Universidade Federal da Paraíba - UFPB

Centro de Informática
Programa de Pós-Graduação em Informática

Luiz Carlos Rodrigues Chaves

Analisando a mobilidade de pesquisadores através de registros curriculares na Plataforma Lattes

João Pessoa - PB

Fevereiro/2016

Luiz Carlos Rodrigues Chaves

Analisando a mobilidade de pesquisadores através de registros curriculares na Plataforma Lattes

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Informática da Universidade Federal da Paraíba como parte dos requisitos necessários para obtenção do grau de Mestre em Informática.

Orientador: Alexandre Nóbrega Duarte

João Pessoa - PB

Fevereiro/2016

C512a Chaves, Luiz Carlos Rodrigues.
Analisando a mobilidade de pesquisadores através de registros curriculares na Plataforma Lattes / Luiz Carlos Rodrigues Chaves.- João Pessoa, 2016.
112f. : il.
Orientador: Alexandre Nóbrega Duarte
Dissertação (Mestrado) - UFPB/CI
1. Informática. 2. Computação distribuída. 3. Mobilidade de pesquisadores. 4. Análise de currículo. 5. Mineração de dados. 6. Visualização de dados.

UFPB/BC CDU: 004(043)

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de **LUIZ CARLOS RODRIGUES CHAVES**, candidato ao título de Mestre em Informática na Área de Sistemas de Computação, realizada em 22 de fevereiro de 2016.

Ao décimo segundo dia do mês de fevereiro do ano de dois mil e dezesseis, às dezessete horas, no Centro de Informática - Universidade Federal da Paraíba (unidade Mangabeira), reuniram-se os membros da Banca Examinadora constituída para julgar o Trabalho Final do **Sr. Luiz Carlos Rodrigues Chaves** vinculado a esta Universidade sob a matrícula 2014108028, candidato ao grau de Mestre em Informática, na área de "*Sistemas de Computação*", na linha de pesquisa *"Sinais, Sistemas Digitais e Gráficos"*, do Programa de Pós-Graduação em Informática, da Universidade Federal da Paraíba. A comissão examinadora foi composta pelos professores: **Dr Alexandre Nobrega Duarte (PPGI-UFPB),** Orientador e Presidente da Banca, **Dr Alisson Vasconcelos de Brito (UFPB),** Examinador Interno e **Dr Francisco Petronio Alencar de Medeiros (IFPB)**, Examinador Externo à Instituição. Dando início aos trabalhos, o professor Presidente da Banca cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou a palavra ao candidato para que o mesmo fizesse, oralmente, a exposição do trabalho de dissertação intitulado *"Mineração e Análise de Localidade da Plataforma de Curriculos Lattes"*. Concluída a exposição, o candidato foi arguido pela Banca Examinadora que emitiu o seguinte parecer: "*aprovado*". Assim sendo, eu, Clauirton de Albuquerque Siebra, Coordenador do Programa de Pós-Graduação em Informática - PPGI, lavrei a presente ata que vai assinada por mim e pelos membros da Banca Examinadora.

João Pessoa, 22 de fevereiro de 2016.

Clauirton de Albuquerque Siebra

Prof Dr Alexandre Nobrega Duarte
Orientador (PPGI-UFPB)

Prof Dr Alisson Vasconcelos de Brito
Examinador Interno (PPGI-UFPB)

Prof Dr Francisco Petronio Alencar de Medeiros
Examinador Externo à Instituição (IFPB)

Dedico este trabalho aos meus pais Luiz Gonzaga Chaves Tito
e Maria Solange Rodrigues Chaves e aos meus irmãos Natan Rodrigues Chaves,
Caroline Rodrigues Chaves e Juliana Rodrigues Chaves pelo amor,
confiança e apoio nos momentos difíceis. E a Poliane Karenine
Batista pelo apoio, compreensão e amor em todos os momentos.

Agradecimentos

A Deus por me ajudar a superar as barreiras da vida, e por me mostrar que sempre existe um caminho de sabedoria e amor a ser trilhado.

A todos os meus familiares por estarem sempre ao lado e ao qual amo muito, pelo carinho, paciência e incentivo.

Ao professor Dr. Alexandre Nóbrega Duarte pela honra de ter me concedido a orientação, pelos relevantes comentários que ajudaram a produzir este trabalho, pela gentileza e compreensão diante de minhas limitações, por toda disponibilidade e dedicação demonstrada durante a realização desse trabalho. De fato você foi para mim o Orientador.

Aos professores da PPGI pela orientação, colaboração, ensinamentos e auxílio na concretização desse trabalho.

Aos colegas de mestrado pelo auxílio e discussões dentro e fora da sala de aula.

Aos colegas de trabalho no IFPB pela compreensão e ajuda prestada permitindo que este trabalho se tornasse viável.

E a todos que, direta ou indiretamente, colaboraram para que o presente trabalho se tornasse realidade.

Resumo

A pesquisa científica possui importante papel na evolução do conhecimento humano e por isso tem sido cada vez mais incentivada como objeto de estudo. Na literatura inúmeros são os focos de análise que exploram e apontam as correlações e motivações desta evolução, mas se sabe que a pesquisa científica tem sido demonstrada como um elo bastante forte para tal desenvolvimento, pois os pesquisadores quando usados de forma estratégica podem transferir principalmente seu conhecimento para fundamentar a resolução de problemas. Além disso, trabalhos recentes vêm mostrando que a mobilidade de um pesquisador pode garantir aspectos positivos na produtividade, colaboração e internacionalização da pesquisa científica. Portanto, neste estudo resolveu-se analisar os padrões de mobilidade entre os pesquisadores atuantes ou capacitados no Brasil em instituições de ensino e pesquisa através da extração e tratamento de métricas dependentes da localidade presente em seus registros curriculares. E devido a abrangência no cenário nacional a plataforma Lattes foi adotada como principal fonte de dados sobre o histórico de deslocamento dos pesquisadores. Contudo, a análise da mobilidade só foi possível graças a metodologia de extração e tratamento dos dados deste trabalho que criou o fluxo de deslocamento de todos os doutores do Lattes usando seus registros de nascimento, formação e atuação para assim criar as métricas e visualizações utilizadas nos resultados obtidos. Com essa metodologia foi possível identificar os centros de formação de recursos humanos mais influentes para a comunidade científica nacional e a rede formada por instituições de formação e de atuação dos pesquisadores cadastrados na plataforma Lattes. Inclusive descrevendo como os deslocamentos variam ao longo do tempo. Além de identificar alguns padrões de mobilidade interessantes, como a tendência de deslocamento mínimo realizado pelos pesquisadores nos mais diversos contextos geográficos de visualização.

Palavras-chave: Mobilidade de pesquisadores, Análise de currículo, Mineração de Dados, Visualização de dados.

Abstract

Scientific research has an important role in the evolution of human knowledge and it has been increasingly encouraged as an object of study. In literature many works are focusing on analysis that exploring and point correlations and motivations of this evolution, but it is known that scientific research has been demonstrated as a very strong link to this development because researchers when used may transfer especially his knowledge to support the problems resolution. In addition, recent work has shown that the mobility of a researcher can ensure positive aspects in productivity, collaboration and internationalization of scientific research. Therefore, in this study it was decided to analyze the patterns of mobility between active or trained researchers in Brazil in teaching and research institutions through the extraction and treatment of locality dependent metrics present in their curriculum vitae records. Because of the scope on the national scene the Lattes Platform was adopted as the primary data source on the researchers mobility history. However mobility analysis was only possible because to extraction methodology and data processing from this work that created the mobility flow of all Lattes doctors using their birth, training and work records in this order to create the metrics and visualizations used in results obtained. With this methodology we could identify the training centers of the most influential human resources for the national scientific community and the network of training institutions and researchers registered activities in the Lattes platform. Even describing how the mobility vary over time. In addition to identifying some interesting patterns of mobility, such as the minimum displacement trend conducted by researchers in various geographical contexts view.

Keywords: Mobility of researchers, Curriculum analysis, Data mining, Data visualization

Lista de Figuras

Lista de Símbolos

ARS : Análise de Rede Social

API : *Application Program Interface*

BDR : Banco de Dados Relacional

C&T : Ciência & Tecnologia

CT&I : Ciência, Tecnologia & Inovação

CAPTCHA : *Completely Automated Public Turing test to tell Computers and Humans Apart*

CG : Centralidade de Grau

DD : Distância do Deslocamento

DTD : *Document Type Definition*

FNF : Fluxo de Nascimento para a primeira Formação

FFF : Fluxo de Formação para outra Formação

FFT : Fluxo de Formação para o Trabalho

GM : Grafo de Mobilidade

HTML : *Hypertext Markup Language*

HTTP : *Hypertext Transfer Protocol*

IDH : Índice de Desenvolvimento Humano

IGC : Índice Geral de Cursos

JAXB : *Java Architecture for XML Binding*

KDD : *Knowledge Discovery in Databases*

MFC : Mapa de Fluxo Circular

MMD : Modelo de Mineração de Dados

MMP : Multigrafo de Mobilidade dos Pesquisadores

MC : Mapa de Calor

ND : Número de Deslocamentos

NL : Número de Localidades

NI : Número de Instâncias

ORM : *Object-Relational Mapping*

PNPG : Programa Nacional de Pós-Graduação

PNE : Plano Nacional de Educação

PIB : Produto Interno Bruto

RN : Registro de Nascimento

RF : Registro de Formação

RT : Registro de Trabalho

RUF : *Ranking* Universitário Folha

SQL : *Structured Query Language*

SIG : Sistema de Informação Geográfica

URL : *Uniform Resource Identifier*

XSD : *XML Schema Definition*

XML : *Extensible Markup Language*

Conteúdo

Capítulo 1

Introdução

Este capítulo introduz a pesquisa deste trabalho começando com a Seção 1.1 destinada a abordar os principais aspectos que influenciaram a sua criação. Na Seção 1.2 é apresentado o seu problema de pesquisa. Já na Seção 1.3 são delineados os objetivos do trabalho em questão. Finalizando com a Seção 1.4 que descreve a organização dos demais capítulos desta dissertação.

1.1 Motivação

A pesquisa científica possui importante papel na evolução do conhecimento humano e por isso tem sido cada vez mais incentivada como objeto de estudo, não apenas nos setores acadêmicos, mas também nos diversos setores da economia como forma de se obter inovação e melhorias na logística produtiva, para assim conquistar melhor competitividade no atual cenário globalizado (SEIDL, 2011). Neste contexto, inúmeros são os focos de análise que exploram e apontam as correlações e motivações desta evolução, mas se sabe que a pesquisa científica é um elo bastante forte para tal desenvolvimento, pois os pesquisadores quando usados de forma estratégica na economia podem transferir principalmente seu conhecimento para fundamentar as estratégias de resolução de problemas tão frequentes nesse cenário (ZELLNER, 2003).

Uma das abordagens, descrita em Filippo, Casado e Gómez (2009), sugere que existe uma forte relação entre a mobilidade e produtividade dos pesquisadores de um país, e os mesmos autores afirmam que a dinâmica de um pesquisador e suas diversas colaborações

podem garantir mais trocas de conhecimento enriquecendo seu domínio interdisciplinar para se adquirir novos horizontes de pesquisa.

Além disso, a maioria do fomento dos pesquisadores e suas instituições são fortemente baseadas em questões de desempenho via produtividade de publicações. O que de certa forma representa um aspecto da mobilidade do pesquisador, pois o fluxo de trabalhos apresentados, acessíveis geralmente de forma pública por meio dos registros curriculares, quando encandeado de modo cronológico pressupõe que o pesquisador se deslocou entre as cidades dos eventos. (CAÑIBANO; OTAMENDI; ANDÚJAR, 2008; FILIPPO; CASADO; GÓMEZ, 2009).

No trabalho de Lepori, Probst et al. (2009) afirmam-se que existem pesquisas que analisam a mobilidade geográfica de pesquisadores por meio de publicações, mas garante que de um modo geral os outros registros de localidade, como a de formação e atuação, podem gerar mais informações ou características substanciais sobre o pesquisador. Em sua pesquisa, por exemplo, o autor mostra que na Suíça, onde se fala mais de um idioma, o fluxo de saída do país para obtenção de formação acadêmica tem forte influência do idioma predominante na região de origem do pesquisador.

Já segundo Jonkers e Tijssen (2008) esses deslocamentos na formação de um pesquisador apontam uma correlação direta com algumas de suas características, como no seu grau de cooperação internacional ou produção científica. E o mesmo trabalho assegura que se um país recebe pesquisadores de fora para sua qualificação isso não necessariamente implica em grandes perdas na evolução científica de tal país, pois o fato de se perder mais um pesquisador nacional pode ser substituído pelo benefício de se aumentar as chances de intercâmbio em futuras pesquisas e publicações.

No contexto brasileiro existem inúmeras pesquisas sobre os dados curriculares de seus pesquisadores. Em Júnior, Carolo e Negri (2013) afirma-se que o fomento de alguns fundos setoriais podem garantir um efeito positivo na produção acadêmica. Já Digiampietri et al. (2014b) ao avaliar a produtividade dos programas de pós-graduação em ciência da computação no Brasil cria um conjunto de métricas para viabilizar um mecanismo de comparação entre si. E Vieira e Wainer (2013) tenta cruzar uma relação entre as publicações e alguns índices de qualidade de publicação.

Até o momento muito já se fez no sentido de coletar os dados, principalmente se tratando

com dados de publicação e o fluxo de seus registros, para estratificar métricas e analisar suas correlações com questões inerentes à produtividade. Mas ainda existe um horizonte de pesquisa que pode ser mais explorado para apontar as característica da pesquisa via a mobilidade de seus envolvidos, incluindo os demais registros de localidade de seus currículos, como as formações, atuação e nascimento, para assim aprofundar as análises usando mais aspectos espaciais e temporais, tal como os trabalhos realizados por Schich et al. (2014), Koblin (2009), Abel e Sander (2014), Ratti et al. (2010), Rae (2009). Então, tal análise poderia identificar os padrões de mobilidade observados entre os pesquisadores atuantes ou formados no Brasil de modo a apontar os principais centros de formação e atuação, junto com a influência destes centros. Inclusive descrevendo como os deslocamentos variam ao longo do tempo.

1.2 Problema de Pesquisa

Para possibilitar a análise de mobilidade de pesquisadores inicialmente se fez necessário identificar os locais que constituíram sua jornada científica através de informações que estão disponíveis em seu currículo vitae. Contudo, a tarefa de obter tais informações dos pesquisadores nem sempre é uma questão trivial devido ao padrão e consistência dos dados declarados como registro de localidade que podem não estar normalizado, isso quando tais dados são informados. Além disso, os dados de localidade disponíveis nos currículos apresentam apenas o nome da localidade e não sua efetiva localização, sendo então necessário um passo adicional para traduzir este nome num par de coordenadas. Entretanto, o êxito de tal informação ajuda a compreender a dinâmica da pesquisa nacional sendo possível identificar a rede de formação e atuação dos pesquisadores no Brasil e no Mundo.

No contexto brasileiro uma importante base de registro de atividades de pesquisa está acessível na plataforma Lattes do CNPq (MENA-CHALCO et al., 2014). Em tal base é possível encontrar todos os pesquisadores ativos em projetos científicos fomentados pelo poder público, todos os pesquisadores atuantes em programas de pós-graduação, além de todos os alunos de iniciação científica, mestrado e doutorado no país. Os currículos nesta plataforma apresentam muitas informações associadas ao percurso, como a participação em projetos de pesquisa, formação acadêmica, atuação profissional, orientação de trabalhos, e

publicações.

Devido a sua abrangência no cenário nacional, a plataforma Lattes foi adotada como principal fonte de dados sobre o histórico de deslocamento dos pesquisadores formados ou atuantes no Brasil. Mas, mesmo que os registros de mobilidade sejam acessíveis, não é claro extrair informações sobre questões relevantes de mobilidade devido ao formato dos dados, à distribuição da informação em inúmeros arquivos e ao grande volume de currículos da plataforma, o que acarreta num alto custo de processamento e de tempo de resposta. Fora que tais fatos associados com o código de validação para requisitar o conteúdo dos currículos e o acesso remoto via Web acrescentam uma complexidade a mais no tempo de resposta para se gerar a mobilidade e suas interpretações.

Logo, a tentativa de se chegar o mais próximo do banco de dados dos pesquisadores cadastrados na plataforma, sem o seu acesso direto, implicou na elaboração de uma metodologia de extração dos dados abertos disponibilizados na plataforma para garantir um acesso mais eficiente e eficaz de tais dados localmente. Graças a essa metodologia é que foi possível se extrair todos os currículos e assim gerar as interpretações sobre os deslocamentos que serão apresentadas neste trabalho.

Com isso algumas questões de mobilidade puderam ser analisadas como: a relevância de algumas instituições desde o escopo mais local até ao nível global; a dependência externa para formação de pesquisadores; identificar quais são as rotas de deslocamento mais relevantes; verificar se existe associação entre o tamanho do currículo e a mobilidade; apontar quais são as instituições que mais concentram pessoas no início ou final do ciclo de formação; descrever como o deslocamento se comporta no decorrer dos anos; ou até mesmo analisar se existe alguma correlação da mobilidade com indicadores qualitativos das instituições.

1.3 Objetivos

O objetivo geral deste trabalho consiste em coletar os registros de localidade de pesquisadores para criar uma representação de sua mobilidade que possa ser analisada por meio de descrição estatística das métricas espaciais, temporais e de ligação dos fluxos de deslocamento.

Tal propósito ajudará na identificação de padrões de deslocamento entre os estados brasi-

leiros e entre diferentes países, inclusive possibilitando identificar o impacto da participação das grandes universidades e cidades brasileiras na formação de pesquisadores e profissionais ativos.

Diante disso, para se alcançar o objetivo geral foram definidos os seguintes objetivos específicos:

1. Coletar registros de localidade dos currículos de pesquisadores na Plataforma Lattes;
2. Elaborar um modelo que possa representar a mobilidade dos pesquisadores coletados numa estrutura de grafo com um conjunto de métricas;
3. Criar visualizações e interpretações que auxiliem na compreensão dos padrões e intensidades da mobilidade entre os pesquisadores;
4. Investigar alguma relação dos artefatos gerados pelo modelo com informações relacionadas a pesquisa.

1.4 Estrutura da Dissertação

Este trabalho está divido em seis capítulos, começando por este capítulo de introdução que exibiu a motivação, o problema alvo da pesquisa e o objetivo do trabalho. Enquanto que o restante deste trabalho encontra-se estruturado da seguinte forma. O Capítulo 2 descreve a fundamentação teórica apresentando os conceitos necessários para compreensão do trabalho que envolveu a ciência dos dados, o processo de mineração de dados, a estrutura e fonte dos dados da pesquisa nacional, as estratégias de armazenamento da informação, e a definição de uma representação da estrutura de mobilidade dos pesquisadores. No Capítulo 3 são apresentados alguns trabalhos relacionados ao aspecto da extração de dados curriculares e análise de mobilidade, e sobre as estratégias de modelagem e visualização da mobilidade. Já no Capítulo 4 é descrito a metodologia de extração e análise da mobilidade do presente trabalho, descrevendo o papel de cada etapa do modelo de mineração de dados proposto. Com o Capítulo 5 os principais padrões e resultados gerados na pesquisa são apresentados juntamente com algumas interpretações. Por fim, o Capítulo 6 exibe as considerações finais a respeito do trabalho desenvolvido, junto com as contribuições, limitações da pesquisa e as frentes de trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Para definir melhor o problema descrito nesta dissertação optou-se em iniciar a sua devida explicação antes de se apresentar os resultados obtidos, que contém uma série de conceitos básicos e necessário para a correta compreensão das questões de mobilidade. Dessa forma, este capítulo apresenta uma breve fundamentação dos assuntos desta pesquisa, dando ênfase aos pontos mais relevantes para a compreensão dos resultados.

Esses assuntos foram subdivididos em seções, a fim de permitir uma melhor compreensão dos mesmos. Sendo assim, a Seção 2.1 apresenta uma visão geral sobre a Ciência dos Dados, ressaltando os principais estágios envolvidos nos projetos desta área de pesquisa, dando ênfase aos processos de extração. A Seção 2.2 apresenta os conceitos básicos sobre a estrutura da pesquisa no Brasil, enfatizando os fatos e instituições que ajudaram a compor o atual cenário da pesquisa nacional, e ainda exibindo as fontes de dados, uma vez que para realizar as análises de mobilidade dos pesquisadores era necessário saber onde estava a informação de seus registros. Na Seção 2.3 são apresentados os detalhes da Plataforma Lattes visto que seus dados formam a base para a construção do grafo de mobilidade deste trabalho. Já a Seção 2.4 apresenta os principais aspectos relevantes na persistência do XML do Lattes, listando todos os procedimentos utilizados para auxiliar em sua compreensão. Por fim, na Seção 2.5 a mobilidade dos pesquisadores foi detalhada e delineada, listando todo um conjunto de métricas e visualizações para auxiliar na compreensão de padrões de deslocamento.

2.1 Ciência dos Dados

O tema principal desta pesquisa se enquadra na Ciência dos Dados, que é uma área que envolve a computação e vem apresentando ampla evolução e destaque na área da tecnologia da informação, pois atualmente, devido aos avanços tecnológicos, é comum encontrar cenários onde instituições coletam dados em grande escala gerando um aumento gradativo da complexidade de acesso e processamento dos mesmos (MCAFEE et al., 2012).

Mediante o acesso desse grande volume de dados fica o questionamento de como tratá-lo de modo eficiente, ou mesmo por onde se começa a extrair informações que possam ser relevantes. Segundo Dhar (2013) por meio da ciência dos dados seria possível encontrar informação que não são aparentes a listagem de dados convencionalmente tabular. Portanto, por meio da análise de padrões seria possível determinar tendências ou até mesmo realizar previsões (WALLER; FAWCETT, 2013), possibilitando novos desafios para os sistemas de gerenciamento de informação (AGARWAL; DHAR, 2014).

Esse horizonte de possibilidade de extração de informações pode ser útil em inúmeros cenários. Em ambientes empresariais poderia haver um maior auxílio com uso de sistemas de tomada de decisão (PROVOST; FAWCETT, 2013), na segurança ajudaria a descoberta de vulnerabilidades ocultas em imagens (SHAH, 2015), no entretenimento auxiliaria a sugerir músicas ou até encontrar a letra de alguma música que se está ouvindo (WANG, 2006). Já no cenário da saúde, dados de exames ou prontuários poderiam auxiliar no combate e alertas contra possíveis doenças ou riscos eminentes a saúde de um paciente (HERSH, 2014). Além disso, com as crescentes iniciativas de abertura de dados, principalmente governamentais, poderia-se analisar a eficiência da aplicação do dinheiro público ou indícios de fraude (UHLIR; SCHRÖDER, 2007).

