popula%C3%A7%C3%A3o-de-cada-estado-pelo-pa%C3%ADs

[2] Lombas ML de S. A mobilidade internacional acadêmica: características dos percursos de pesquisadores brasileiros. Sociologias. 2017;19(44):308–33.

[3] Demartini M. Falta de oportunidades mantém cientistas brasileiros no exterior [Internet]. Exame. 2017 [citado 14 de abril de 2021]. Disponível em: https://exame.com/ciencia/falta-de-oportunidades-mantem-cientistas-brasileiros-no-exterior/

[4] Aveiro TMM. O programa Ciência sem Fronteiras como ferramenta de acesso à mobilidade internacional. #Tear [Internet]. 15 de dezembro de 2014 [citado 23 de julho de 2020];3(2). Disponível em: https://periodicos.ifrs.edu.br/index.php/tear/article/view/1867

[5] Dias TMR. Um estudo da produção científica brasileira a partir de dados da Plataforma Lattes. 2016 [Doutorado em Modelagem Matemática e Computacional-Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte]. [Belo Horizonte]: CEFET-MG; 2016.

[6] Marques F. Registros valiosos [Internet]. Pesquisa Fapesp. 2015 [citado 21 de julho de 2021]. Disponível em: https://revistapesquisa.fapesp.br/registros-valiosos/

[7] Digiampietri LA, Mena-Chalco JP, Pérez-Alcazar JJ, Tuesta EF, Delgado KV, Mugnaini R, et al. Minerando e Caracterizando Dados de Currículos Lattes. In: Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM) [Internet]. SBC; 2012 [citado 19 de julho de 2021]. Disponível em: https://sol.sbc.org.br/index.php/brasnam/article/view/6868

[8] Silva FM, Smit JW. Organização da informação em sistemas eletrônicos abertos de Informação Científica & Tecnológica: análise da Plataforma Lattes. Perspect ciênc inf. abril de 2009;14:77–98.

[9] Lane J. Let's make science metrics more scientific. Nature. março de 2010;464(7288):488–9.