Mas para que o processo de engenharia e manipulação dos dados permita a existência e extração de informação é necessário que muitas áreas da computação se integrem a outras áreas do conhecimento, tais como matemática e estatística (JONES, 2013). É por isso que muitas representações desta ciência se assemelham ao proposto pela Figura 2.1 em que Palmer (2015) ilustra a junção da computação, das ciências exatas e do conhecimento dos dados que se pretende analisar.

A Figura 2.1 também lista o aprendizado de máquina e o processamento de dados como

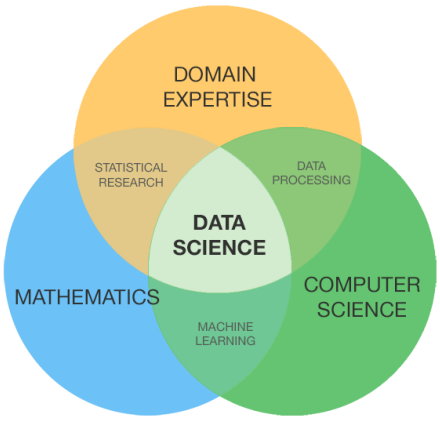


Figura 2.1: Representação da Ciência dos Dados em função das áreas de atuação. Fonte: (PALMER, 2015)

campos da computação que são relevantes para a Ciência dos Dados. Entretanto a mineração de dados, inteligência artificial e a descoberta de conhecimento em base de dados (KDD, do inglês *Knowledge Discovery In Databases*) também são utilizadas nesta ciência (JONES, 2013; DHAR, 2013; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

2.1.1 Extraindo Informação

Para realizar a interpretação dos dados, em muitos cenários, é preciso inicialmente seguir um conjunto de processos sobre os dados. Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), esta ação está alinhada com a utilização de KDD para descobrir informações úteis sobre os dados, usando algoritmos de mineração para encontrar padrões. Conforme Fayyad, Piatetsky-Shapiro e Smyth (1996) existe um processo a ser seguido a partir dos dados até se chegar à informação desejada definida por meio de algumas etapas, tal como a Figura 2.2. Então, mediante a compreensão dos dados a serem analisados e definidos os objetivos da análise, o KDD realiza estas etapas de forma sequencial fazendo:

1. Seleção: que inicia o processo de interpretação por meio da obtenção de um subconjunto de amostras e variáveis sobre os dados que contribuam para o fornecimento da informação desejada;

2. Pré-processamento: que mediante os dados selecionados realiza a sua limpeza removendo possíveis ruídos, valores redundantes ou *outliers*, além de tratar dados que são ocultos;

3. Transformação: que realiza a mudança dos dados com redução de dimensão ou aplicando a transformação de seus valores. Então é possível balancear, normalizar e combinar variáveis;

4. Mineração de dados: no qual efetua a aplicação de processamento sobre os dados transformados por meio de técnicas estatística e mineração, tais como caracterização, sumarização, classificação, regressão, descoberta de relação e clusterização, cujo o resultado seja uma potencial informação ou padrão pretendido;

5. Interpretação ou Avaliação: que é a última etapa que tem como intuito discutir os padrões e resultados gerados na etapa anterior e que geralmente é associada com visualizações e documentações.

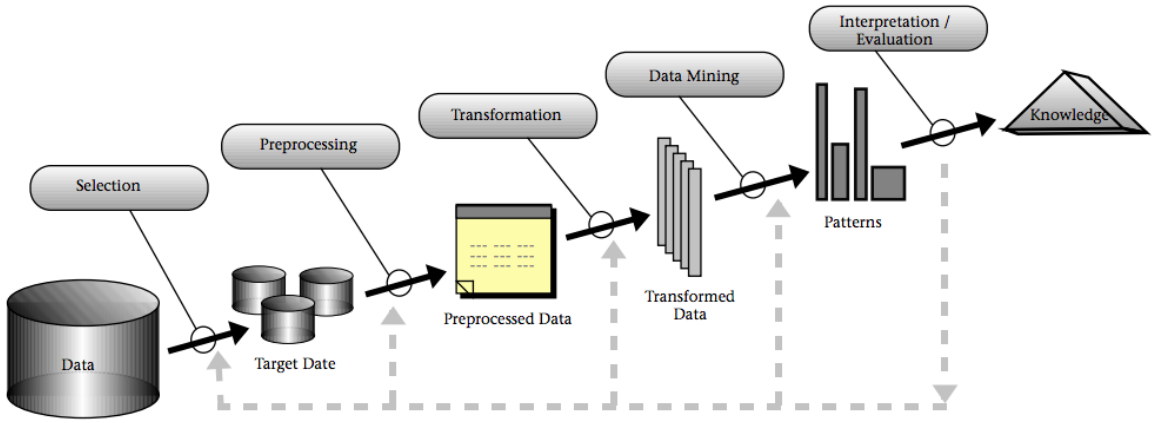


Figura 2.2: Etapas do processo de KDD. Fonte: (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)

Mas nem sempre existe acesso direto aos dados para realizar as etapas do KDD. Nestes casos são necessários utilizar algumas técnicas que ajudem na obtenção de dados, como a utilização de *Web Crawler* para coletar conteúdos remotos na *Web*.

Na realidade o *Web Crawler*, ou *spider*, é um programa que tem a capacidade de baixar páginas Web de forma automática a partir de sua URL (CHUN, 1999). Através de configurações, o *Web Crawler* pode compreender a estrutura de um portal na *Web* para obter suas principais páginas (THELWALL, 2001), evitando inclusive requisições de páginas duplicadas.

Geralmente é comum encontrar o *Web Crawler* programado para funcionar em *multi-thread* ou de modo distribuído para se obter maior desempenho no *download*. Mas, tal artifício deve ser utilizado de forma escalável para evitar degradação nas taxas de *download* (BOLDI et al., 2004; THELWALL, 2001; HEYDON; NAJORK, 1999).

Quando se armazena os conteúdos obtidos pelo *Web Crawler* acaba-se atingindo um benefício de independência da fonte original dos dados, pois o acesso direto e recorrente do mesmo conteúdo com *Web Crawler* pode enfrentar alguns problemas como: alta latência das requisições, limitações de conectividade, limitação de acesso diário das páginas e possibilidade de mudanças nas interfaces dos portais. Então, para não prejudicar a análise dos dados presente nos portais, faz-se necessário o armazenamento da página Web de forma integral ou parcial, estruturando os dados em arquivos ou em bases de dados.

Em alguns casos, o *Web Crawler* pode se beneficiar de convenções ou mesmo de vulnerabilidades de portais para obtenção de dados (HUANG et al., 2003). Por exemplo, alguns portais usam identificação numérica linear na URL para se obter as páginas, logo um laço em programação facilmente poderia ajudar neste cenário. Já em outros casos existem páginas de pesquisa que listam todas as possíveis páginas de um portal. Mesmo na situação em que as páginas de consulta não permitam a listagem integral de sua estrutura, dependendo do caso, as consultas podem ser realizadas mediante *SQL Injection*[1] para se obter as páginas na integra.

Inclusive quando se encontra barreira no uso do *Web Crawler* imposto pelo uso de CAPTCHA[2] na requisição de uma página, é possível usar técnicas de reconhecimento de padrão em imagens para burlar a segurança imposta por este artifício e assim acessar a página pretendida.

[1]Quando sistemas usam Base de Dados Relacional e não fazem o devido tratamento na cláusula WHERE, curingas ou caracteres específicos podem gerar resultados inesperados permitindo acessos ou mudanças não autorizados no banco.

[2]Este acrônimo deriva da expressão em inglês *Completely Automated Public Turing test to tell Computers and Humans Apart*, também conhecido como teste de Turing, seu criador, cujo objetivo consiste em aplicar um desafio cognitivo que seria dificilmente interpretado pelo computador, mas de fácil compreensão humana. Muito utilizado para evitar negação de serviço na Web por grandes volumes de requisições que podem ser disparadas por algum *Web Crawler* mal intencionado (HUANG et al., 2003).

2.1.2 Inferência de Localidades

A geocodificação de uma localidade permite definir com exatidão o seu posicionamento na superfície terrestre por meio de alguma coordenada, como a latitude e longitude (GOLDBERG, 2008). Os registros de geocodificação geralmente estão disponíveis em bases de Sistema de Informação Geográfica (SIG), mas é possível encontrar esses registros em inúmeras bases de dados que contêm informações geográficas (CHRISTEN, 2012).

Christen, Churches e Willmore (2004) e Zandbergen (2008) afirmam que em algumas bases existem nomes de localidades que nem sempre determinam sua efetiva localização, pois é preciso que a localidade esteja disponível em algum SIG para determinar o seu posicionamento. Já Sengar et al. (2007) abordam que a geocodificação de uma localidade a partir de um nome envolve um grande desafio por conta: das variações de grafia inerentes a um local, das inconsistências de ortografia e digitação, e variações no formato das localidades entre os países. Além disso, Christen e Belacic (2005) constatam que as ruas podem ser renomeadas, ou novos locais sempre são construídos ao longo dos anos. Portanto, o tratamento e padronização dos endereços, bem como a verificação se eles realmente existem, são, portanto, importantes passos no processo de mineração de tais dados.

Para contornar o desafio da efetiva inferência de localidade, várias propostas são sugeridas. Christen e Belacic (2005) sugerem uma metodologia baseada em fatores probabilísticos modelados por meio de cadeias de Markov. Já Fu, Christen e Boot (2011), para ligar registros históricos, utilizam um pré-processamento de limpeza das localidades para em seguida realizar a sua comparação a uma base normalizada. Enquanto que Goldberg (2011) realiza comparações baseadas num esquema de níveis de pontuação na associação. Ao passo que Goeken et al. (2011), para resolver o mesmo problema, usam máquina de vetores de suporte para classificar potenciais ligações.

Em Winkler (2006) é listado um conjunto de técnicas como a padronização de localidades, comparação de texto e métodos de associação de localidades a partir de uma base normalizada. Mas o mesmo autor relata uma importante característica neste problema definindo que nomes de alta frequência tendem a ser mais próximos da versão normalizada de uma localidade (WINKLER, 1995). Logo, a utilização dos registros de alta frequência com a associação de localidades normalizadas em SIG pode se tornar uma alternativa para ajudar na convergência da geocodificação de um conjunto de localidade declaradas em sistemas de

currículo vitae de pesquisadores.

Mas mediante a caraterística citada acima se torna importante a seleção de bases normalizadas de geolocalização que listem os nomes de localidades com suas respectivas coordenadas. Roongpiboonsopit e Karimi (2010) afirmam que esse processo pode ser feito por meio de alguma base de referência disponível em bases de SIGs, como o ArcGIS (STEINER et al., 2003). Mas além disso, no mesmo trabalho os autores destacam que inúmeros SIGs tentam incorporar as técnicas de inferência de localidades para aumentar a acurácia de suas consultas por meio de textos de localidade. Inclusive os SIGs que possuem serviços disponibilizados na Web, como a API do Google Maps[3], Yahoo[4] e Open Street Map[5], por meio de serviços na Web (ROONGPIBOONSOPIT; KARIMI, 2010).

Uma alternativa ao uso dos SIG seria utilizar as bases de geolocalização abertas, como o Geonames[6] que é uma base geográfica que acumula mais de dez milhões de registros com acesso via serviços na Web e extratos em arquivos, sendo bem mais leve em relação a outras bases semelhantes como o OpenStreetMap[7]. Trabalhos como Popescu, Grefenstette e Moëllic (2008) mostram que esses tipos de base na realidade já são resultados de mineração sobre inúmeras fontes abertas de dados geográficos.

Por fim, uma outra alternativa aplicada na inferência de localidade é o uso de bases semânticas na Web, pois essas bases usam ontologias para compor as informações geográficas (DING et al., 2009; JAIN et al., 2010). Inclusive algumas bases estão usando esta abordagem para melhor compor sua estrutura, como o caso do Geonames (HAHMANN; BURGHARDT, 2010). Além do Geonames, o Freebase[8] e DBpedia[9] também se aplicam a esta abordagem (YU et al., 2014; GOVAERTS; DUVAL, 2009; MENEZES et al., 2014). Então usando o DBpedia seria possível obter informações sobre a Universidade de São Paulo (USP) através da URL:

http://dbpedia.org/page/University_of_São_Paulo

que retorna a estrutura semântica deste conceito segundo a base. A partir deste conceito

[3]https://www.google.com.br/maps
[4]https://developer.yahoo.com/maps/
[5]http://wiki.openstreetmap.org/wiki/API
[6]http://www.geonames.org/
[7]https://www.openstreetmap.org/about
[8]https://www.freebase.com/
[9]http://wiki.dbpedia.org/

é possível esboçar parte de sua estrutura semântica na Figura 2.3 que a situa no Brasil em pleno estado de São Paulo pelas relações *dbpedia-owl:country* e *dbpedia-owl:state*.

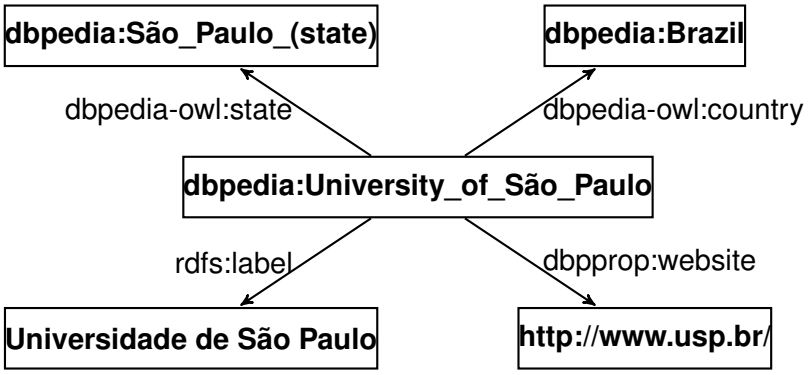


Figura 2.3: Parte da representação da USP via ontologia do DBpedia.

2.2 Estrutura de Pesquisa no Brasil

A análise da evolução e mobilidade dos pesquisadores no Brasil está muito associada à história da pesquisa nacional. Dessa forma, é importante ressaltar os fatos que nortearam a atual conjuntura da pesquisa brasileira, listando as instituições envolvidas e as fontes de dados que retratam e aferem essa conjuntura.

2.2.1 Origem e Características

Os primeiros registros sobre pesquisa no Brasil se iniciam no século XX com a atuação de alguns centros de educação superior e técnica no Brasil. Mas o foco principal destes centros não era a pesquisa e sim a profissionalização de pessoas direcionado-as para a política, administração pública e criação de trabalhadores envolvendo principalmente direito, medicina e engenharia (ROMÊO-UNU; IBMEC, 2004). Porém, a partir do momento que se inicia as primeiras frentes de crescimento da indústria nacional, juntamente com a entrada de investimentos externos e privado, se expande a necessidade de desenvolver a Ciência e Tecnologia (C&T) nacional por meio da formação da mão de obra especializada, e aumento no ensino superior e profissionalizante (SANTOS, 2003).

Com isso, sugiram as primeiras tentativas de reduzir a dependência externa de C&T que culminaram na década de 50 com fatos marcantes para pesquisa nacional. Por exemplo, a cri-

ação da rede universitária federal por meio da federalização de alguns centros universitários existentes nesse período, no intuito de padronizar e associar a pesquisa e o ensino nacional. Além da criação de órgãos como a CAPES[10] e o CNPq[11] para amparar respectivamente o ensino superior e a pesquisa no Brasil (MOROSINI; SOUZA, 2009).

Mesmo assim, percebe-se que não houve uma expressiva tentativa e incentivo do setor privado de reverter o quadro de dependência internacional por meio do investimento em C&T, pois seu custo era maior que a aquisição dos produtos manufaturados prontos advindos do exterior. Portanto, o resultado foi uma forte dependência do setor público para a implantação da pesquisa nacional, que ainda se matem nos dias atuais. O que difere de países desenvolvidos onde as iniciativas privadas começaram e mantiveram um importante espaço nas pesquisas de seus países, motivados principalmente em desenvolver seus interesses em solucionar problemas práticos de uma sociedade que vivia o desenvolvimento industrial (VERHINE, 2008).

Na década de 60, o Brasil contava com 38 cursos de pós-graduação, sendo 11 de doutorado e 27 de mestrado, porém ainda não havia uma plena padronização e integração entre as instituições. Em alguns casos o professor catedrático era o dirigente máximo, enquanto o assistente e o associado trabalhavam sob sua direção. Mas junto com o regime militar veio o intuito de delinear a pesquisa nacional para garantir avanços econômicos por meio da evolução da C&T, surgindo assim importantes marcos para pesquisa nacional, que foram o Parecer Sucupira e a Lei de Reforma Universitária advindos respectivamente pelo parecer C.E.Su. n° 977 de 1965 e pela Lei n° 5.540 de 1968 (VELLOSO et al., 2003).

No parecer Sucupira decretou-se a padronização das instituições de ensino superior num formato de estrutura departamental. Para assegurar a criação de programas de pós-graduação foi determinado que os professores assistentes deveriam ter o grau de mestre e os adjuntos o de doutor. Tais programas de pós-graduação também deveriam seguir o modelo composto de uma combinação de créditos, exames e uma dissertação supervisionada. Este fato acarretou na necessidade da formação de professores e pesquisadores brasileiros no exterior, assim como os acordos de intercâmbio cultural e científico que traziam pesquisadores de vários países ao Brasil (VELLOSO et al., 2003).

[10] http://www.capes.gov.br/
[11] http://www.cnpq.br/

Já na Lei da Reforma Universitária a pós-graduação ganha destaque na estrutura organizacional da universidade e recebia o objetivo de qualificar professores para o ensino superior, capacitar pessoal para atuar nos setores público e privado, e estimular a produção de conhecimento científico vinculada ao desenvolvimento do país (MOROSINI; SOUZA, 2009).

Em 1975, o país ainda possuía 50 instituições de ensino superior nas quais funcionavam 195 cursos de mestrado e 68 de doutorado (ROMÊO-UNU; IBMEC, 2004). Contudo, para progredir ainda mais os rumos da pesquisa no Brasil, surgiu nesse ano o Plano Nacional de Pós-Graduação (PNPG) no intuito evitar o processo de expansão desordenado da pós-graduação (SANTOS; AZEVEDO, 2009).

Até o momento o PNPG já se encontra na sexta edição sendo gerido pela CAPES por intermédio do Ministério da Educação, e em cada documento é possível extrair inúmeras informações importantes da pós-graduação do Brasil (CAPES, 2010). Por exemplo, em toda a sua história o PNPG vem gerando importantes pontos de inflexão na pesquisa e pós-graduação nacional. Dentre as inúmeras metas e diretrizes criadas pode-se listar: a expansão e fortalecimento dos programas de pós-graduação com integração do ensino à pesquisa; o estímulo a criação de centros de pesquisa e a capacitação de pesquisadores e professores; e a criação de um sistema nacional de avaliação dos programas para qualificar as pesquisas realizadas. Mas inúmeros desafios marcam o atual PNPG, dentre eles pode-se listar: a redução da assimetria regional; o estímulo a internacionalização das pesquisas e o surgimento de novas áreas de conhecimento; a elevação dos padrões de pesquisa nacional comparando-se aos países desenvolvidos; e a criação de políticas de valorização na formação do magistério na educação básica.

Também é documentado nos documentos do PNPG que várias outras ações importantes foram implantadas para atender as necessidades de cada momento. Por exemplo, na década de 90 estudos mostraram a necessidade da sociedade pelo mestrado que envolvesse mais as pesquisas aplicadas nas cadeias produtivas do mercado, o que culminou com a criação do mestrado profissional. Mas além da CAPES, outros importantes fundos advêm do Ministério de Ciência e Tecnologia do governo federal, por meio das agências CNPq, FAPs e FINEP. Essas instituições fornecem um importante amparo às instituições de pesquisa em Ciência, Tecnologia e Inovação (CT&I), através do financiamento de pesquisadores com a CNPq e FAPs, e os projetos de inovação em pesquisa de empresas com a FINEP (CAPES, 2010).

Até o Plano Nacional de Educação (PNE)[12] 2011/2020 define várias estratégias para evoluir a pesquisa nacional. Entre as várias proposições de metas do PNE, a 14ª declara que se deve elevar gradualmente o número de matrículas na pós-graduação stricto sensu, de modo a atingir a titulação anual de 60.000 mestres e 25.000 doutores até o final de 2020. O grande desafio para essa meta é a manutenção do crescimento do número de bolsas de estudo e a indução de novos cursos de doutorado em áreas estratégicas para o desenvolvimento do país.

Concluindo, o resultado de todas essas ações resultou na marca de 4.101 cursos de pós-graduação em 2009, sendo 2.436 para mestrados acadêmicos, 243 para mestrados profissionais e 1.422 para doutorados. Também houve uma expressiva evolução do número de discentes, matriculados e formados; docentes atuantes; e concessão de bolsas, no Brasil e exterior. Em termos quantitativos percebe-se que desde a década de 90 as instituições privadas vêm ocupando um espaço cada vez maior, além de uma crescente participação das empresas na produção de CT&I (CAPES, 2015).

2.2.2 Fonte de Dados

Como já foi dito, uma importante fonte de dados no Brasil sobre pesquisa está acessível na plataforma Lattes do CNPq, onde se encontra boa parte dos pesquisadores ativos na pesquisa nacional com inúmeras informações sobre participação em projetos de pesquisa, formação acadêmica, atuação profissional, orientação de trabalhos e publicações. Mas no CNPq também é possível encontrar resultados de editais de bolsas, auxílios, programas e prêmios de amparo à pesquisa, como por exemplo as bolsas de produtividade, bolsas de pós-graduação, auxílio de eventos e de financiamento de projeto de pesquisa.

Outra importante fonte de dados disponível no Brasil está acessível na CAPES. O próprio PNPG fornece estatísticas que estratificam e projetam indicadores da pesquisa nacional, muitos deles acessíveis pelo sistema Geocapes[13]. Outro sistema desse órgão é o Sucupira[14], no qual os programas de pós-graduação no Brasil informam e publicam dados cadastrais, linhas de pesquisa, discentes, docentes, projetos de pesquisa, trabalhos concluídos e participantes externos. Também se encontram na CAPES editais de bolsa, programas e prêmios.

[12] http://pne.mec.gov.br/
[13] http://geocapes.capes.gov.br/geocapesds/
[14] https://sucupira.capes.gov.br/sucupira/

Outros dados importantes publicados pela CAPES são os conceitos de programas de pós-graduação do Brasil, que é considerado um instrumento de classificação das instituições de ensino e pesquisa nesta modalidade. Mas além deste mecanismo de avaliação existem outros processos que inclusive são independentes do poder público, como o RUF[15] que é uma avaliação anual do ensino superior do Brasil feita pelo jornal Folha de São Paulo desde 2012, ou o Webometrics[16] que é uma avaliação semestral de instituições de pesquisa e ensino em escala mundial que existe desde 2004.

Para este estudo a CAPES e CNPq servirão de importante valia na obtenção de dados, mas outras instituições também podem ser consideradas como fontes de dados. Um caso seria o INEP[17], que em seu portal oferece o censo da educação superior no Brasil[18] com importantes informações sobre as instituições de ensino superior e seus cursos, e o censo escolar no DataEscolaBrasil [19]. Também é possível encontrar resultados de avaliações educacionais que contêm listas de instituições e cursos como o ENADE[20], ENEM[21] e Prova Brasil[22]. Inclusive outra possibilidade de comparação de instituições de ensino superior por meio do conceito IGC[23].

Além disso, o Ministério da Educação oferece outros sistemas e dados que também podem oferecer locais de formação e atuação dos pesquisadores. O e-Mec[24] e Sistec[25] enriquecem a lista de cursos e instituições no Brasil. Programas de financiamento também podem oferecer dados como o programa Ciência sem fronteiras[26], que oferece dados sobre seus investimentos, inclusive pessoas e instituições envolvidas, ou no programa Mais Educação[27], que é um fundo focado para instituições públicas do ensino fundamental. Inclusive até resultado de indicadores educacionais, como o IDEB[28], podem ser utilizados como fonte de dados.

[15] http://ruf.folha.uol.com.br/2014/
[16] http://www.webometrics.info/
[17] http://www.inep.gov.br/
[18] http://portal.inep.gov.br/basica-levantamentos-acessar
[19] http://www.dataescolabrasil.inep.gov.br/dataEscolaBrasil/
[20] http://portal.inep.gov.br/enade
[21] http://enem.inep.gov.br/
[22] http://portal.inep.gov.br/web/saeb/aneb-e-anresc
[23] http://portal.inep.gov.br/educacao-superior/indicadores/indice-geral-de-cursos-igc
[24] http://emec.mec.gov.br/
[25] http://sistec.mec.gov.br/login/login
[26] http://www.cienciasemfronteiras.gov.br/
[27] http://portal.mec.gov.br/index.php?option=com_content&id=16690&Itemid=1115
[28] http://ideb.inep.gov.br/

Em todas essas possibilidades de fonte de dados citadas nesta seção, poucas são as fontes que oferecem seus dados em formato aberto e que seja de fácil o seu uso e processamento. Boa parte pode ser usada como fonte de dados de modo indireto por meio de extração de localidades em páginas HTML usando *Web Crawler*, ou PDF usando analisadores deste tipo de arquivo.

2.3 Plataforma Lattes

A Plataforma Lattes[29] é um sistema de informação do CNPq para gerenciar dados relacionados a pesquisadores e instituições do país criado em 1999 a partir de um projeto desenvolvido pelas Universidades Federal de Santa Catarina (UFSC) e Federal de Pernambuco (UFPE) em parceria com empresas privadas. Atualmente é dividido em quatro módulos: o Currículos Lattes, o Diretório de Instituições, o Diretório dos Grupos de Pesquisa e o Painel do Lattes. Muitas instituições de fomento, universidades e institutos de pesquisa do país usam a plataforma para registrar dados sobre a pesquisa nacional (CNPQ, 2015).

Seu maior destaque se dá ao cadastro de currículo vitae por meio do Currículo Lattes que como já foi citado possui um conjunto de registros da vida pregressa e atual da maioria dos pesquisadores atuantes no cenário nacional. Devido a grande quantidade de informações existente no Currículo Lattes, no qual já ultrapassa quatro milhões de registos, muitos programas de financiamento na área de ciência e tecnologia acabam utilizando-o para análise de mérito e competência do pesquisador (CNPQ, 2015).

Já o Diretório de Instituições cadastra no CNPq informações e hierarquia organizacional de instituições para auxiliar na obtenção de dados em programas promovidos pela agência. O Diretório dos Grupos de Pesquisa constitui-se no inventário dos grupos de pesquisa científica e tecnológica em atividade no país. Por fim, o Painel do Lattes oferece um conjunto de análises descritivas dos dados da plataforma dispostos graficamente em séries históricas. Entretanto, existe um conjunto limitado de análises e restrição direta aos dados da análise. Portanto, desses módulos o mais importante para a corrente análise de mobilidade será o Currículo Lattes que será descrito a seguir (CNPQ, 2015).

[29] http://lattes.cnpq.br/

2.3.1 Currículo Lattes

Dado a finalidade deste sistema, Lane (2010) frisa que a utilização do Currículo Lattes como base de dados dos pesquisadores é hoje uma das melhores maneiras de analisar o perfil acadêmico dos pesquisadores brasileiros, já que a criação do currículo se torna obrigatória em várias atividades que envolvem pesquisas. Além de ser um importante instrumento para gerar estatísticas que orientam as políticas públicas de educação (GUERRA, 2012).

Qualquer pessoa pode se cadastrar na plataforma para alimentar seus dados curriculares, e mediante o cadastro todos os currículos passam a possuir duas identificações, uma no formato numérico de 16 caracteres e a outra no formato alfanumérico de 10 caracteres, que serão chamadas respectivamente de ID16 e ID10.

Essas identificações serão úteis para acessar os dados curriculares da plataforma através dos dois formatos públicos existente na Web, que é o HTML e o XML. Então, para acessar um dos formatos basta usar os identificadores combinados a alguns padrões de URL. No caso do HTML, seu acesso pode ser feito mediante as URLs:

http://buscatextual.cnpq.br/buscatextual/visualizacv.do?id=<ID10>

http://lattes.cnpq.br/<ID16>

e para acessar o XML o padrão da URL seria:

http://buscatextual.cnpq.br/buscatextual/download.do?&idcnpq=<ID16>

sendo também necessário em ambos acessos, atualmente, um preenchimento de CAPTCHA para se acessar a visualização pretendida.

No caso do autor desta dissertação o <ID16> é 7165875430419020 e o <ID10> é K4241290Z6, que podem ser obtidos na própria plataforma através da página de consulta de currículo. O <ID10> para acesso da versão HTML é identificado na página de consulta da plataforma ao listar os *links* da consulta sobre o nome do autor. Mediante este <ID10> se gera a URL de acesso ao currículo na versão HTML, presente na Figura 2.4, e que por consequência viabiliza a extração do <ID16> além da URL de acesso do XML.

Toda a estrutura que norteia o padrão do XML do Lattes foi definida pela comunidade LMPL[30], cujo objetivo principal foi a criação de ontologias que auxiliam no intercâmbio de informações entre agências de fomento à pesquisa e instituições ligadas a CT&I (CNPQ,

[30]http://lmpl.cnpq.br/lmpl/

Figura 2.4: Acesso de identificadores de curriculum na Plataforma Lattes.

2015). Tal ação permitiu a criação da ontologia do Currículo Lattes por meio de um DTD[31].

2.3.2 Acesso aos Dados

Atualmente o acesso direto aos dados do Lattes só é possível as instituições que já possuem protocolo de cooperação técnica firmado com o CNPq mediante um ofício a sua presidência, devidamente assinado por algum dirigente máximo de instituição de pesquisa, contendo a exposição de motivos e destinação a ser dada aos dados a serem extraídos. Este acesso pode ser feito por meio do espelhamento ou extração de curriculum no formato XML (CNPQ, 2015).

A existência de tais dados é cobiçada por muitas intuições para aferir e controlar dados dos seus envolvidos. Entretanto, por ser este um processo burocrático e potencialmente demorado, muitos pesquisadores têm buscado alternativas para extrair dados da plataforma Lattes, copiando os dados a partir dos formatos públicos existentes (MENA-CHALCO et al., 2014; DIGIAMPIETRI et al., 2014a; DIGIAMPIETRI et al., 2014b).

Dos formatos disponíveis citados no Lattes a melhor escolha seria o XML devido a associação semântica dos dados e porque algumas informações são disponibilizadas exclusivamente neste formato (DIGIAMPIETRI et al., 2012a; DIGIAMPIETRI et al., 2012b).

Contudo, esta estratégia exige o uso de um *Web Crawler* para copiar o XML associado a um analisador de CAPTCHA, para informar os caracteres do código de segurança, no sentido

[31]Este arquivo é usado para definir a estrutura de um XML, e a URL do DTD do Lattes é http://www.cnpq.br/documents/313759/b6f13489-2166-4cb4-8be5-8ab3fb5ab106

de acessar as informações contidas no currículo.

2.4 Armazenamento de XML

A partir do momento que se obtêm um conjunto de arquivos XML, muitas estratégias podem ser utilizadas para seu armazenamento e acesso, principalmente com o auxílio de algum banco de dados. Neste cenário percebe-se que o arquivo pode ser tratado de forma integral ou segmentada. Já em relação ao tipo de banco de dados, existe a possibilidade de se usar o XML, Relacional ou NoSQL (JAGADISH et al., 2002; HAN et al., 2011). Neste trabalho foi utilizado o armazenamento do XML integral e segmentado em Banco de Dados Relacional (BDR).

2.4.1 XML Integral em BDR

Alguns BDRs possuem suporte à manipulação de arquivo XML devido ao fato dos registros de tabela serem exportados neste formato, e no geral também se pode fazer o inverso, ou seja, inserir registros contidos em XML nas tabelas. Já em outros BDR o XML é tratado como um tipo de dado, além de permitir o seu acesso granular por meio de consultas SQL usando Xpath[32] (TATARINOV et al., 2002).

O benefício deste armazenamento se dá por meio das otimizações que o BDR aplica na persistência de seus registros, como compressão no armazenamento e agilidade na consulta. Inclusive, essa solução garante que grandes volumes de registros XML sejam persistidos, diferente de um sistema de arquivo tradicional que não foi idealizado para receber milhares de registros de arquivo em um único diretório. Contudo, enquanto se possui uma pequena quantidade de XML, as consultas granulares geram tempos aceitáveis. Mas a partir do momento que se aumenta o número de XML percebe-se que as consultas granulares acabam se tornando lentas, durando até dias dependendo do caso, principalmente quando se trabalha com recursos escassos de *hardware*.

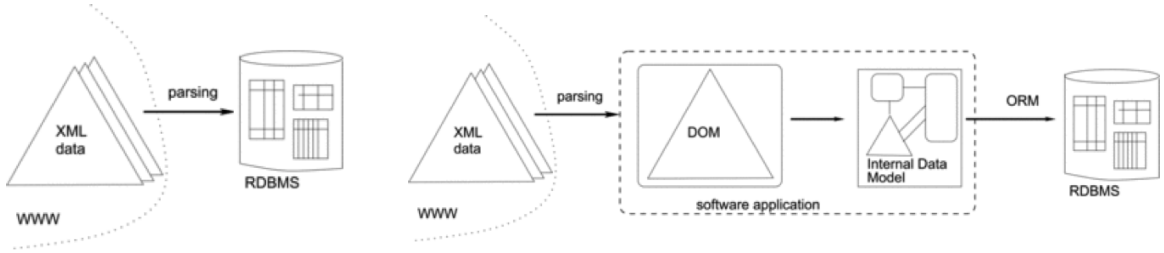
Em parte, isso se justifica pelo fato de que o processamento granular do XML eventualmente envolve complexos processamento de texto ou até mesmo sua manipulação em

[32]Este termo é um acrônimo de *XML Path Language* e garante ao XML um mecanismo de seleção de seus elementos e atributos.

DOM[33], sendo um processo ineficiente quando se repete a cada consulta (JAGADISH et al., 2002).

2.4.2 XML Segmentado em BDR

Para contornar o citado problema de complexidade na extração de informação granular do XML no BDR, Dukovich et al. (2008) e Klettke e Meyer (2001) sugerem usar o acesso granular mediante o uso de tabelas no BDR que possuam os dados e estrutura do XML, ou seja, diante do mapeamento dos elementos e atributos desejados do XML em tabelas e colunas do BDR, ilustrado no esquema da Figura 2.5. De acordo com os autores, este mapeamento pode ser feito de duas maneiras: manualmente ou automaticamente.



(a) manual

(b) automático

Figura 2.5: Esquema de Mapeamento dos XML em BDR. Fonte: (DUKOVICH et al., 2008)

O processo manual, representado pela Figura 2.5a, seria o mais complexo pelo fato de que a criação de tabelas e a realização das consultas do XML para alimentar as tabelas seriam manuais. No contexto dos arquivos XML do Lattes esse processo se agrava mais ainda devido a sua estrutura que possui uma grande dimensão, incluindo mais de 300 elementos e quase 2.000 atributos. Essa situação de arquivos XML com grande extensão não é exclusividade do Lattes, outros projetos acabam forçando a necessidade de mapeamento automático do XML devido as proporções de sua estrutura, como no caso dos dados de bioinformática citados por Dionisio e Dahlquist (2008).

Logo o processo automático se torna mais otimizado porque o XML é mapeado em objetos que por sua vez refletem tabelas de BDR, conforme a Figura 2.5b, o que evita qualquer processo manual de criação e alimentação das tabelas. Os próprios autores Dukovich et

[33]Este termo é um acrônimo de *Document Object Model* e padroniza uma representação de documentos de marcação, como XML e HTML, em objetos.

al. (2008) sugerem a utilização do *framework* HyperJAXB[34] que funciona basicamente integrando a biblioteca JAXB[35], para carregar o XML em memória através de objetos, junto com o Hibernate[36], para persistir os objetos do JAXB em tabelas de BDR, por meio de um mapeamento automático de ambos, ou seja, utilizando o mapeamento objeto relacional (ORM, do inglês *Object-relational mapping*). Além disso, por ser um *framework*, pode ser facilmente aglutinado como um componente do projeto da mesma forma que o XMLPipeDB[37] fez para tratar os dados de bioinformática em Dionisio e Dahlquist (2008).

A princípio, esta transformação foi possível graças a disponibilização do DTD que define o XML do Lattes na própria plataforma, porque o HyperJAXB3 utiliza como entrada o arquivo de definição do XML a ser persistido e as várias instâncias desta definição em arquivo XML. Pois caso não houvesse este DTD, seria necessário percorrer todos os XMLs com o intuito de encontrar todas as possíveis entidades, atributos e seus relacionamentos para inferir sua definição, o que demandaria mais tempo de conversão. Mesmo assim, essa definição inferida seria válida apenas para o conjunto que o gerou, podendo não ser válida para novos arquivos XML. Outra questão é que o arquivo de definição de XML que o HyperJAXB trata é o XSD[38], o que exigiu a conversão do DTD para XSD com as devidas adaptações de tipo, para que fosse possível armazenar todos os dados existentes do Lattes.

2.5 Mobilidade dos Pesquisadores

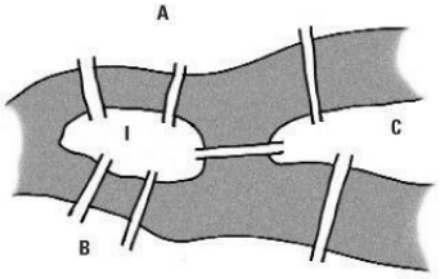
A ideia de mobilidade dos pesquisadores deste trabalho surgiu historicamente do estudo de Leonhard Euler sobre deslocamento entre um conjunto de pontes que haviam na cidade de Königsberg da extinta Prússica em pleno século XVIII (CRILLY, 2007). Na época, essas pontes cortavam o Rio Pregel para conectar duas grandes ilhas, que juntas formam um complexo com sete pontes conforme mostra a Figura 2.6a.

Por muito tempo os habitantes daquela cidade perguntavam-se se era possível cruzar as sete pontes numa caminhada contínua sem passar duas vezes por qualquer uma delas. Para

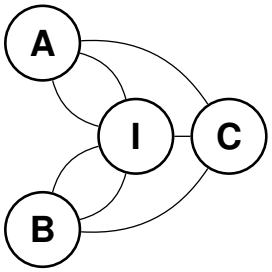
[34] https://github.com/highsource/hyperjaxb3
[35] https://jaxb.java.net/
[36] http://hibernate.org/
[37] http://xmlpipedb.cs.lmu.edu/
[38] Este termo é um acrônimo de *XML Schema Definition* e propõe um mecanismo de definição de estrutura do XML mais avançado que o DTD.

resolver esse problema Euler idealizou a estrutura de um multigrafo não orientado conexo, ilustrado na Figura 2.6b, em que as arestas e nós abstraiam respectivamente as pontes e os locais que elas ligavam. Esta representação difere de um grafo simples porque existem arestas múltiplas, e é considerado conexo porque é possível estabelecer um caminho de qualquer vértice para qualquer outro vértice do multigrafo (CRILLY, 2007).

A análise em sua abstração permitiu que ele afirmasse que não havia possibilidade de encontrar um caminho sem repetições nas pontes, provando que se o multigrafo não direcionado apresentasse apenas um vértice com número de arestas ímpar seria impossível encontrar um percurso sem repetições (CRILLY, 2007).



(a) Ilustração das pontes. Fonte: (CRILLY, 2007)



(b) Multigrafo

	A	I	B	C
A	0	2	0	1
I	2	0	2	1
B	0	2	0	1
C	1	1	1	0

(c) Matriz de adjacência

Figura 2.6: Problema das sete pontes de Königsberg

Essa abstração é considerada um dos primeiros estudos sobre a teria dos grafos (CRILLY, 2007). Mas além da representação visual do grafo por meio de nós e arestas existem outras organizações que refletem esta estrutura. Entre elas existem a matriz de adjacência, a matriz de incidência e a lista de adjacência, no qual cada uma possui suas peculiaridades e benefícios (CORMEN, 2009).

No caso do problema das sete pontes de Königsberg o multigrafo pode ser adaptado para a matriz de adjacência da Figura 2.6c, onde as colunas e linhas estão representando os quatro pontos conectados pelas pontes. Essa matriz é quadrada e possui a coluna principal nula porque não existe laços no multigrafo, já os pontos que são maiores que um indicam a existência de arestas múltiplas entre os nós, enquanto que os demais valores nulos indicam ausência de ligação. Nessa matriz também é possível identificar um espelhamento de valores na coluna principal, devido ao fato das arestas não possuírem direcionamento.

Diante deste potencial de abstração dos nós e arestas para representar cenários, percebe-

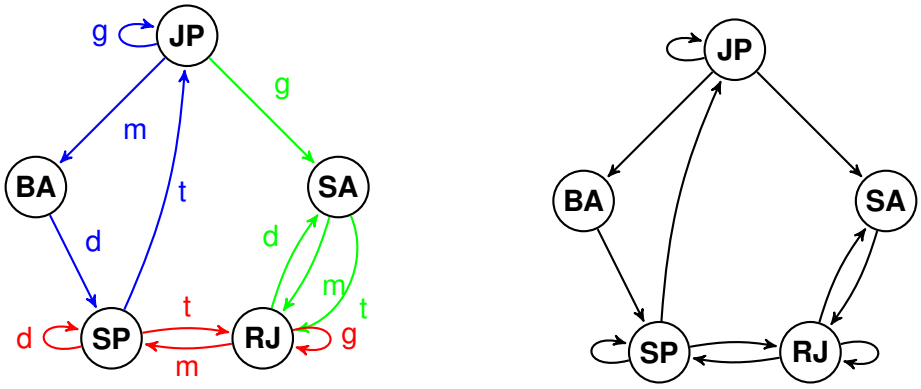
se que desde então os grafos são utilizados para os mais diversos fins, como na descrição de conexões de redes sociais (DIGIAMPIETRI et al., 2014a), ou até na representação do fluxo do comércio global (KLETTKE; MEYER, 2001). Outra aplicação recorrente dos grafos é a representação de deslocamentos de indivíduos, pois no geral, os nós podem indicar os pontos onde as pessoas passaram e as arestas esboçam sua trajetória de deslocamento entre os pontos (RAE, 2009; RATTI et al., 2010; SCHICH et al., 2014). Ou seja, uma outra proposição para representar o problema de mobilidade abordado por Euler, focando nas pessoas e não no percurso em si.

Portanto, nesta pesquisa, para estruturar a mobilidade dos pesquisadores do Lattes basta adaptar a estrutura de Euler para permitir orientação no fluxo e existência de laços. A primeira adaptação é necessária porque os pesquisadores se deslocam entre as cidades com Registro de Nascimento (RN), Registro de Formação (RF) e Registro de Trabalho (RT) conforme a ordem cronológica de cada acontecimento de sua formação e atuação. Então o pesquisador deve iniciar seu deslocamento orientado pela cidade com RN, passando pelas cidades com RF e terminando com a cidade de RT. Já a existência do laço seria para representar a mudança de formação ou trabalho sem deslocamento de lugar. Em outras palavras, esta estrutura é a representação de um multigrafo orientado ou dígrafo.

Com essa adaptação seria possível organizar o Multigrafo de Mobilidade dos Pesquisadores (MMP) considerando que os nós, representando os locais, podem ser visualizados através do contexto de sua instituição, cidade, estado, região, país ou continente; enquanto que as arestas, representando os deslocamentos, apontam os fluxos de cada pesquisador existente no percurso entre nós de um contexto específico. Esta abordagem é similar à forma que Schich et al. (2014) e Ratti et al. (2010) utilizou na abstração de mobilidade, mas difere pelo fato de que, a priore, considera-se apenas como nós, ou pontos de mobilidade, os RT, RF e RT declarados pelos pesquisadores, pois assim viabiliza a simplificação da composição e análise do MMP.

Por exemplo, na Figura 2.7a quando se monta a mobilidade no contexto das cidades, de três pesquisadores, indicada pelas três cores, através da formação nos níveis de graduação, mestrado, doutorado, e a atuação profissional designadas respectivamente pelas letras g, m, d e t na Figura 2.7a se compõe um multigrafo orientado. No qual os nós são compostos pelas cidades que possuem RN, RF e RT. Já as arestas são compostas pelos Fluxos de Nascimento

para a primeira Formação (FNF), de Formação para outra Formação (FFF) e da última Formação para o local de Trabalho (FFT) de cada pesquisador entre duas cidades. E a ordem de cada fluxo depende da posição cronológica de tais eventos para cada pesquisador.



(a) sem agrupamento (b) com agrupamento

Figura 2.7: Ilustração do Grafo de Mobilidade de três pesquisadores entre as cidades de João Pessoa (JP), Salvador (SA), Brasília (BA), Rio de Janeiro (RJ) e São Paulo (SP)

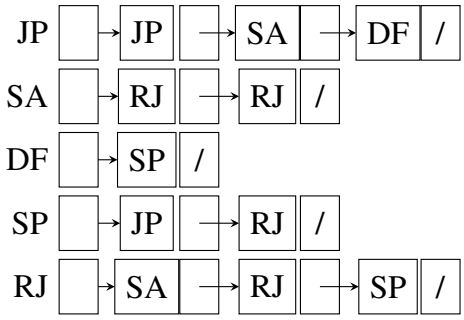
Um fato importante é que no currículo obtido da plataforma Lattes é possível coletar outros registros de localidade que poderiam compor a mobilidade do pesquisador. Por exemplo, os registros de participação em projetos de pesquisa e os registros das apresentações de trabalho científico podem implicitamente referenciar uma localidade respectivamente por meio do nome e ano do evento, e através do nome da instituição. Entretanto, por uma questão de simplificação do modelo de mobilidade do pesquisado desta pesquisa foram considerados apenas os RN, RF e RT na composição da mobilidade, conforme já foi citado.

Também vale salientar que por questões de otimização da visualização do MMP, quando se existe múltiplas arestas no mesmo sentido de uma cidade para outra, pode-se realizar a junção de todas em uma única aresta. Este agrupamento pode ocorrer entre várias ocorrências de cada um dos três tipos de fluxo, de um ou mais pesquisadores existentes em um percurso. Por exemplo, na Figura 2.7a existem dois fluxos de Salvador para o Rio de Janeiro que podem ser agrupados conforme a figura Figura 2.7b.

O agrupamento é um recurso opcional para visualização dessa mobilidade, mas se faz necessário principalmente em alguns casos que acumulam alta densidade de aresta, como por exemplo, no deslocamento de doutores, no contexto da cidade, a Plataforma Lattes registra 72.905 fluxos de São Paulo até o Rio de Janeiro, sendo 775 FNF, 41.805 FFF e 30.325 FFT. Portanto, neste cenário o agrupamento de arestas ajuda a entender melhor o padrão da malha

de instituições de ensino e pesquisa através da visualização de uma única aresta.

Além dos nós e arestas, a mobilidade presente nas Figuras 2.7a e 2.7b podem ser visualizadas através das estruturas de lista e matriz de adjacência presentes, respectivamente, nas Figuras 2.8a e 2.8b. Na lista, sua organização é feita pelos nós encandeados com cada um dos nós ao qual ele se conecta; já na matriz, a leitura deve ser feita considerando que a ordem dos fluxos ocorre dos nós descritos na abcissa até os nós das ordenadas.



(a) lista de adjacência

	JP	SA	DF	SP	RF
JP	1	0	0	1	0
SA	0	0	0	0	1
DF	1	0	0	0	0
SP	1	0	1	0	1
RJ	0	2	0	1	1

(b) matriz de adjacência

Figura 2.8: Representação da Mobilidade

Na Figura 2.8 as arestas múltiplas, que representam a ocorrência de mais de um fluxo, seja de vários tipos ou de uma ou mais pessoas, são visíveis pela repetição de nós na lista e com valores maiores que um na matriz. Já os laços são visualizados pela ocorrência do próprio nó em sua lista encandeada e na matriz por valores não nulos na coluna principal. Logo, a matriz pode ser uma ótima alternativa para exibir os deslocamentos com arestas múltiplas sem agrupamentos, pois sua representação é feita apenas numericamente no ponto de ligação dos nós.

A princípio neste trabalho essas estruturas são processadas de modo *ad-hoc*, mas nada impede que elas sejam persistidas e acessadas em alguma base de armazenamento de grafo, tipo o Neo4J[39] (WEBBER, 2012). Por outro lado, sua composição pode correlacionar ou filtrar seus nós e arestas usando algum critério específico, podendo exibir: a mobilidade de pessoas usando uma área de formação, como ciência da computação; um local específico, como pessoas oriundas de estados que possuem baixos índices de PIB e IDH; ou até mesmo limitando que as arestas sejam selecionadas com fluxos de apenas um tipo, como o FFF. Contudo, quando se deseja representar mais de um tipo de fluxo de modo simultâneo nas

[39]http://neo4j.com/

estruturas da Figura 2.8 se deve utilizar mais de uma lista ou matriz por tipo.

Em relação à disposição dos nós da Figura 2.6b, percebe-se que seu posicionamento foi feito de tal modo que as arestas coincidam com as pontes na Figura 2.6a, mas em alguns casos a distribuição circular dos nós permite identificar melhor as arestas, como na Figura 2.7. Entretanto, nos casos com grandes quantidades de fluxos ou quando o grafo for completo, ou seja, um nó possui ligação com todos os demais, esta abordagem pode não ser útil para compreensão de todos as característica e padrões da mobilidade, pois algumas informações só ficam mais visíveis apenas quando se extrai algumas métricas ou visualizações aprimoradas de mobilidade.

A título de exemplo, pode-se verificar que na mobilidade de todos os registros dos doutores no contexto dos estados brasileiros, representada na Figura 2.9, é difícil identificar que no multigrafo não existe completude, pois 11 estados não possuem conexões de entrada e saída para todos os demais. Além disso, esta visualização não favorece a identificação de que os 9 maiores fluxos entre os estados ocorrem sem sair de si mesmo, e nem que a maioria das rotas de fluxo se encontra na região sudeste. Por isso que se faz necessário selecionar métricas e visualizações que ajudem a melhor apresentar esses fatos.

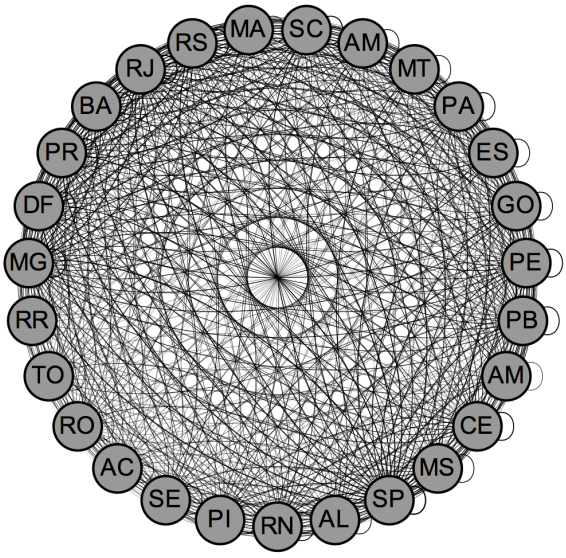


Figura 2.9: Mobilidade dos pesquisadores no contexto dos estados brasileiros

2.5.1 Métricas

Devido a dificuldade de interpretar informações ocultas ou embaralhadas visualmente no dígrafo da Figura 2.9, fica claro que é necessário selecionar algumas métricas para viabilizar a análise de mobilidade. A princípio, selecionou-se valores de mobilidade em termos espaciais, temporais e de ligação no grafo. Estas medidas serão utilizadas como métricas para caracterizar o deslocamento e também podem ser filtradas por critérios específicos.

Através da análise da ligação de um grafo é possível indicar padrões que já foram amplamente investigados pela Análise de Rede Social (ARS) (SCOTT, 2012). Inúmeras métricas de ARS podem ser usadas para ajudar a descrever e caracterizar o grafo, e um bom exemplo seria a Centralidade de Grau (CG). Esta métrica consiste no número de ligação incidentes em um nó de um contexto específico, de tal forma que um nó v pode ser definido pela Equação 2.1 (SCOTT, 2012):

$$CG(v) = \sum_{u=1}^{n} a_{vu} \tag{2.1}$$

em que a_{vu} indica a aresta entre o nó v e o u, de forma que u representa todos os nós do grafo enumerados de 1 até seu número de nós n. Portanto, se a aresta a_{uv} existir seu valor será computado, caso contrário será desconsiderado. Se o grafo for orientado é possível filtrar a contabilização no nó para considerar apenas fluxos de entrada ou de saída, gerando respectivamente as CG de entrada (CG_e) e saída (CG_s).

Então, a partir da CG é possível identificar que na Figura 2.9 o multigrafo possui n igual a 27 e não é considerado completo porque nem todo nó possui as 26 arestas de saída e entrada para os demais nós, incluindo mais uma aresta de seu laço, ou seja, CG igual a 53.

Mas um fator que pode alterar o resultado do CG é a utilização de agrupamentos, pois caso seja utilizado este recurso, o cálculo deve ser feito considerando a aresta agrupada como sendo um e não o número de arestas agrupadas.

Entretanto, apesar da composição coletiva das arestas agrupadas na mobilidade dos fluxos gerados por meio do Lattes, existe um aspecto espacial e de individualidade nos fluxos, pois cada pessoa contribui com suas localidades em todo o seu trajeto. Portanto, é possível analisar, a partir de um indivíduo específico, o seu Número de Deslocamentos (ND) e sua Distância do Deslocamento (DD).

Logo, definido que o multigrafo orientado de todos os indivíduos é $G = (V, A)$, no qual V representa os vértices e A as arestas, então a representação dos fluxos de um indivíduo i será definido por $g(i) = (v, a)$, onde v e a representam os vértices e arestas do seu fluxo, de tal forma que $g(i) \subset G$. Então, para um indivíduo i o seu ND, que contabiliza quantos saltos um indivíduo fez em seu fluxo, é definido por $ND(i)$ na Equação 2.2:

$$ND(i) = \sum_{j=1}^{\alpha} 1 \tag{2.2}$$

onde α é o número de aresta em $g(i)$, ou seja, representa a.

Já a DD de um indivíduo i é definida por $DD(i)$ que computa a distância de todas as aresta de i através da Equação 2.3:

$$DD(i) = \sum_{j=1}^{n} d_j \tag{2.3}$$

onde n representa o $ND(i)$, e d_j a distância da aresta entre os vértices enumeradas de 1 até n. Entretanto, como os vértices da aresta d_j refletem posições no globo, logo seria necessário usar a fórmula de Haversine para definir o seu real valor, visível em $d_h(v_1, v_2)$ através da Equação 2.4:

$$d_h(v_1, v_2) = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{2.4}$$

onde calcula, no globo de raio r, a distância em sua superfície entre dois pontos, v_1 e v_2, de latitudes φ_1 e φ_2, e longitudes λ_1 e λ_2. Portanto, a aresta d_j pode ser substituída pela fórmula de Haversine para se calcular o DD(i) por meio da Equação 2.5:

$$DD(i) = \sum_{k=1}^{m-1} d_h(v_k, v_{k+1}) \tag{2.5}$$

onde m representa o número de vértices de $g(i)$, v_k o vértice enumerado segundo a ordem cronológica dos eventos de i a partir de 1 até $m - 1$, e $d_h(v_k, v_{k+1})$ a distância entre os vértices v_k e v_{k+1} definida pela Equação 2.4. Neste caso, a $DD(i)$ pode ser expressa em unidade derivada do metro.

Um fato importante é que como se sabe que o ND contabiliza os saltos no fluxo de um

indivíduo, e implicitamente pode-se refletir o Número de Locais (NL) visitados por meio da Equação 2.6:

$$NL(i) = ND(i) + 1 \tag{2.6}$$

e quando se exclui da contagem de $ND(i)$ as arestas que representam deslocamentos repetidos, obtêm-se o ND distintos (ND_d), e por consequência também se obtém o NL distintos (NL_d).

Por fim, existe o aspecto temporal aplicado às formações, pois os cursos no Lattes geralmente são registrados com sua duração, logo é possível analisar a mobilidade na escala temporal e, com isso, se extrair o ano de início e fim de uma formação para calcular o seu Intervalo de Duração (ID). Então, para se obter o ID de todos os eventos de i é necessário definir $ID(i)$ na Equação 2.7 como sendo:

$$ID(i) = \sum_{j=1}^{n} \Delta I_j \tag{2.7}$$

onde n representa o número de arestas de $g(i)$ e ΔI_j a duração do evento na aresta j enumerada de 1 até n. Vale salientar que $ID(i)$ não incluem as brechas de intervalos de tempo que podem aparecer entre os eventos de cada aresta, pois esta inclusão na realidade representaria o tempo da primeira formação até a última, ou seja, o ID total (ID_t), sendo necessário apenas para esse cenário os dois anos para seu cálculo.

Outra possibilidade de medida temporal consiste no Número de Instâncias (NI) contabilizado em um intervalo de tempo. Por exemplo, é possível determinar o NI de doutores que se formam num ano específico ou nas últimas décadas. Para isto é preciso definir $NI(i)$ na Equação 2.8 como sendo:

$$NI(i) = \sum_{j=a}^{b} I_j \tag{2.8}$$

onde I_j representa o número de eventos em $g(i)$ num intervalo de tempo enumerado de a até b, e por fim se soma o $NI(i)$ de todos os $g(i)$ de G.

Entre as métricas citadas acima é possível se aplicar agrupamentos dos valores obtidos em cada $g(i)$, seja através de média, ou mesmo somatório como no último caso retratado em

$NI(i)$. Além disso, as métricas que são geradas por meio de somatório podem realizar sua contabilização de modo instantâneo, parcial ou integral, como por exemplo, no NI as três possibilidades poderiam ser geradas respectivamente em um ano específico, um intervalo de tempo e em todos os anos registrados. Finalmente, quando os valores de uma métrica tendem a possui variação muito extrema, pode-se aplicar alguma transformação na escala de valores para melhorar sua visualização, por exemplo, na CG se poderia aplicar a escala logarítmica ($\text{CG}_{[\log]}$).

2.5.2 Visualização

A ideia de visualização de mobilidade se assemelha aos da métrica no sentido de que auxiliam a compreensão dos fluxos, tornando-se em alguns casos complementar e até indispensável para determinadas interpretações (STEELE; ILIINSKY, 2010). A grande questão sobre a importância de se criar visualizações aprimoradas ocorre pelo fato de que a exibição desordenada dos nós e arestas do multigrafo orientado nem sempre ajuda a entender certos padrões ou comportamentos da mobilidade de forma clara e objetiva, como foi demonstrado na Figura 2.9.

Além disso, gráficos clássicos como *boxpot*, histograma e de dispersão ajudam a entender melhor as tendências e distribuição de valores de uma métrica, porém não ajudam a ilustrar literalmente o fenômeno do deslocamento. Portanto, no sentido de instruir melhor esse fenômeno, foram utilizados alguns gráficos que enaltam de forma mais objetiva a ideia de mobilidade, permitindo auxiliar a sua compreensão e visualização (ABEL; SANDER, 2014; SCHICH et al., 2014; RATTI et al., 2010). Entre os gráficos existentes, foram selecionados: o Grafo de Mobilidade, o Mapa de Calor e o Mapa de Fluxo Circular, para atingir tal propósito.

Vale destacar que a maioria das ferramentas e estratégias utilizadas para geração dessas visualizações estão detalhadas no Apêndice A.

Grafo de Mobilidade

O Grafo de Mobilidade (GM) se trata de uma ilustração do multigrafo direcionado, mas difere no posicionamento dos nós e na personalização da aparência dos nós e arestas. Deste

modo, a ilustração clássica do multigrafo da Figura 2.9 pode ser transformado na representação da Figura 2.10a, onde os nós são posicionados conforme sua coordenada geográfica (SCHICH et al., 2014; RATTI et al., 2010).

Quanto à estrutura, basicamente existe a equivalência nos nós e arestas, no qual os nós representam os estados com RN, RF e RT nos currículos dos doutores, enquanto que as arestas rastreiam os FNF, FFF e FFT entre as entidades envolvidas. O mesmo procedimento pode ser aplicado para os demais contextos de país, continente, região e instituição.

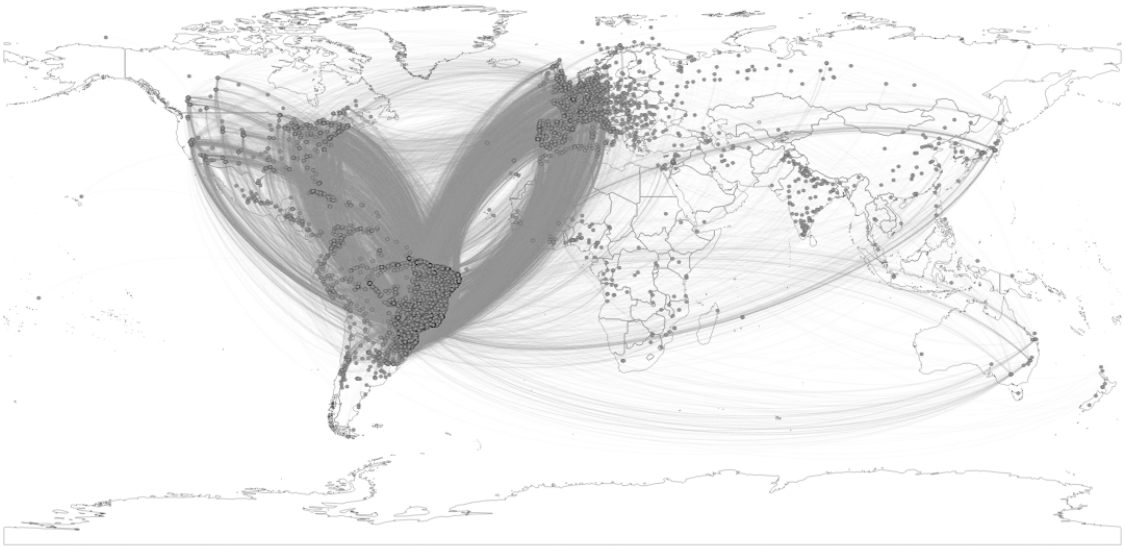
No GM da Figura 2.10a também se faz necessário a utilização de agrupamento de fluxos para facilitar a visualização da rede de mobilidade, principalmente nos casos em que a densidade de fluxos é grande como na Figura 2.10b, que exibe a mesma mobilidade no contexto da cidade, onde mesmo com a utilização do agrupamento ainda exibe as transições mas de modo condessado.

Um benefício do posicionamento dos nós consiste no fato de associar as interpretações a aspectos geográficas, ou seja, fica mais claro na Figura 2.10a que existe uma maior quantidade de fluxos na região sudeste do Brasil, que corresponde com a realidade nacional. Mas tal interpretação deve ser validada com o uso de métricas, pois o mesmo raciocínio aplicado na Figura 2.10b pode levar a interpretação que a maioria dos fluxos entre os doutores envolve instituições no exterior, e que de fato não corresponde com a realidade, pois menos de 10% dos fluxos envolvem localidades no exterior. Uma explicação para esse fato na Figura 2.10b é porque os fluxos externos acabam sobrepondo visualmente o fluxo nacional, o que chama mais atenção, especialmente, porque esse fluxos possuem uma grande extensão.

No entanto, existe uma alternativa de mesclar as métricas no GM, através da associação de valores de algumas métricas no tamanho e cor dos nós, arestas e textos descritivos. Na Figura 2.10a é possível ver essa associação no qual o número de aresta agrupadas é expresso pela variação do tamanho e cor da arestas, já a CG do nó é expressa pela variação do tamanho e cor dos nós, e o tamanho do texto descritivo dos nós. Resultado, fica mais claro identificar que o fluxo mais intenso no Brasil se concentra na região sudeste, porque lá apresenta uma alta densidade de arestas com cores e larguras mais intensas; e nós com cores e tamanhos mais intensos. Já o estado de São Paulo possui mais ligações por causa do texto descritivo de maior tamanho.

Esta associação ajuda consideravelmente a interpretação da mobilidade, pois na Figura

(a) estados brasileiros



(b) cidades

Figura 2.10: Mapa da mobilidade dos doutores cadastrados na plataforma Lattes

2.10b, que não utiliza esta abordagem, fica difícil identificar que a cidade de São Paulo possui o maior fluxo. Porém, mesmo com o aprimoramento do GM usando esta abordagem, ainda existem limitações nesta visualização, ou seja, até então não é possível exibir que o multigrafo da Figura 2.10a não é completo. Além disso, a dependência de posicionamento espacial gera, nos casos de alta densidade de fluxo, impossibilidade de localização das arestas e nós de modo explícito. Outra fato relevante é que por questões de nitidez na visualização da mobilidade da Figura 2.10 as setas que representam a orientação do fluxo são omitidas de sua extremidade, mas nada impede que as setas possam ser utilizadas no GM.

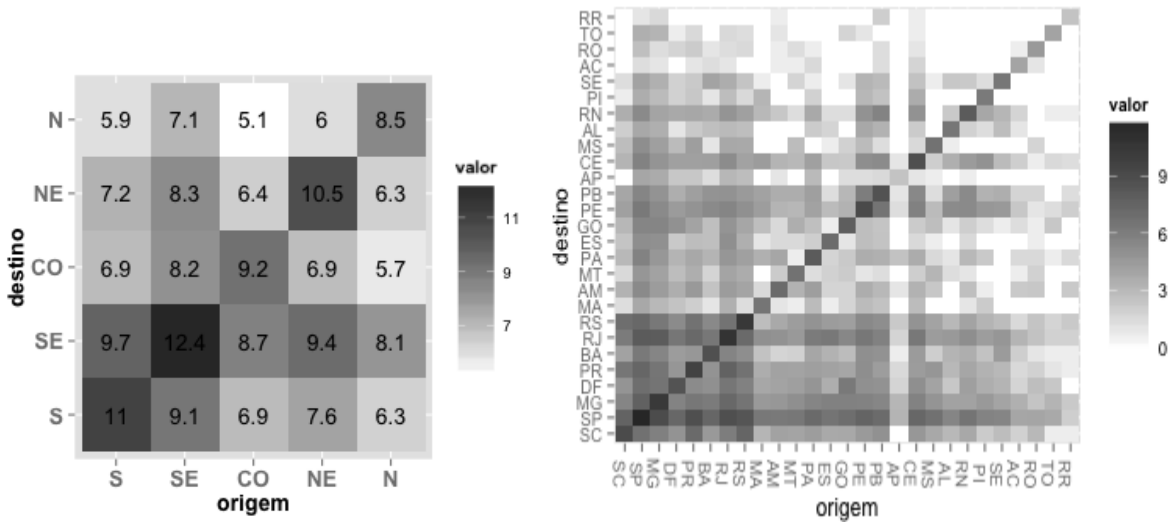
Outra estratégia aplicada no GM, devido a personalização de tamanhos de elementos na sua estrutura, é a transformação na escala de valores para determinar uma melhor visualização, como a conversão dos valores na escala logarítmica. Tal fato é importante porque em alguns casos a dispersão de valores chega a ser muito extrema e com padrões variados de crescimento. Um exemplo dessa personalização de escala foi aplicado na Figura 2.10a para determinar os tamanhos e cores de nós, arestas e textos descritos por meio da escala logarítmica. Por fim, toda a possibilidade de aplicação de filtros nas arestas e nós podem ser usadas para se interpretar alguma informação específica.

Mapa de Calor

O Mapa de Calor (MC) é um gráfico que representa a variação de valores de uma métrica presente na relação binária de elementos num conjunto de entidades. Esta relação binária pode ser é expressa por meio de uma matriz que associa elementos dispostos nos eixos da abcissa e ordenada. O ponto da relação binária dos elementos na matriz indica a variação da métrica usando uma escala de cores que cresce de forma gradativa entre o valor mínimo até o máximo (SCHICH et al., 2014).

Inúmeras são as aplicações do MC, e entre elas pode-se encontrar o MC abstraindo a mobilidade devido a sua semelhança com a estrutura da matriz de adjacência do multigrafo de mobilidade, presente na Figura 2.8b, pois em termos visuais o MC apresenta a mesma aparência com exceção do uso da escala de cores. Então, para construir o MC de todos os FFF dos doutores do Lattes no contexto das regiões brasileiras, se enumera as regiões nos eixos da abcissa para expressar o ponto de saída de um fluxo existente para a ordenada como

o ponto de chegada, resultando na Figura 2.11a[40].



(a) regiões brasileiras (b) estados brasileiros

Figura 2.11: Mapa de calor dos doutores cadastrados na plataforma Lattes (log).

Pelas cores na escala é possível apontar que o maior deslocamento ocorreu da região sudeste para si mesma e o menor ocorreu do centro-oeste para o norte. De outro modo é possível observar esse mesmo fato através dos valores que estão expressos na associação junto as cores, no qual o fluxo máximo possui 12,4 e o mínimo 5,1. Contudo, esses valores não indicam a quantia exata de fluxos porque neste cenário foi utilizada a escala logarítmica para que os valores não ficassem com muita variação e dificultando possíveis comparações.

É interessante observar que neste MC das regiões brasileiras é possível identificar a completude da mobilidade, pois todos os pontos possuem valores não nulos indicando que a partir de um nó se encontra fluxo para qualquer outro nó. Além disso, devido a organização da abcissa na parte inferior do gráfico, percebe-se que os laços ocupam a coluna secundária da matriz.

Outra ilustração do MC na abstração da mobilidade está presente na Figura 2.11b, que exibe o mesmo fluxo no contexto dos estados brasileiros. Por questões visuais não foi expresso nesse gráfico o valor da relação binária na matriz, mas analisando as cores fica evidente perceber que os laços, ou seja, as formações ou atuações profissionais no mesmo estado, possuem valores mais substancias se comparados aos fluxos de saído e entrada de um

[40]As regiões Norte, Nordeste, Centro-Oeste, Sudeste e Sul estão respectivamente representados pelas suas siglas: N, NE, CO, SE e S

estado para os demais.

Usando a escala é possível intensificar o valor do fluxo na associação entre dois elementos, mas dependendo do ponto esse valor pode ter significados diferentes. Para os MCs da Figura 2.11 o valor da associação expressa no ponto da abcissa é o CG_s, já para a ordenada indica o CG_e.

A vantagem do MC em relação ao GM consiste em garantir mecanismos mais claros e explícitos para localizar os fluxos e nós devido: à estrutura de associação da matriz; à facilidade em identificar a origem e destino de um fluxo com a projeção do ponto na associação da matriz nos eixos; além da simplicidade em identificar o volume do fluxo devido a variação de cores na matriz de associação em função da escala de cor. Todas as observações aplicadas no GM sobre contextos, filtros e escala também podem ser aplicadas ao MC, mas comparado ao GM, o MC não possui uma percepção visual geográfica dos nós.

Além desta abordagem de abstração da mobilidade, o MC pode ajudar na percepção da evolução de valores de uma métrica num intervalo de tempo, como por exemplo, na Figura 2.12 é possível analisar a evolução do número de doutores formados nas 10 instituições de maior $CG_{[log]}$ desde 1950, no qual cada linha representa uma das instituições e a coluna um ano específico.

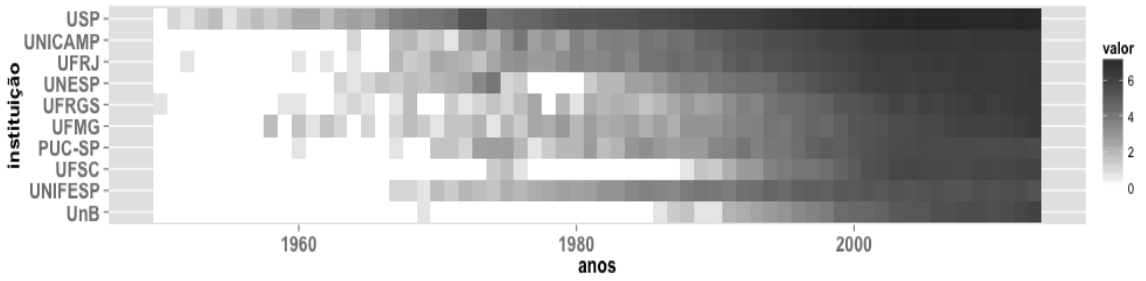


Figura 2.12: Mapa de calor da evolução da formação de doutores nas 10 instituições com os maiores registros (log).

Mapa de Fluxo Circular

O Mapa de Fluxo Circular (MFC) é um gráfico que expressa transições por meio de arestas a partir de um conjunto de pontos distribuídos radialmente. Cada ponto é expresso por uma cor diferente e contém três fluxos, o de saída, permanência e chegada em relação a um ponto,

quando existe. Todos os fluxos de um ponto computam o seu peso em relação ao total para se determinar o tamanho proporcional dos fluxos e pontos (ABEL; SANDER, 2014).

Este gráfico possui relação direta com a mobilidade pois se trata de uma representação do fluxo, sendo uma boa representação da estrutura de lista de adjacência. Então um exemplo gráfico de sua estrutura para representar a transição das pessoas que foram fazer o doutorado entre os estados brasileiros em 2013 está presente na Figura 2.13.

Na Figura 2.13 é possível apontar que São Paulo é o estado com maior peso no fluxo nacional, ocupando um pouco mais de um quarto do fluxo total, e que pela proporção, os fluxos de saída, em vermelho no estado, são menores que os fluxos de chegada, de outras cores, mas juntos os fluxos não são maiores que o fluxo de permanência. Em uma comparação global também pode-se perceber o predomínio dos fluxos internos em relação aos de entrada e saída.

Comparado ao MC o MFC possui a substituição da percepção da intensidade do fluxo com cores e proporção das arestas, e se beneficia na comparação global dos nós, que visualmente são dimensionadas conforme sua parcela de contribuição da mobilidade geral. Mas ambos descartam a questão geoespacial dos nós e não são indicados quando a quantidade de pontos for grande por causa da dificuldade de visualização de sua estrutura e, consequente, obstáculo de interpretação das informações de mobilidade. Além disso, todas as observações aplicadas no GM sobre contextos, filtros e escala também podem ser aplicados ao MFC.

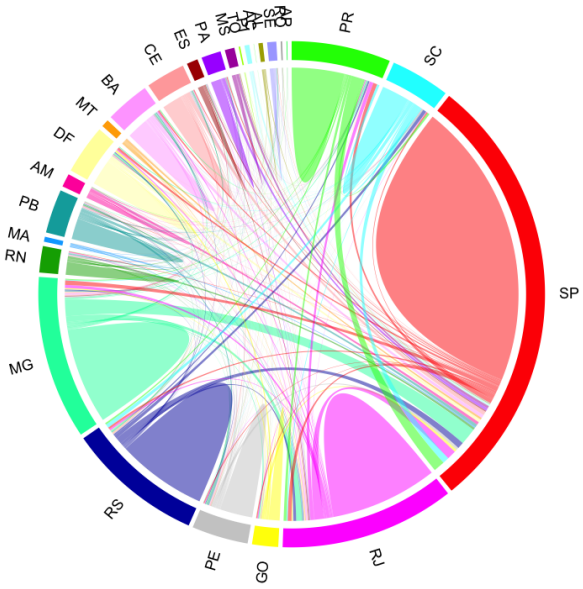


Figura 2.13: Mapa de Fluxo da formação dos doutores entre os estados brasileiros.

2.6 Considerações Finais

A intenção deste capítulo foi de listar um conjunto mínimo de conhecimento acerca dos assuntos abordados nesta dissertação. Nesse sentido, foram abordados inicialmente aspectos inerentes à Ciência dos Dados para uma melhor compreensão dos procedimentos utilizados neste trabalho, com enfoque nos processos de extração.

Além disso, foi apresentado a estrutura da pesquisa no Brasil citando os fatos marcantes que possibilitaram o surgimento do atual panorama e apontando onde se obtêm os dados da pesquisa nacional. Também foi abordado como se estrutura a Plataforma Lattes, fora as estratégias utilizadas para persistir os currículos. Por fim, foi delineada a concepção de mobilidade dos pesquisadores, com ênfase na listagem de métricas e visualizações que ajudam a entender tal concepção.

No próximo capítulo serão apresentados os principais trabalhos relacionados à esta pesquisa, analisando os principais resultados encontrados nos mesmos, bem como um comparativo destes resultados com esta dissertação.

Capítulo 3

Trabalhos Relacionados

Neste capítulo são descritos alguns trabalhos relacionados à pesquisa realizada nesta dissertação. Na Seção 3.1 são apresentadas as pesquisas que investigam a extração de dados curriculares para se analisar a mobilidade dos pesquisadores. A Seção 3.2 destina-se a descrever os trabalhos que realizam a modelagem e visualização da mobilidade de uma forma geral. Já a Seção 3.3 apresenta uma breve discussão sobre os principais resultados encontrados pelos autores e suas contribuições para esta pesquisa.

3.1 Extração de dados curriculares e análise de mobilidade

A ideia de se obter dados de pesquisadores, para analisar sua mobilidade, é algo que tem sido objeto de estudo entre inúmeras pesquisas, e vem conquistando importantes resultados nos últimos anos. Segundo Cañibano, Otamendi e Solís (2011), a análise de mobilidade dos pesquisadores possui registro desde a década de 70, com o dilema da fuga dos pesquisadores de alguns países. Contudo, alguns autores afirmam que, o atual contexto globalizado de incentivo à colaboração internacional e transferência de conhecimento, permitiu um maior fluxo dos pesquisadores, fazendo que os atuais trabalhos foquem mais nas análises da mobilidade.

Cañibano, Otamendi e Solís (2011) também afirmam que existem algumas estratégias que são utilizadas para delinear a mobilidade dos pesquisadores: a princípio algumas pesquisas utilizam a abordagem de uma consulta direta a um conjunto de pesquisadores, via formulários, sobre informações que possam descrever a mobilidade, como a formação, atuação e publicação; outra estratégia consiste em obter o acesso às consultas já realizadas, por

algumas instituições e até países, para avaliar o desempenho dos consultados no intuito de distribuir fundos de pesquisa, ocorrendo quase que de forma periódica; outra opção seria utilizar as bases curriculares que possibilitam o acesso, direto ou indireto, dos registros de formação e atuação do pesquisador; por fim, existe a possibilidade de inferir a mobilidade do pesquisador baseada nos registros de filiação de instituição declarados nas publicações de artigo, e que são acessíveis entre os principais portais de publicação como o Scopus[1], *Web of Science*[2] e IEEE Xplore[3]. Cañibano, Otamendi e Solís (2011) e Cañibano, Bozeman et al. (2009) exibem exemplos de cada uma dessas estratégias.

Ainda em Cañibano, Otamendi e Solís (2011) a fonte de dados de mobilidade foi extraída usando a abordagem da plataforma de currículos espanhol *Scientific Information System of Andalusia*[4], que coletou registros de uma amostra de 10.000 doutores em um período de quatro décadas. Tal pesquisa mostrou-se eficaz na análise por meio de currículos digitais da plataforma, e resultou em algumas descobertas em termos de frequência, duração e destino dos deslocamentos, através das áreas de conhecimento, estágios da carreira e períodos de tempo. O principal resultado apresentado foi que, dependendo da área de conhecimento, a mobilidade possuía fluxos definidos e diferentes entre si.

No trabalho de Moed e Halevi (2014), existe uma dedução dos deslocamentos entre formação e atuação através dos registros de filiação dos artigos disponíveis no Scopus. A princípio utilizou-se um filtro por artigo contento 17 países no período de 2000 até 2012, pois neste sistema é possível se chegar nas instituições através dos países, e por consequência, em seus autores filiados. O resultado foi a coleta de uma amostra de 100.830 pesquisadores, com artigos que possuem filiação nos países selecionados.

Essa abordagem e técnica de inferência de mobilidade usando o Scopus também foram recorrentes em Moed, Plume et al. (2013), que analisam os fluxos que passam pelo Reino Unido; e em Roberge e Campbell (2012), com fluxos que passam pelo Canadá. Entretanto, nesta abordagem existe uma limitação da filiação para se definir a formação de modo cronológico e específico, pois não existe informações associadas à filiação que possam definir que tal registro pertença a um período da formação do pesquisador. Entretanto, é possível

[1] http://www.scopus.com/
[2] http://webofknowledge.com/
[3] http://ieeexplore.ieee.org/Xplore/home.jsp
[4] https://sica2.cica.es/

subtender que existe uma evolução na titularidade ou atuação, diante das várias filiações apresentadas. Contudo, pesquisas como a de Kawashima e Tomizawa (2015) podem ajudar no nível de especificidade das filiações, pois os mesmos, associam ou validam as filiações de publicação de um pesquisador usando outras bases que contêm informações do próprio pesquisador, como a utilização dos registros curriculares. Nesse trabalho, o autor demonstrou que tal técnica apresentou resultados interessantes, e foi validado utilizando-se as filiações do Scopus junto com os registros da plataforma de currículo japonês KAKEN[5].

Diante da análise dos fluxos, Moed e Halevi (2014) afirmam que existe uma relação entre a migração coletada nas filiações de cada autor e o desempenho dos índices do pesquisador em seu perfil no Scopus. Além disso, os autores destacam que grande parte da rede de coautoria formada pelos artigos, possui filiação nos Estados Unidos e China. Os resultados também sugerem que a similaridade da língua entre os países do deslocamento é um fator importante na migração internacional.

Esta última conclusão também foi deduzida por Lepori, Probst et al. (2009), que determinou que existe uma influência entre o destino de atuação no exterior de um pesquisador suíço, e o idioma predominante da sua região de origem. A diferença é que em Lepori, Probst et al. (2009) se utilizaram apenas 67 pesquisadores na análise através de seus registros curriculares e filiações de publicações no *Web of Science*.

O índice de publicação *Web of Science* também foi utilizado por Yamashita e Yoshinaga (2014) e Ponds, Oort e Frenken (2007) para inferir a mobilidade usando as filiações. Contudo, além do *Web of Science*, outros portais de publicação também podem ser utilizados para análise de mobilidade via filiação (FILIPPO; CASADO; GÓMEZ, 2009).

Yamashita e Yoshinaga (2014) tentam comparar a mobilidade de alguns artigos, com alta e baixa citação em inteligência artificial, entre 2004 e 2006, e concluíram que, dependendo da origem de filiação do pesquisador, o artigo irá obter padrões distintos, por exemplo, artigos da China são mais citados no próprio país, enquanto que os artigos da Índia são mais citados no exterior. Vale destacar que os artigos mais citados, geralmente, possuem coautoria de pessoas com mais experiência no exterior e voltaram para o seu país de origem, principalmente, no caso da China. Além do que, os artigos mais citados, geralmente, possuíam pesquisadores com tempo de experiência de mais de seis anos.

[5]https://kaken.nii.ac.jp/en/

Por outro lado, Ponds, Oort e Frenken (2007), utilizam as filiações do *Web of Science* no intuito de montar a rede de coautoria, e assim indicar o grau de colaboração das pesquisas científicas. Para isso, foi selecionado algumas áreas específicas do conhecimento, através de termos de consulta, mas filtrando os artigos que possuíam mais de um registro de filiação, no período de 1988 até 2004, e com no mínimo um registro na Holanda. Através dessa coleta, os autores elaboraram uma rede de colaboração de algumas áreas na Holanda, e constataram que a porcentagem de pessoas holandesas na rede era inferior quando comparada aos demais autores da cooperação internacional, contendo, em grande proporção, pessoas dos demais países da União Europeia, além do mais, cada rede de colaboração por área apresentou padrões diferentes entre si. Quando os autores distinguiram os tipos de instituições perceberam que as universidades possuíam maior proporção na colaboração, em relação a outros tipos de instituições de pesquisa, em todas as áreas. Por fim, o trabalho demostrou que a proximidade entre regiões favorece mais a colaboração em relação à semelhança de instituição.

Já o aspecto da análise e comparação entre produtividade e mobilidade dos pesquisadores é investigado nos trabalhos de Cañibano, Otamendi e Andújar (2008) e Jonkers e Tijssen (2008), onde o primeiro trabalho utilizou o contexto do programa de capacitação espanhol *Ramón y Cajal*[6], para coletar a mobilidade via currículos, disponíveis no Ministério de Educação e Ciência do país. O estudo propôs avaliar os padrões de mobilidade destes pesquisadores para procurar evidências de ligações entre a sua mobilidade e produtividade, então, por meio de dados estatísticos das publicações e os deslocamentos de cada pesquisador, percebeu-se que existia uma relação entre esses dois fatores. Da mesma forma, o trabalho de Furukawa, Shirakawa e Okuwada (2011) propôs um objetivo semelhante, mas usando filiações de artigos do evento do *IEEE Transactions on Pattern Analysis and Machine Intelligence*, durante o período de 1997 até 2009.

Jonkers e Tijssen (2008) analisam o impacto da atuação na produtividade de chineses quando passam um tempo atuando e publicando fora da China e depois voltam para o país. Para isso, os autores utilizaram uma amostra contendo alguns pesquisadores seniores, atuantes em instituições de ponta em Pequim e Xangai. A mobilidade na pesquisa foi extraída de registros curriculares publicados nos relatórios anuais de cada instituição e nas filiações de artigos do *Web of Science*. Contudo, tal procedimento foi associado a uma entrevista, como

[6]http://icc.ub.edu/index.php?m=job&c=ramon_cajal&op=frm_ramon_cajal

forma de solucionar alguns casos omissos ou de incerteza na mobilidade dos pesquisadores. O resultado evidenciado pelos autores foi que existe sim um aspecto positivo na produtividade quando os pesquisadores possuem vivência e contatos no exterior. Esta conclusão também foi importante por causa de suas implicações para o debate sobre a fuga e retorno de pesquisadores, pois ela mostra que a perda de capital humano pode ser parcialmente compensada por novas e intensas relações entre pesquisadores chineses e colegas de trabalho no exterior. O benefício direto é o fortalecimento dos laços entre o sistema de pesquisa chinês e de outros países, incorporando mais relevância da China dentro do sistema científico global.

Também foi possível encontrar no trabalho de Baruffaldi e Landoni (2012) uma análise semelhante do retorno dos pesquisadores ao seu país de origem, usando uma consulta de registros sobre alguns pesquisadores da Itália e de Portugal. O mesmo fato também pode ser observado no trabalho de Gibson e McKenzie (2014).

Por fim, outros trabalhos focam na análise da carreira profissional de pesquisadores (GAUGHAN, 2009; DIETZ et al., 2000), em aspectos de mobilidade em regiões especificas (CONCHI; MICHELS, 2014), e na relação do sobrenome de um pesquisador e sua região de origem (LEWISON; KUNDRA, 2008).

3.2 Estratégias de modelagem e visualização da mobilidade

Uma vez que se sabe como é extraído a mobilidade, é importante utilizar técnicas apropriadas de modelagem e visualização dos deslocamentos no intuito de garantir boas interpretações, pois o uso dos dados depende muito da forma como eles são apresentados. Nesse sentido, muitos trabalhos tentam focar em aspectos visuais e de métricas para compreender melhor os deslocamentos e os fluxos em suas várias perspectivas:

- Mobilidade humana (RAE, 2009);
- Migração em países (ANDRIS, 2011) ;
- Fluxos de ligações telefônicas (RATTI et al., 2010);
- Redes de coautoria (DIGIAMPIETRI et al., 2014a);
- Redes sociais (HEER; BOYD, 2005);

- Transporte urbano (MAYERES; OCHELEN; PROOST, 1996);
- Tráfego da Internet (BHATTACHARYA et al., 2008);
- Tráfico humano (FEDORSCHAK et al., 2014);
- Migração de animais (ROBINSON et al., 2009);
- Fluxo do comércio internacional (KLETTKE; MEYER, 2001);
- Fluxo de epidemia (GUBLER, 2011)

A maioria desses trabalhos possuem uma característica em comum, que é a análise de dados envolvendo registros de localidade. De tal forma que o desafio da maioria, nesse cenário, consiste em exibir informações por meio de mapas e gráficos, uma vez que informações tabulares e sem organização podem não garantir boas interpretações (CHEN; HÄRDLE; UNWIN, 2007).

Por isso, alguns trabalhos tentam exibir informações usando mapas com: o GM, para expressar melhor os fluxos e seus padrões (SCHICH et al., 2014); os vetores de deslocamento, para indicar as várias tendências de deslocamento por região (RESEARCH, 2012); e os mapas do tipo *choropleth*, para associar a intensidade de uma variável em determinadas regiões (BARKER et al., 2013). Mas além dos mapas, outros formatos de visualização podem auxiliar na compreensão da mobilidade, como: o MFC (ABEL; SANDER, 2014), diagramas de Vonoroi (BALCAN et al., 2009), MC (SCHICH et al., 2014), *sankey* (HIRST, 2013), *treemap* (OEC, 2015), diagrama de arcos (SANCHEZ, 2013), ou até junção de mapas com gráficos convencionais, como o gráfico em barra ou pizza localizado em pontos do mapa (GIMENEZ, 2015).

Entre esses trabalhos, que envolvem a análise de dados com localidades, pode-se citar o de Christen (2012), que utiliza a geocodificação para ajudar as empresas a entender melhor onde vivem seus clientes, e onde sugerir a abertura de novas lojas. Além de auxiliar as agências de estatística a indicar o planejamento de novos equipamentos públicos, tais como escolas, hospitais, centros comerciais ou estradas, conforme o crescimento populacional. Outro trabalho seria o de Kumar et al. (2011), que utilizam dados de redes sociais públicas,

interpoladas em mapas, para auxiliar socorristas e a defesa civil no amparo de desastres naturais. Já no trabalho de Koblin (2009), destaca-se as principais rotas e pontos de densidade da aviação civil utilizando GM.

A relevância de se tratar os dados geográficos também pode ser vista no trabalho de Barker et al. (2013), que cria um modelo de predição do índice de diabetes em regiões dos Estados Unidos, através de dados da doença no país. Esse resultado possibilita estimar a probabilidade de habitantes, de um determinado local, possuírem ou não a doença, contribuindo assim com o processo de intervenção precoce. Também no contexto da saúde, Rushton et al. (2006), utilizam a geocodificação na pesquisa do câncer para identificar correlações geográficas entre casos de câncer e fatores ambientais próximos, tais como fábricas de produtos químicos, ou as fontes de radiação, que podem influenciar na ocorrência de certos tipos de câncer.

Mas, o foco deste trabalho se fundamenta na questão da mobilidade dos pesquisadores e apresenta aspectos semelhantes aos encontrados nos mapas de análise de Schich et al. (2014), que detalha o deslocamento nos últimos dois mil anos, entre o local de nascimento e morte, de um conjunto de pessoas notáveis da história mundial. Para isso, foram coletados os tais registros de localidade, de 150.000 pessoas notáveis, nas fontes do Freebase, *General Artist Lexicon* e *Getty Union List of Artist Names*. Entre os resultados da pesquisa, existem algumas associações entre a taxa de entrada e saída de pessoas de algumas cidades ao longo dos anos, a visualização da mobilidade utilizando o GM, e a definição da distribuição de probabilidade de algumas variáveis extraídas.

Contudo, vale salientar que as visualizações de mobilidade em GM de Schich et al. (2014) extrapolaram a essência do deslocamento, fornecendo outros tipos de informações correlatas, como por exemplo, ao rastear o deslocamento de algumas pessoas notáveis na história mundial, acabou-se exibindo alguns fatos históricos da humanidade. Entre esses fatos históricos estão a formação dos grandes centros europeus e a colonização norte americana, que estavam envolvidos com a mobilidade destas pessoas notáveis.

Outro trabalho que também ajuda a compreender a mobilidade das pessoas é o de Abel e Sander (2014), que tenta caracterizar os fluxos bilaterais entre 196 países, de 1990 até 2010, fornecendo uma visão abrangente dos fluxos migratórios internacionais, disponibilizados em dados da Organização das Nações Unidas. Como resultado, os autores sempre constataram

que existe uma estabilização no fluxo migratório mundial, que atinge 0,6% da população. Também existe a representação dos fluxos entre regiões do globo usando principalmente o MFC para visualizar a relação de imigração e emigração no mundo, além de identificar que os maiores fluxos ocorrem entre o sul e o oeste asiático, da América Latina para a América do Norte e dentro da África.

Já no trabalho de Rae (2009), existe um conjunto de representações, incluindo o GM, para ilustrar o deslocamento familiar via os dados do censo no Reino Unido de 2001. Os registros de origem e destino das mudanças de cada família se mostraram eficazes para identificar as principais rotas de deslocamento desse tipo de fluxo que contribuíram para entender melhor a nova configuração, urbana e rural, do Reino Unido. Enquanto que o trabalho de Ratti et al. (2010) através do uso de vinte milhões de registros, de destino e origem, das chamadas telefônicos feitas no Reino Unido, propôs uma nova abordagem de divisão geográfica baseado na rede de interação e comunicação das pessoas, diferente das estratégias geográfica ou econômica existentes. Seu grande diferencial foi propor um algoritmo que pudesse identificar as principais aglomerações da rede de comunicação.

3.3 Considerações Finais

Dos trabalhos analisados sobre a mobilidade dos pesquisadores, boa parte idealizou a utilização de uma amostra para determinar o seu fluxo entre as intuições, que de uma forma ou de outra, participaram da atuação do pesquisador em algum estágio de sua vida. Mas algo em comum é observado. Na maioria dos casos, a amostra geralmente consiste de pesquisadores que possuem um estágio evoluído de sua formação, e com ampla experiência na pesquisa, como no caso dos doutores.

Tal fato pode ser justificado por dois motivos principais, esse tipo de pesquisador possui uma média maior de interação com instituições do que outros tipos. A maior experiência e vivência em projetos de pesquisa implicam que os dados dos doutores possuem mais qualidade na declaração de currículo e, geralmente, contêm publicações com maior fator de impacto. Estes fatos resultam em um cuidado especial ao se interpretar tais resultados deste tipo de amostra, pois inevitavelmente pode existir um viés de que as conclusões se encaixem melhor na amostra, e não ao todo.

Quanto às estratégias de inferência de mobilidade listadas, percebe-se que a abordagem da coleta de informações curriculares possui um benefício, que consiste em tratar os dados desejados direto da fonte, ou seja, dos pesquisadores, mas sua execução é considerada mais complexa devido à dependência de voluntários no processo. Quando se obtém o acesso aos registros de consultas curriculares, realizadas por instituição ou país, de fato o processo fica mais abrangente que o anterior, devido ao número de pesquisadores que participam dessas coletas, mas nem sempre essas informações são públicas. Já a inferência construída pelas filiações em publicações exige que os pesquisadores possuam trabalhos que sejam reconhecidos em tais portais de publicação a cada atuação profissional ou formação, para assim compreender melhor seus fluxos, ou seja, é algo dificilmente alcançado por todos.

Logo, diante das limitações existentes, para inferir a mobilidade de pesquisadores, a estratégia do uso do currículo vitae se torna interessante, pois os registros de filiações são declarados de forma específica, tornando claro o que cada local representa no ciclo de formação e atuação do pesquisador. Além disso, o número de registros nas plataformas de currículo pode alcançar uma margem bastante representativa em alguns sistemas. Mesmo assim, ainda existe limitação quanto à qualidade dos dados declarada manualmente, ou seja, estes dados podem não ser normalizados ou conter erros.

Já no quesito de visualização, percebe-se que muitas alternativas de representar a mobilidade já foram exploradas, e que cada estratégia apresenta seus aspectos peculiares e positivos. Mas no geral, o grande desafio que a maioria tentou resolver foi permitir representar processos especiais, que em muitos casos acumulam um grande volume de fluxos, e dinâmicos, de um modo simplificado e fácil, para assim ajudar qualquer pessoa leiga a assimilar sua compreensão.

Quanto ao presente trabalho, existe a preferência do uso de registros de currículo vitae, mas diferente dos trabalhos anteriores por utilizar uma metodologia de garantia de qualidade dos registros de localidade, para que não se comprometa as interpretações de mobilidade. Já o número de registros utilizados neste estudo, também é considerado relevante se comparado aos trabalhos citados, pois não usa um filtro por área específica. Contudo, neste trabalho se utiliza todo o universo de pesquisadores seniores e com alta titulação no âmbito nacional, com a possibilidade de utilizar todo o contexto de pesquisadores do Brasil.

Na questão da visualização, inúmeros gráficos recentes e relevantes, sobre a mobilidade,

como o GM, MFC e MC apresentados em Schich et al. (2014), Abel e Sander (2014), Rae (2009), Ratti et al. (2010), foram utilizados neste trabalho, e garantiram uma melhor exibição dos aspectos de deslocamento dos pesquisadores, diferentemente dos demais trabalhos deste tipo mobilidade. O que ajudou a identificar outros fatos associados, como o destaque dos grandes centros políticos e econômicos na polarização das pesquisas. No entanto, analisar a causalidade destes fatos não será alvo desta pesquisa.

No próximo capítulo, serão apresentados todos os detalhes da metodologia utilizada para a análise do deslocamento dos pesquisadores da Plataforma Lattes.

Capítulo 4

Construção do Modelo de Mineração

Neste capítulo são apresentadas as principais etapas envolvidas na construção do Modelo de Mineração de Dados (MMD) proposto neste trabalho. A Figura 4.1 apresenta de maneira sucinta a metodologia aplicada para se chegar nas interpretações de mobilidade geradas pelo MMD. Quanto a sua composição, o MMD foi inspirada no modelo de etapas do KDD em Fayyad, Piatetsky-Shapiro e Smyth (1996), mas com uma pequena alteração para iniciar o KDD, devido a inclusão de mais uma etapa antes do seu início, para se obter os dados de mobilidade do Lattes que não estão prontos e precisam ser coletados.

4.1 Obtenção dos dados

A intenção original desta etapa consiste em prover uma cópia dos dados curriculares da Plataforma Lattes através de uma base estruturada. Primeiramente, foi proposto um processo de obtenção dos arquivos XML do Lattes e em seguida, foi realizado um armazenamento em uma base de dados diferente, para garantir independência do Lattes. Por último, foi definido um novo mecanismo de acesso granular às informações contidas no XML de forma mais eficiente.

4.1.1 Obter os Arquivos XML

Para se obter os dados do Lattes foi determinado que, dos formatos disponíveis na plataforma, a melhor escolha seria o XML, devido a estruturação semântica dos dados e porque

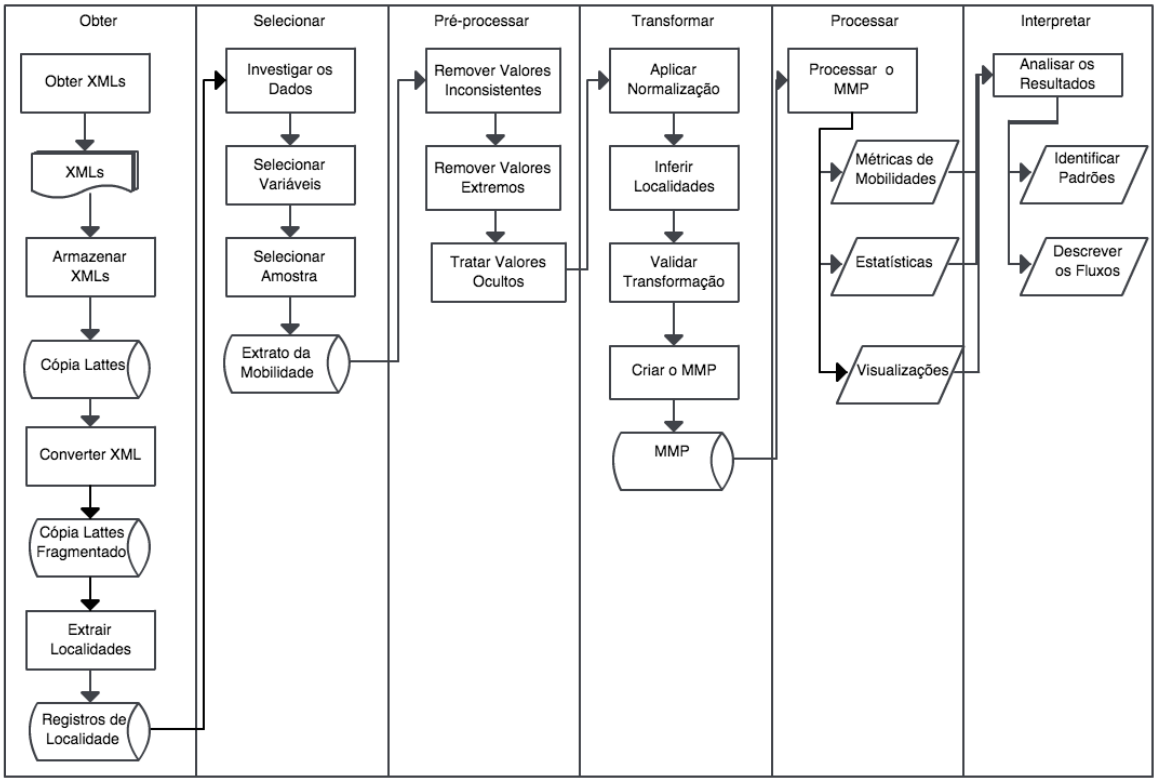


Figura 4.1: Fluxo de extração de informação do MMD.

algumas informações são exclusivas apenas neste formato. Então, conforme já foi descrito no Capítulo 2, uma alternativa de coleta deste arquivo é fornecida pela funcionalidade de consulta de pesquisadores da plataforma, que junto com cada resultado da lista de pesquisa retornada se encontra o ID16 do pesquisador.

Contudo, para fazer a cópia integral do Lattes, era necessário se obter todos os arquivos XML, logo, foi essencial procurar alguma estratégia para se coletar todos os identificadores dos pesquisadores existentes na plataforma. Através de sua análise, foi possível explorar uma alternativa sobre a página de consulta de pesquisadores usando o princípio do *SQL Injection*, com um carácter vazio que listava indiretamente todos os ID16 existentes na plataforma. Portanto, a coleta desses ID16 associada a um *Web Crawler*, conseguiria simular um usuário baixando todos os arquivos XML da plataforma, atingindo o propósito inicial.

Então, para criar um *Web Crawler* foi necessário analisar os fluxo das requisições HTTP a partir de um ID16 até se obter um currículo em XML. Porém, para evitar o *download* de modo automático, o próprio Lattes exige na requisição que se interprete visualmente um CAPTCHA, para evitar eventuais negações de serviço. Contudo, mediante a aplicação de

algoritmos de reconhecimento de padrão dos CAPTCHAs, gerados pelo Lattes, é que se atingiu uma taxa de acerto do seu reconhecimento com eficiência média de 75%, e nos casos de erro, se repetia a interpretação até se chegar ao sucesso.

Todo o processo e análise descritos culminaram no Algoritmo 1, que permitiu a criação do *Web Crawler* para a cópia integral dos arquivos XML do Lattes, através da função $copiarXMLs()$ na linha 1. Essa metodologia garantiu que cada currículo pudesse ser acessível de forma local, sem a dependência dos servidores do Lattes, precavendo qualquer mudança da estrutura e acesso à plataforma no momento das análises, e até a dependência de disponibilidade do Lattes.

É importante ressaltar que, devido a cópia integral, o MMD pôde suportar a extração de qualquer dado do currículo, antecipando qualquer eventual necessidade de outros dados que não estejam relacionados diretamente com os registros de mobilidade do presente trabalho.

```
1  Function copiarXmls()
2      ids16 = obterTodosOsId16DaPaginaDeConsulta()
3      while !todosOsCurriculosForamSalvos() do
4          for id16 ∈ obterIdsNaoExistentesNoBanco(ids16) do
5              id10 = obterID10NoCurriculoHtml(id16)
6              xml = obeterXml(id10)
7              salvarDados(id16, id10, xml, ultimaAtualizacao,
8              bolsaCnpq, formacoes, tamanhoXml)
9          end
10     end
11 Function obterXml(id10)
12     captcha = obterImagemCaptcha(id10)
13     do
14         xmlComprimido = decodificarCaptcha(captcha)
15         xml = descomprimirXml(xmlComprimido)
16     while !xmlValido(xml) & numeroDeTentativa < k
17     return xml
```

Algorithm 1: Processo de extração de XML da Plataforma Lattes.

Concluída a criação do *Web Crawler*, o próximo passo foi planejar o fluxo de *download* e o armazenamento do XML. Tal esforço resultou no extrato de *download* da Figura 4.2, realizado do dia 29/09/2014 até 25/10/2014, ou seja, quase um mês de extração.

Esse processo exigiu a superação de alguns desafios, principalmente devido: ao fluxo de solicitações, à dependência de acesso Web, à otimização da concorrência em múltiplos

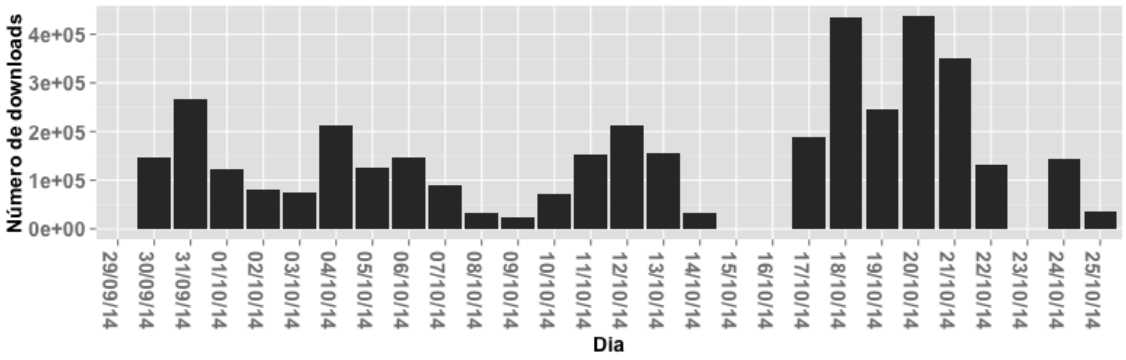


Figura 4.2: Execução de *download* de currículos por dia.

downloads, e às limitações de *hardware*. Como forma de contornar os problemas citados e acelerar o processo de coleta, Boldi et al. (2004), Thelwall (2001), Heydon e Najork (1999), sugerem a utilização de *threads* para aumentar a concorrência do uso de *hardware*, enquanto se espera que um XML seja entregue pela Web. No entanto, para não degradar o desempenho de consumo de memória, processamento, acesso ao disco e rede, devido ao grande volume de dados a ser consumido, também foi necessário ajustar alguns limites da concorrência com uso de *thread pool* e, de forma semelhante, a persistência dos arquivos XML também exigiu um controle da quantidade de acessos ao banco, por meio de um *connection pool*.

Além disso, houveram ajustes na quantidade de *threads* no acesso ao Lattes e na conexão ao banco local como forma de otimizar a velocidade de *download*, pois altos números de conexão ao Lattes geravam perdas e alta latência da resposta, e muitos acessos ao banco, também prejudicavam o desempenho do mesmo devido às operações de disco. Inclusive nas *threads*, foi identificado que alguns fatores, como excesso de saída no console para fins de acompanhamento do estado da extração, deveriam ser eliminados pois prejudicavam a fluidez do *download* em lote.

Também existia a preocupação de que o alto fluxo de requisição se enquadrasse como algo atípico, podendo até ser interpretado como algum princípio de negação de serviço pelos administradores do Lattes, ou da rede local de acesso à Web, acarretando em alertas nos sistemas de detecção de intrusão[1]. Tal situação poderia promover uma reação com limitações nas requisições advindas da rede de origem do *Web Crawler*, o que acarretou em constante monitoramento do fluxo de *download*.

[1]Este sistema deriva da aplicação em inglês *Intrusion Detection System* que representa o sistema que tenta descobrir em uma rede acessos não autorizados.

Não ficou claro se essa prática era adotada pelos responsáveis do Lattes ou da rede local, mas se sabe que, ao aumentar os níveis de requisição, fatalmente se perdia o ritmo do *download*, e em algumas situações, houve a perda de acesso à plataforma apenas na rede do *Web Crawler*. Quanto ao melhor pico de download, foi alcançado na marca de 436.300 currículos/dia, quase 11% do total. Mas é importante notar também no gráfico que em alguns dias não houve *download* ou apresentou valores baixos, motivados principalmente pela indisponibilidade do Lattes e do acesso Web, além das questões de degradação do sistema operacional com alguns excessos de *swap*.

Toda problemática citada acima pode favorecer a possíveis perdas no processo de cópia dos currículos, o que exigiu que a extração e cópia do XML deveria ocorrer com um tempo limitado. Nos casos de insucesso, deveria-se colocar as requisições com perda no final da fila de *download*, para ser extraída e copiada em um instante posterior.

Outra observação importante é o fato do *Web Crawler* ter sido executado na mesma máquina do banco que foi usado para guardar os currículos. Isso fez com que ambos competissem pelos mesmos recursos de *hardware* e, considerando o volume de dados, o ideal é que a execução do *Web Crawler* e a persistência do XML fossem realizadas em máquinas separadas, ou com maior capacidade. Mesmo assim, ainda havia a possibilidade de se usar múltiplas instâncias do *Web Crawler* para otimizar ainda mais o fluxo do *download* (BOLDI et al., 2004; THELWALL, 2001; HEYDON; NAJORK, 1999).

4.1.2 Armazenar e Converter os Arquivos XML

Uma vez que se obteve o XML, o próximo passo consistiu em sua persistência. Quanto a escolha do banco, foi decidido usar o Postgres[2], devido a existência do tipo de dado XML, com opção de sua consulta via Xpath; robustez na indexação da consulta; e na taxa de compressão deste Banco de Dados Relacional (BDR). Logo, por meio de cada XML, extraiu-se a seguinte tupla para ser armazenada no Postgres, conforme a linha 8 do Algoritmo 1:

$$(id16, id10, xml, ultimaAtualizacao, bolsaCnpq, formacoes, tamanhoXml)$$

no qual os dois primeiros valores se tratavam das identificações dos pesquisadores no Lattes, o terceiro valor era o próprio arquivo XML, e os demais valores eram informações cole-

[2]http://www.postgresql.org/

tadas do HTML e do XML através do MMD. De todos os dados, as quatro últimas informações eram processadas para facilitar a filtragem de currículos mediante tais aspectos, ou seja, caso fosse necessário obter apenas pessoas com titulação de mestrado, bastava utilizar o valor $formacoes$ que descrevia uma lista de formações e suas respectivas quantidades, comportando-se como um índice para acessar o XML do currículo.

Outro fato importante que ocorreu nessa etapa foi um aparente vazamento de memória, ocasionado devido ao funcionamento de forma inadequada do *Object-Relational Mapping* (ORM) utilizado para guardar a tupla, e que possivelmente foi gerado e agravado graças ao grande volume de dados trafegado para o BDR e seu grande acesso concorrente. Tal vazamento foi identificado por causa de alguns excessos de *swap* no sistema operacional e lentidão na taxa de *download*, que fizeram com que as suspeitas caíssem sobre o ORM. Especificamente, no seu processo automático de limpeza das tuplas já persistidas na memória, pois seu escalonamento de limpeza automática não estava suportando o grande volume de dados gerado no *Web Crawler*. Então, mesmo que a ativação manual da limpeza do ORM não seja necessária em cenários normais onde a limpeza funciona de forma satisfatória, neste trabalho, essa ativação se fez necessária e permitiu maiores marcas de *download* nos últimos dias de extração dos currículos.

Contudo, havia também questões de desempenho no acesso ao BDR. Antes de tudo, é importante frisar que a forma mais simples de se extrair um conjunto de dados de um lugar para outro é selecionando o dado um a um da origem, para ir inserindo um a um no destino. Logo, a velocidade com que se lê é a mesma com que se escreve. Na prática, esta técnica se mostrou muito ineficiente nesta etapa, pois disparar inúmeros acessos contento um único XML gerava muitas requisições ao BDR sem muita ação efetiva. Após esta percepção, chegou-se a conclusão que a inserção em lote seria mais eficiente, e acabou sendo a forma desejada de operar a inserção dos arquivos XML no banco. Então, sempre que possível, as escritas e leituras no BDR eram realizadas em lote porque era mais eficiente do que inúmeros acessos singulares.

Além disso, também existe a possibilidade de otimizar o desempenho da persistência no DBR realizando, a priori, configurações que aloquem mais recursos da máquina, ou ponderando as questões de arquitetura de armazenamento do XML no BDR, como proposto por Burzańska et al. (2010), Florescu e Kossmann (1999).

Dessa forma, o MMD gerou a base Cópia Lattes com todos os currículos da plataforma. Entretanto, tal cópia não garantiu boa eficiência no acesso granular dos dados curriculares, pois devido as proporções da base, uma pequena consulta de informação, como o nome de todos os pesquisadores, poderia facilmente ultrapassar um dia de execução. Isso exigiu um processo de transformação do XML para alguma estrutura mais fragmentada, e que permitisse melhor tempo de acesso as suas informações, como o proposto por Dukovich et al. (2008), através do cenário da Figura 2.5b usando o HyperJAXB.

Então, o uso do HyperJAXB3, atrelado com os arquivos XML da base Cópia Lattes e o XSD gerado pelo DTD do Lattes, resultou na geração de 321 tabelas de forma automática, disponibilizando todos os registros dos arquivos XML nas suas respectivas tabelas no BDR. Com isso, seus dados passaram a ser acessíveis de forma mais fácil, se comparado com o acesso via Xpath na base Cópia Lattes, devido as consultas de Hibernate usando o mapeamento dos objetos de entidades do XML. Como na consulta a seguir dos nomes dos cursos de pós-doutorado de um currículo com id específico:

```
1 //Leitura baseado no JPA do Hibernate
2 EntityManger loadManager = entitnyManagerFactory.createEntityManager();
3 CurriculoVitaeType curriculum = loadManager.find(CurriculoVitaeType.class, id);
4 DadosGeraisType dg = curriculum.getDADOSGERAIS();
5 FormacaoAcadecimcaTitulacaoType f = df.getFORMACAOACADEMICATITULACAO();
6 f.getPOSDOUTORADO().getNOMECURSOINGLES();
```

ou diretamente via consultas nativas do DBR, como no esquema da consulta SQL que apresenta cinco junções de tabela na Figura 4.3, ou seja, um consulta mais complexa.

Tal resultado gerou outra base, que refletiu o esquema proposto pela Figura 2.5b, que é a Cópia Lattes Fragmentada, e representou um significativo avanço no acesso de dados curriculares específicos, em relação à base Cópia Lattes, pois neste momento não era preciso processar todo o XML, mas sim as tabelas que possuíssem a informação desejada. Mesmo assim, ainda havia uma dificuldade aparente no acesso de tais informação, pois a estrutura e navegação dos dados nas tabelas ainda refletia o XML, como se percebe no código de acesso dos cursos de um currículo de id específico. Além disso, existiam dados que não estavam armazenados de forma eficiente, como os nomes de localidade, que se repetiam a cada currículo ao invés de serem armazenados de forma centralizada, conforme os conceitos de normalização de BDR.

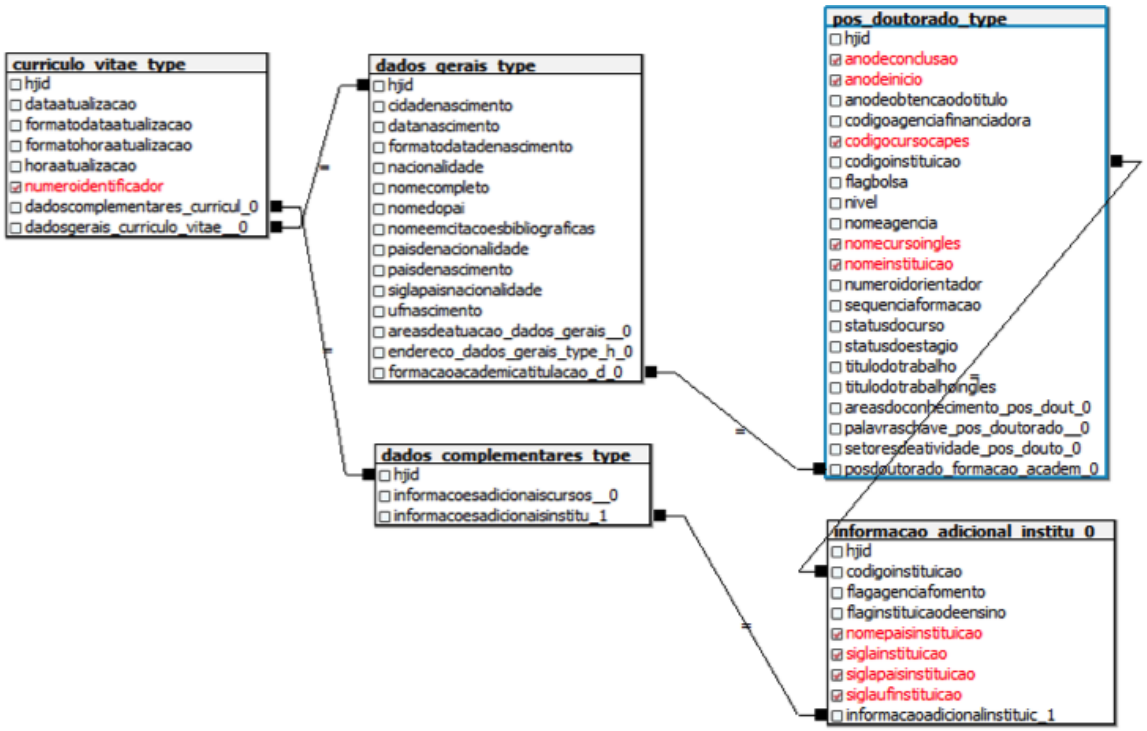


Figura 4.3: Esquema SQL da consulta dos nomes dos cursos de pós-doutorado de um currículo com id específico.

Visando resolver essa questão, foi decidido pré-selecionar alguns dados que possuíssem os Registo de Nascimento (RN), Registro de Formação (RF) e Registro de Trabalho (RT) que compõe o MMP para reduzir a quantidade de dados curriculares na base Cópia Lattes Fragmentada, e seu complexo mecanismo de consulta. A princípio, foram coletados todos os registros de localidade nos currículos formando a tupla:

$$id16, cidade, estado, pais, local, anoInicio, anoFinal, tipoDeRegistro$$

no qual armazenava a identificação do pesquisador, sua posição geográfica, o nome do local que envole o registro, o período de duração, e o tipo do registro seja ele um RN, RF e RT, de tal forma que nos casos de RF se informava especificamente o tipo de formação. Todas as tuplas foram armazenadas em uma única tabela na base Registro de Localidade, resultando em um acesso quase que instantâneo de qualquer informação de localidade dos pesquisadores da Plataforma Lattes.

Claro que o ideal seria ter o acesso ao banco de dados do Lattes para se extrair as informações granulares de localidade, mas o que se percebeu é que, mesmo com o uso dessa alternativa descrita até então, a da manipulação dos arquivos XML para gerar a base Registro

de Localidade, o resultado obteve acesso eficaz e eficiente aos dados pretendidos no presente trabalho. Mesmo que se houvesse a aprovação burocrática do módulo de Extração do Lattes, supostamente haveriam problemas quanto aos limites de acesso, e quanto ao formato dos dados disponibilizados que seria feita usando o mesmo arquivo XML que foi obtido no atual processo de coleta, e não através de um serviço de consulta granular a seu banco de dados (CNPQ, 2015).

Essa aparente limitação do acesso direto à base de dados e o incentivo do uso das cópias dos currículos em XML na Plataforma Lattes, podem ser explicados por questões de distribuição e processamento do seu grande volume de dados. Para melhor entender tal fato, é preciso visualizar que em um cenário de pequeno volume de dados, um computador, que geralmente possui capacidade de processamento superior ao de armazenamento, pode atender a necessidade de extração de pequenas demandas de informação usando mais processamento do que armazenamento, caso seja necessário. Entretanto, em um cenário de grandes volumes de dados, pode ser inviável depender do processamento para se obter informações dos dados usando essa mesma estratégia, pois isso demandaria um tempo não aceitável para geração da resposta, e pode ser agravado, principalmente, nos casos de acesso concorrente e escalável em número de usuários (CHANG et al., 2008; DEAN; GHEMAWAT, 2008).

Portanto, a disponibilização de cópias de informações em grande volume de dados pode se tornar mais eficiente do que processá-los a cada instante, principalmente, quando se sabe previamente a informação desejada, e se pretende evitar a complexidade de escalonamento de processo entre inúmeros usuários que a plataforma possa atender. Então, a estratégia de fornecer a visualização da atualização do currículo, seja ele por meio de HTML ou XML, depois de um tempo ao qual a visualização foi feita, pode transparecer que esse artifício seja aplicado na plataforma, porque já se sabe a estrutura da informação curricular requisitada.

Logo, nesse cenário ilustrado seria mais econômico concentrar as atenções na disponibilização das visões estáticas para o amplo acesso de seu público, evitando uma alta centralização da base de dados do Lattes para gerar as visões de forma dinâmica a todo instante, e que exigiria um alto custo de processamento. Em outras palavras, é como se fosse mais econômico disponibilizar cópias de uma informação já processada do que processá-la a cada instante.

Mesmo que a atualização não seja algo constante para todos os currículos, essa abordagem ainda se matem viável, pois quando se analisa o volume diário de atualizações que a plataforma possa receber, logo se percebe que a geração dinâmica direto da fonte poderia ainda assim consumir bastante recursos da plataforma, fatalmente se tornando um gargalo crítico para o sistema.

4.2 Selecionar Dados

A partir do momento que se obtém a base de dados a ser analisada, a base Registro de Localidade, é que se torna possível iniciar as etapas do KDD. Então, nesta etapa será detalhada a investigação dos dados do Lattes para auxiliar na seleção de alguma amostra e de variáveis na base. Por fim, será detalhada a criação da base Extrato de Mobilidade.

4.2.1 Investigar os Dados

Como forma de maximizar o êxito nas interpretações e hipóteses envolvendo as localidades extraídas, é inevitável que se realize uma análise descritiva para se obter uma melhor compreensão de suas características. Então, por meio da cópia local dos currículos extraídos no final de 2014, foi possível realizar algumas análises descritivas e estatísticas dos dados.

Nessa cópia, o número de currículos foi contabilizado com 3.911.585. Comparado com os valores da data de publicação deste trabalho, um pouco mais de 4 milhões de currículos, percebe-se houve um aumento no número de currículos, entretanto, ainda é possível determinar uma aproximação das interpretações dos dados coletados com o cenário da publicação. Além disso, do total extraído, aproximadamente 99% dos currículos são de brasileiros.

Do total das identificações coletadas, 105 foram desconsideradas devido à ausência de XML, ou por se tratar de pessoas com registro de falecimento. Em 3.202 XMLs haviam a presença de um conteúdo inconsistente, apresentado sempre o tamanho peculiar de 57 bytes. Uma parte destes currículos (2.938) na realidade se tratavam de duplicidades de currículo, ou seja, acontecia devido a uma inconsistência gerada em 1.469 pesquisadores na plataforma que possuía dois identificadores. A princípio, todos esses 3.307 currículos foram considerados inválidos e descartados das análises.

Sobre a última data de atualização, percebe-se que 37% dos currículos foram atualizados

em 2014. Já alguns currículos, aparentemente, não são atualizados desde sua criação, conforme a Figura 4.4. Os meses com menos atualizações são os meses de férias (dezembro, janeiro, fevereiro e junho), diferente de setembro e outubro com mais atualizações. Já os primeiros dias dos meses possuem poucas atualizações, até se chegar aos últimos dias onde existe um gradativo aumento no número de atualização.

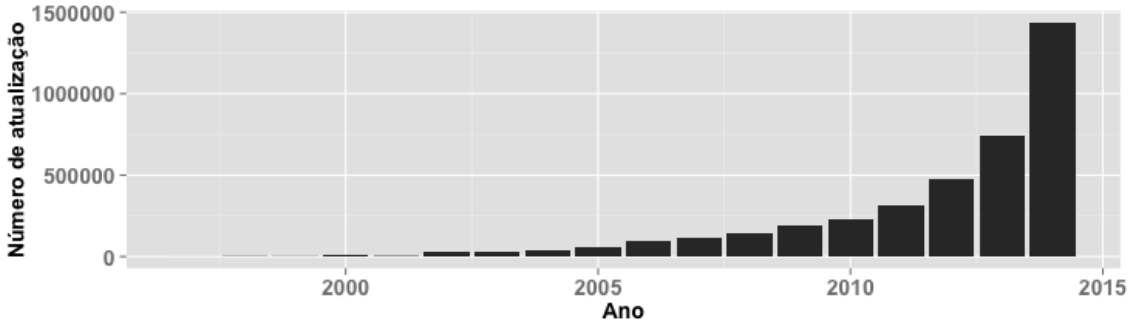
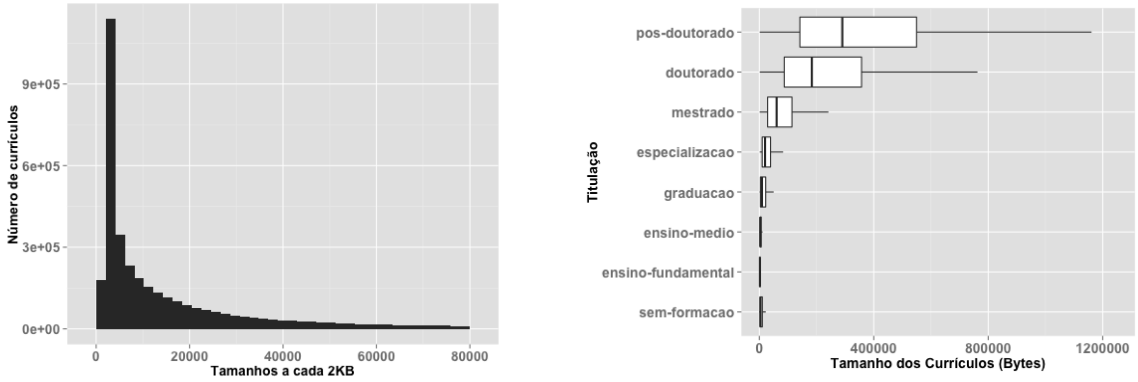


Figura 4.4: Frequência de atualização de currículos por ano.

O tamanho totalizado da base foi de 149 GB, com média de 2 MB por currículo, já no banco esses arquivos são comprimidos no momento do armazenado reduzindo esse tamanho para 32 GB, podendo chegar a 16 GB quando exportado. Em parte, essa eficiência de compressão se justifica pela estrutura do XML, que é bastante redundante. Entre os dez maiores tamanhos de currículos, existem 6 físicos que ocupam as 4 primeiras posições, 2 médicos e 2 engenheiros. O físico experimental detentor do maior tamanho de currículo chega a possuir tanta informação que demora a ser carregado no *browser*[3]. São cerca de 400 artigos completos em periódicos desde 2006 e mais de 20MB de XML.

Já 50% dos currículos apresentam tamanho inferior a 8 KB, conforme mostra a Figura 4.5a. Geralmente, currículos pequenos não possuem tantos dados relevantes, contendo na maioria das vezes apenas a estrutura mínima do XML do Lattes. Por exemplo, para os arquivos abaixo de 1KB, apenas 0,4% informam algum tipo de registro de formação acadêmica. Na medida que o tamanho do currículo cresce, mais dados vão surgindo, inclusive dados relacionados às formações. Até 2KB, cerca de 47% não apresenta formação. Neste aspecto, percebe-se que os currículos pequenos podem ser associados as pessoas no início da formação, conforme mostra a Figura 4.5b, ou que iniciaram o cadastrado na base, mas não preencheram todas as suas informações.

[3]http://lattes.cnpq.br/6961723193391123



(a) Distribuição dos primeiros 75%

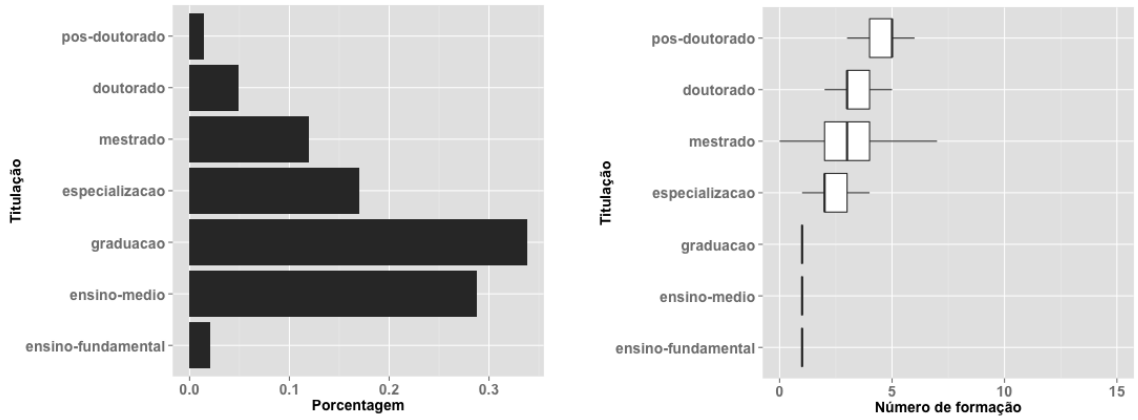
(b) Titulação por tamanho

Figura 4.5: Dados sobre os tamanhos dos currículos.

Quanto à formação, aproximadamente 75% dos currículos possuem até o título de especialização, conforme apresenta a Figura 4.6a. O número de formações aumentam de modo gradativo com a titulação, segundo indica a Figura 4.6b. Ou seja, os pós-doutores são as pessoas que possuem a maior média de registro de formação, e que boa parte das pessoas possuem uma quantidade de dados de formação baixa.

A partir dos dados obtidos, inúmeras outras informações tornaram-se possíveis de serem computadas, inclusive com destaque a alguns fatos curiosos e intrigantes. Por exemplo, com relação aos 126 tipos de bolsa de fomento da CNPq, informadas na base Lattes, apenas 105.774 pessoas são contempladas, cerca de 2,7%. Quanto ao número de formações, pouquíssimas pessoas, preenchem alguns treinamentos como sendo aperfeiçoamento, gerando alguns valores altos no número de formações, como a de um indivíduo que possui mais 160 formações [4]. E a respeito das orientações, 429.233 pesquisadores apresentam algum tipo de orientação, mas nada comparado com a de um pesquisador que possui 2.528 orientações[5], com uma média de 253 orientações por ano. Curiosamente, se considerarmos que em um ano existem em média 253 dias úteis, é como se esse pesquisador orientasse um aluno por dia a 10 anos.

[4]http://lattes.cnpq.br/5772629607957880
[5]http://lattes.cnpq.br/2108232930539953



(a) Porcentagem das titulações

(b) Número de formação por título

Figura 4.6: Dados sobre as formações dos currículos.

4.2.2 Selecionar Amostra e Variáveis

A partir da investigação realizada sobre os dados copiados, iniciou-se o esforço para selecionar uma possível amostra com significância para as análises de mobilidade no MMD. Então, vendo que o número de registros de localidade era relevante para compor o MMP, foi identificado, por meio do padrão do número de formações, que algumas titulações possuíam mais concentração de tais registros. Portanto, optou-se por considerar apenas os currículos de doutores e pós-doutores porque esta amostra de 221.898 currículos concentra mais localidades se comparada com outras possíveis amostras, possuindo em média 6,3 registros por currículo na base Registros de Localidade, sendo que 67% dessa amostra apresenta algum tipo de RT e 89% possuíam RN declarados.

Esta seleção acabou considerando a priori os RN, RF e RT, descartando outras possíveis possibilidades de registro, como forma de simplificar o modelo de mobilidade. Por exemplo, foram descartadas as instituições nas quais o pesquisador realizou orientação de trabalho ou locais de eventos com publicação, mas nada impede que essas outras localidades sejam relacionadas ao MMD no futuro.

Com o uso dessa amostra foi possível extrair aproximadamente 1,3 milhão de localidades na base Registros de Localidade. No entanto, mediante ao êxito desse estudo de caso para analisar o MMD usando esta amostra, é que seria possível ampliar as análises para toda a base.

Para composição e análise do MMP também foi necessário extrair algumas variáveis

que envolvessem: a localização do registro, alguma instituição ou local envolvido, o tipo de registro, a duração do registro e a identificação da pessoa que possui o registro. Para isso, foi estipulado o uso dos valores contidos nas tuplas armazenadas na base Registro de Localidades. Então, as informações referente à $cidade$, $estado$ e $pais$ ajudariam a compor a localização; o nome do $local$ informaria o envolvimento de alguma instituição; os valores do $anoInicio$ e $anoFim$ servem para expressar a duração, o $id16$ ajuda a atribuir uma relação do registro ao pesquisador; e o $tipoDeRegistro$ permite identificar qual tipo de registro ele é, um RN, RF ou RT.

Além dessas foi necessário extrair algumas variáveis dependentes. A partir do momento que se coleta $cidade$, $estado$ e $pais$, ou mesmo o $local$, será possível determinar a efetiva posição geográfica com latitude e longitude. Com o $anoInicio$ e o $anoFim$ é possível calcular algumas métricas que dependem do tempo como o Número de Instâncias, NI, e Intervalo de Duração, ID, que ajudam a ordenar os registros cronologicamente para montar o MMP. Uma vez montando o MMP, também se calcula a métrica da Centralidade de Grau, CG, de algum nós e, caso se isole o fluxo de um pesquisador, é possível calcular o Número de Deslocamentos, ND, e a Distância do Deslocamento, DD.

Quanto ao escopo dessas variáveis, o mesmo será determinado pelo tipo de análise a ser realizada. Algumas situações exigiram um escopo global ou nacional dos registros, já outras casos exigiram algum tempo ou tipo de registro específico.

Por questões de otimização, as variáveis extraídas das tuplas da base Registro de Localidades foram transformadas em algumas entidades no BDR, além de persistirem na base Extrato da Mobilidade. Todo esse processo de transformação garantiu bons resultados, principalmente pelo fato de que muitos valores que estavam sendo armazenados com redundância passaram a se armazenados de modo mais consistente e sem duplicidades no BDR.

4.3 Pré-processar as Variáveis

Essa etapa tem como objetivo iniciar uma preparação dos dados para garantir maior integridade nas interpretações do MMD. Para isso, foi necessário realizar a remoção de valores extremos e inconsistentes, além de tratar os dados ocultos.

A exigência de remover inconsistência se deve ao fato de que o preenchimento de dados

do Lattes ocorre por via manual, o que fatalmente está fadado a erros de digitação. Então, ao preencher uma formação, uma pessoa sem querer poderia informar que seu registro começou em 1880 e não em 1980. Apesar desses fatos ocorrem em pequena proporção, tais valores devem ser tratados pois, na base obtida, é inconsistente considerar que o ano possua valores abaixo de 1900 e acima de 2014. Para essa situação, o registro foi considerado como ruído e desconsiderado na amostra, uma vez que o número de registros excluídos não gerou uma redução impactante na mesma. Outra situação de similar proporção ocorria nos casos em que se informava uma cidade ou instituição que não existia no estado ou país declarado, o que também exigia tratamento.

Quanto à remoção de valores extremos, tornou-se necessário para que os mesmos não afetassem as conclusões, com deturpações nas tendências das métricas. Por exemplo, o ND é baseado na quantidade de registros de localidades que um pesquisador possui. Portanto, no caso citado acima onde o pesquisador possui mais de 160 formações, esse valor representava um ponto fora do padrão dos demais pesquisadores, ou seja, um *outlier*. Assim, para evitar tal situação, fez-se necessário analisar a distribuição dos valores de uma variável para remover os valores extremos. No geral, a aplicação de tal pré-processamento também não implicou em reduções preocupantes na amostra.

O tratamento de dados ocultos também foi aplicado pois, em alguns casos, os registros possuíam uma ocultação parcial ou quase total de seus dados. Quando era total, não existia bem o que fazer a não ser desconsiderá-lo. Nos casos de parcialidade, em algumas situações específicas, havia a chance de recuperar os dados ocultos. A título de exemplo, pode-se citar que alguns RF não declaravam a localização, porém informavam a instituição ao qual pertence. Nessas situações, como a partir da instituição existe a possibilidade de inferir seu posicionamento geográfico, logo foi viável utilizar tal registro na amostra, mesmo sendo uma informação incompleta. Contudo, quando o RF não declarava a instituição e nem a localização envolvida, era inviável definir seu posicionamento geográfico e, por consequência, era descartado.

4.4 Transformar as Variáveis

Com a realização da etapa de pré-processamento, a etapa de transformar as variáveis foi idealizada para continuar o KDD, no intuito de normalizar os dados existentes fazendo alguma mudança em seu formato. Essa etapa essencial para determinar a efetiva localização dos registros. Dessa forma, foi necessário transformar os nomes das cidade e instituição em coordenadas geográficas, pois os nomes não determinam o posicionamento geográfico efetivo necessário para o MMP. Após validar os valores coletados para a mobilidade, foram armazenados todos os dados na estrutura do MMP para se iniciar a etapa do processamento.

Como foi visto, os dados de localidade que são gerados pelo pesquisador deviam possuir uma única instância para todos os cadastrados, como no nome de instituição, que é a priori oferecido por meio de uma listagem padrão na plataforma Lattes. Mas como o sistema pode não abranger todos as instituições, logo se encontra a opção de permitir a inserção manual de instituições ou cidades que ainda não é listado pelo Lattes. Ou seja, nem sempre a mesma instituição é informada de modo normalizado, gerando possíveis duplicidades no seu formato entre os registro de localidade.

É interessante notar que nesta amostra de pessoas, que supostamente tem mais experiência com o Lattes, a declaração do RN como sendo a cidade de São Paulo possui aproximadamente 120 combinações entre nome da cidade, estado e país. Este fato é motivado em grande parte por erros de grafia e combinação inválidas, como declarar que a cidade é de outro país ou estado. Quando a cidade não é informada, a instituição se torna a única alternativa para inferir a efetiva localidade; contudo, o mesmo problema de normalização ocorre com o nome das instituições, como no caso da USP que há aproximadamente 500 combinações encontradas, principalmente devido à combinação do nome da instituição com o nome do curso ou órgão que envolve o RF. Mesmo no caso em que se informam as instituições, deve-se analisar a validade da associação com a localidade, pois pode existir erros de declaração que informam que ela está em um lugar que de fato não está.

Também é importante ressaltar que devido as análises na estrutura das páginas e fluxos HTTP do portal da CAPES e CNPq, foi identificado que as instituições e os cursos apresentavam identificações semelhantes em alguns de seus sistemas. Mais especificamente, isso ocorre nos sistemas da Plataforma Lattes, na Lista de conceitos dos cursos de pós-graduação

e na listagem dos cursos do Sucupira. Isto supostamente indica que existe uma base estruturada de localidades das instituições e cursos, compartilhada entre as duas entidades. Mas como essa base de localidades é fechada e não foi encontrada, a priori, uma alternativa foi a de inferir as localidades com o máximo possível de precisão por meio do nome da instituição ou da cidade, estado e país.

Para resolver a inferência de localidades, primeiro, foi idealizado realizar a normalização desses registros, removendo as duplicidades mais evidentes que estão relacionadas com: a variação das fontes em maiúsculo e minúsculo; excesso de espeço; e utilização de caracteres especiais, como as entidades do HTML, acentuações e pontuação. Tal procedimento de transformação permitiu uma boa redução de duplicatas de uma mesma instância de localidade.

Depois, o próximo passo foi pegar as localidades normalizadas e compará-las com as bases de geolocalização, para assim inferir as suas localizações por meio da obtenção dos pares de coordenada em latitude e longitude. Entre as bases que o MMD utilizou para a inferência dos dois tipos de registro de localidade, cidade e instituição, estão as bases normalizadas de geolocalização do Google Maps e Geonames (ROONGPIBOONSOPIT; KARIMI, 2010); as bases ontológicas com geolocalização do Freebase e DBPedia (DING et al., 2009; JAIN et al., 2010); e as bases de instituições extraídas do portal do Sucupira e dos microdados do INEP conforme a Figura 4.7.

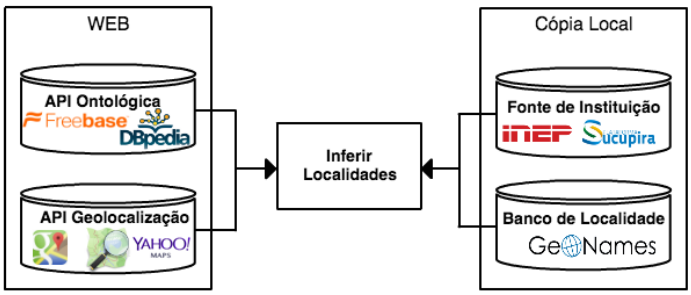


Figura 4.7: Bases utilizadas na inferência de localidades.

A primeira base utilizada para ser comparada com os registros seria a base normalizada de geolocalização usando a API Web do Google Maps, que possibilita definir a latitude e longitude através dos nomes das cidades ou instituições (ROONGPIBOONSOPIT; KARIMI, 2010). Porém, a limitação de requisição diária imposta por esses tipos de API ocasionava um baixo grau de inferência de localidades e um alto tempo de resposta. Além disso, havia

a possibilidade de falhas nas consultas. Por exemplo, se fosse requisitada a localização da universidade X poderia não haver resultado disponível, ou se informava a universidade Y e até algum lugar Z, que não fosse uma universidade, mas que possui nome semelhante a X.

Então, mediante a necessidade de validação dos resultados, a princípio, poderia se utilizar outras bases de geolocalização para reforçar a exatidão das respostas. Contudo, como o problema de comparação dos nomes informados também se tratava de uma questão semântica, foi necessário utilizar as bases ontológicas com geolocalização, como forma de aumentar o sucesso da localização das instituições informadas (DING et al., 2009; JAIN et al., 2010). Logo, a ontologia propiciava a busca de localidades por um tipo específico de lugar, como por exemplo, as universidades, porém, isso ainda não resolvia o problema do tempo das respostas, pois, as limitações de acesso diário também era recorrente nessas bases.

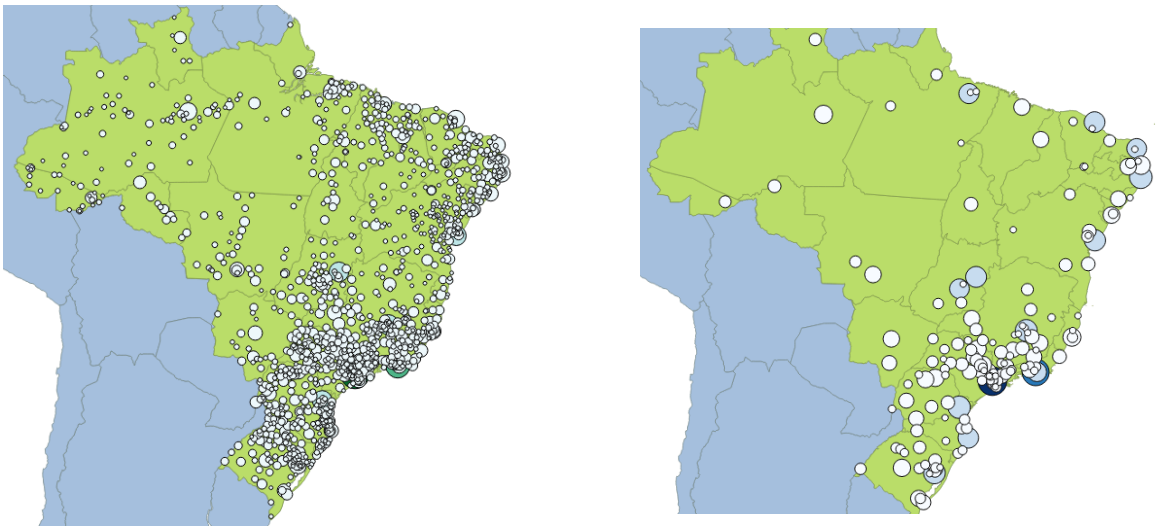
Como forma de resolver o problema do tempo de resposta, foi idealizado procurar bases alternativas que não fossem acessíveis apenas via Web, mas que permitissem acesso aos seus dados via cópias locais ao MMD. Essa solução também impactou em uma eficiência da inferência, pois decidiu-se primeiro processar os nomes de cidades e instituições localmente através das cópias e, no caso de insucesso, se usaria as fontes remotas de localização citadas acima.

Entre as bases pesquisadas que ofereciam cópias, a de localização do GeoNames mostrou-se ser uma solução compacta, viável e ágil. Já quanto a base de instituições, foi necessário fazer algumas análises. Primeiro, houve a suposição de que a maioria das instituições seriam localizadas no Brasil, pelo fato de que os pesquisadores cadastrados no Lattes são predominantemente brasileiros. Desta maneira, seria interessante selecionar uma base de instituições do Brasil, até mesmo porque expandir o mesmo procedimento para todos os países não aumentaria tanto o sucesso da inferência de localidades, principalmente, devido à complexidade de selecionar bases para todos os países. Pensando nisso, foi selecionada a base do censo da educação superior disponibilizada pelo INEP, juntamente com o cadastro dos programas de curso do Sucupira[6], para compor a base de localidades das instituições localmente.

A Figura 4.8 exibe o resultado da extração das instituições combinado com o processamento das localidades usando o GeoNames, que permitiu exibir a concentração de cursos de

[6]https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/dadosCadastrais/dadosCadastraisPublico.jsf

pós-graduação e graduação nas cidades brasileiras.



(a) Graduação

(b) Pós-graduação

Figura 4.8: Concentração dos cursos brasileiros no contexto da cidade.

Outras bases poderiam enriquecer a lista das instituições do Brasil, no entanto, isso demandaria mais tempo e processamento, o que poderia não garantir um aumento no rendimento da inferência de localidade, uma vez que a maioria das instituições de ensino superior são contempladas pelas localidades coletados no Sucupira e INEP. Por uma questão de escopo, também foi considerado que apenas informações a partir da graduação em diante seriam coletadas, fato que permitiu a ausência de alguns RF nas modalidades do ensino fundamental, ensino média e de cursos técnicos, para a geração do MMP. Isso porque o perfil do pesquisador geralmente começa a se definir principalmente no ingresso do ensino superior.

Com as bases locais foram coletadas coordenadas de 27.729 cidades e 2.374 instituições para serem comparadas com os registros da base Extrato de Mobilidade. Na ausência de associação local é que se consulta as bases remotas, de geolocalização e ontológica. Mesmo assim, pode ocorrer falta de associação devido a erros de digitações graves, que comprometem totalmente sua inferência.

Tal fluxo de inferência poderia ser aplicado a cada registro de forma individual, mas graças à proposição de Winkler (1995), que os nomes de alta frequência tendem a ser mais próximos da versão normalizada de uma localidade declarada de forma manual, foi idealizado utilizar apenas os nomes mais frequentes na inferência do MMD.

Mediante a análise da distribuição de frequência dos nomes das instituições, percebeu-se

que a sua distribuição apresentava um padrão no qual 80% dos registros se concentravam nos primeiros 1.000 nomes de maior frequência. Portanto, diante dessa circunstância, inicialmente foi decidido inferir no MMD com apenas 1% dos nomes de maior frequência. O mesmo padrão se repetiu para os nomes de cidades, exigindo a mesma estratégia. Como resultado, tal técnica foi utilizada para otimizar a convergência das coordenadas, através da qual 93% dos registros da base Extrato de Localidades obtiveram suas coordenadas, e foram usados na análise de mobilidade do presente estudo.

Finalmente, com a obtenção, seleção, pré-processamento e tratamento dos dados foi possível montar a base do MMP, que consiste no fluxo dos registros de cada pesquisador, em ordem cronológica de cada acontecimento, agrupados em um multigrafo.

4.5 Processamento e Interpretação dos dados

Diante da base MMP é possível acessar de modo ágil toda a mobilidade da amostra, de tal forma que a base que possuía dados no volume de GB passa ao tamanho de quase 10 MB em formato comprimido. Logo, a ideia nesta etapa consiste em realizar o processamento sobre os fluxos da amostra, no intuito de gerar informações importantes que não são aparentes na plataforma Lattes, devido ao formato como se apresentam os dados, ao seu grande volume e a estratégia de publicação com bloqueios via CAPTCHA.

O processamento gerou alguns artefatos que incluem alguns gráficos, como MFC, GM e MC, em função dos 797.486 deslocamentos entre quase 2.000 instituições distintas existentes nesta amostra. Também foi gerada uma análise estatística de todo o fluxo de deslocamento para se entender quais são os locais e instituições de destaque na mobilidade dos pesquisadores, além das métricas que permitem gerar os gráficos que ajudam nas interpretações dos dados. Coube também a esta etapa, identificar padrões relevantes na mobilidade nos vários estágios de formação da amostra, que são discutidos e analisados na última etapa do MMD e serão apresentados detalhadamente no próximo capítulo. Por fim, existe a possibilidade do MMD tornar os dados e artefatos gerados acessíveis futuramente em uma aplicação Web pública.

4.6 Considerações Finais

Neste capítulo foram apresentadas as etapas do MMD para gerar a base do MMP. A partir da utilização do método foram obtidos os dados sobre mobilidade de uma amostra de pesquisadores, na qual se ajustaram os dados para extrair um conjunto de variáveis e a estrutura de dados que representa a mobilidade.

Diante disso, o próximo capítulo será dedicado a avaliar as variáveis extraídas, juntamente com as análises estatísticas. Além disso, também irá gerar as métricas e visualizações gráficas sobre os deslocamentos, que viabilizará as suas devidas interpretações.

Capítulo 5

Resultados e Discussões

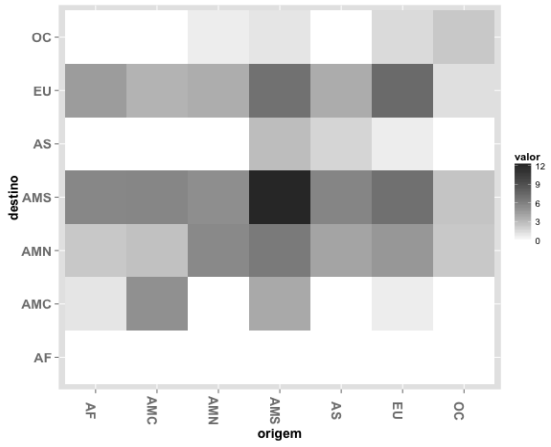
No capítulo anterior foram apresentadas as etapas do Modelo de Mineração de Dados (MMD), para geração da mobilidade dos pesquisadores armazenadas na base Multigrafo de Mobilidade dos Pesquisadores (MMP). Dessa forma, este capítulo descreve o processamento e as análises realizadas nos deslocamentos, em todo o ciclo de formação da amostra selecionada, para exibir alguns dos padrões e dos resultados mais relevantes gerados pelo MMD.

Sendo assim, a Seção 5.1 apresenta os fluxos que acontecem a partir do nascimento até a primeira formação entre os pesquisadores, apontando quais instituições e locais se destacam nesses deslocamentos. Na Seção 5.2 é realizada a comparação dos fluxos entre as formações da amostra, começando da primeira titulação até a última, incluindo também as análises sobre os padrões desses deslocamentos, destacando a influência do poder público na evolução da comunidade científica, o que permitiu a ampliação e a distribuição da formação dos doutores no contexto nacional. Já a Seção 5.3 exibe as informações sobre os fluxos a partir da conclusão da formação acadêmica para algum local de atuação profissional, também informando quais são os locais que se destacam nesses deslocamentos. Por fim, a Seção 5.4 discute a relação das métricas extraídas do grafo de mobilidade com alguns conceitos existentes, como a internacionalização e a associação com conceitos de qualidade das instituições.

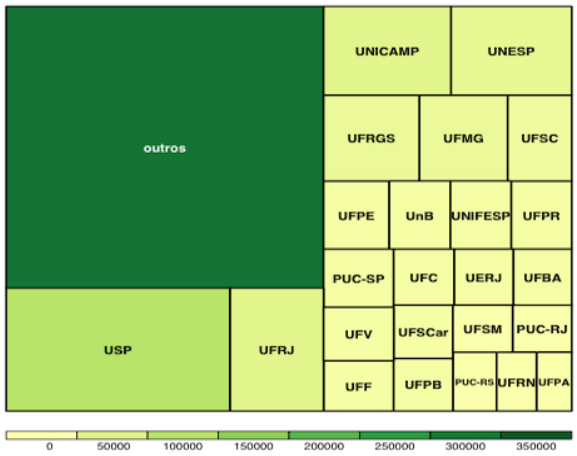
5.1 Origem dos Doutores

Para analisar a origem da amostra coletada, foi preciso inicialmente filtrar os deslocamentos baseado nos Fluxo de Nascimento para a primeira Formação, os FNF, ocupando 24% do total. Diante dessa operação, foi possível perceber que a origem dos doutores é 95% de brasileiros, o que se aproxima da proporção global dos currículos na plataforma. Quanto à localização, a maioria distribui-se entre partes das regiões sul, sudeste e nordeste do Brasil, ao passo que no exterior, a maioria das origens advém da Europa e América do Norte, conforme apresenta a Figura 5.1a.

Também foi possível identificar que 40% das primeiras formações estão presentes nas cidades de origem do doutor, onde as cidades de São Paulo, Rio de Janeiro, Porto Alegre e Belo Horizonte sempre ocupam o posto entre as que mais ofereceram doutores e os retêm no início de carreira. Já a USP, UFRJ, UNICAMP, UNESP e UFRGS são as universidades que mais atraem os doutores no início de formação em termos absolutos ao longo dos anos, acumulando cerca de 24% dos doutores conforme ilustra a Figura 5.1b.



(a) MC no contexto do continente



(b) Proporção da escolha da primeira instituição na formação de doutores.

Figura 5.1: Informações sobre o FNF.

Coincidentemente, estas instituições estão localizadas em regiões de grande concentração habitacional em relação as demais, possuindo grande relevância no contexto da pesquisa nacional. Entretanto, caso se analise a origem das pessoas atraídas por essas instituições, percebe-se que nem todas elas vieram da cidade da instituição. Por exemplo, na USP 47% dos doutores que iniciaram sua formação lá vieram da cidade de São Paulo e 83% do estado

de São Paulo. Por outro lado, quando se analisa o número de pessoas que saem da cidade de São Paulo para iniciar sua trajetória de formação fora da cidade, encontra-se um valor de 40%.

5.2 Formação de Doutores

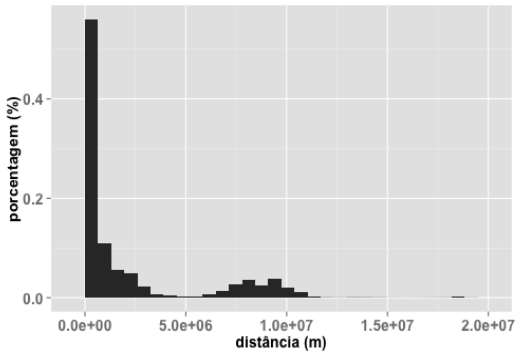
De maneira similar à análise da origem, para se analisar a formação dos doutores foi necessário se coletar apenas os Fluxos de Formação para outra Formação, os FFF, que ocupam 57% do total, ou seja, a maioria dos fluxos. Diferente do que se pode aparentar na Figura 2.10b, do Grafo de Mobilidade no escopo da cidade, o FFF estrangeiro totaliza apenas 9% e apresenta uma centralização no Brasil, com maior Centralidade de Grau de Saída, CG_S, em relação à Centralidade de Grau de Entrada, CG_E. Já os FFF nacionais, que afetam a maioria dos registros de localidades, possuem uma característica em comum, independente do contexto da região, estado, ou cidade. Geralmente, os fluxos concêntricos com Distância do deslocamento, DD, nulo se destacam mais do que os fluxos externos, com DD maior que zero.

Por exemplo, se compararmos a mobilidade entre as regiões brasileiras na Figura 2.11a, verifica-se que o deslocamento na própria região, representada pela coluna secundária, é maior do que entre as regiões distintas. Fato que também ocorre para todos os estados na Figura 2.11b e na maioria das cidades com grande CG, indicando uma possível característica na preferência das pessoas pelos menores percursos.

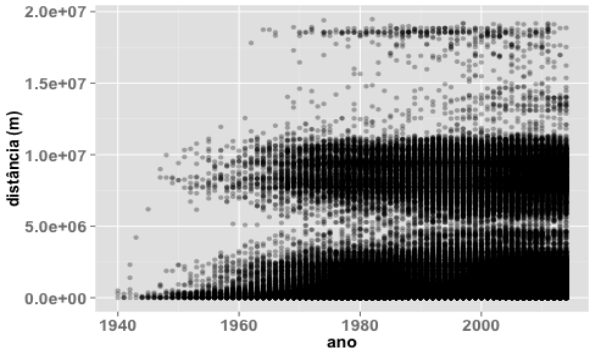
Uma outra forma de reforçar ainda mais essa característica da mobilidade dos pesquisadores no FFF, é que a maioria das pessoas, cerca de 65%, chega a passar por no máximo duas instituições distintas entre as várias formações, com o Número de Localidades, NL, igual a 2. Já 27% tem o NL igual a 1, ou seja, ausência de deslocamento com DD total de 0 Km. Tal característica não é algo exclusivo do FFF, mas também dos demais tipos de fluxo.

No escopo global, boa parte dos FFF são constituídos de pequenos deslocamentos. Na própria Figura 5.2a, 67% das rotas apresentam a DD quase nula, enquanto que 87% não ultrapassam a distância de 1.000 Km. Caso fosse analisada a distribuição da DD na escala temporal, por meio da Figura 5.2b, percebe-se que a partir da década de 70 os deslocamentos mantiveram um padrão uniforme, embora apresentem algumas variações nas intensidades

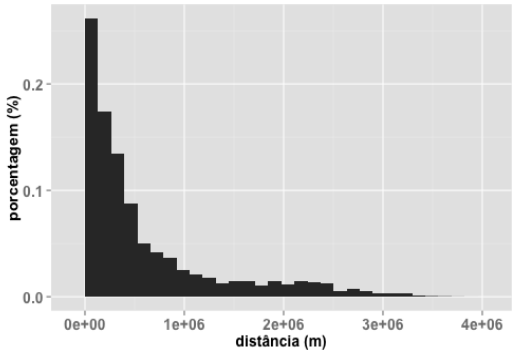
dos fluxos. Quando se extrai apenas as rotas no escopo nacional, obtém-se a frequência das distâncias das rotas na Figura 5.2c, e sua distribuição na Figura 5.2d, permitindo evidenciar que, no Brasil, boa parte dos deslocamentos também apresentam as DD quase nulas.



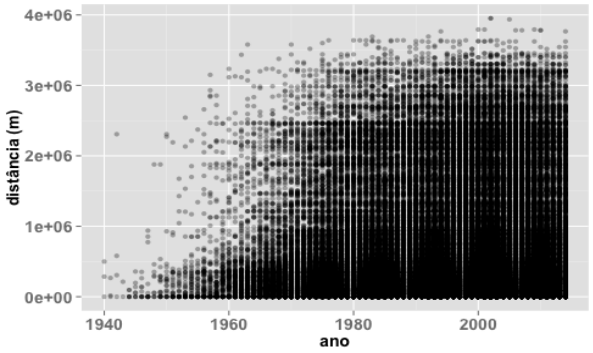
(a) Histograma dos deslocamentos global



(b) Distribuição dos deslocamentos global por ano



(c) Historgrama dos deslocamentos no Brasil



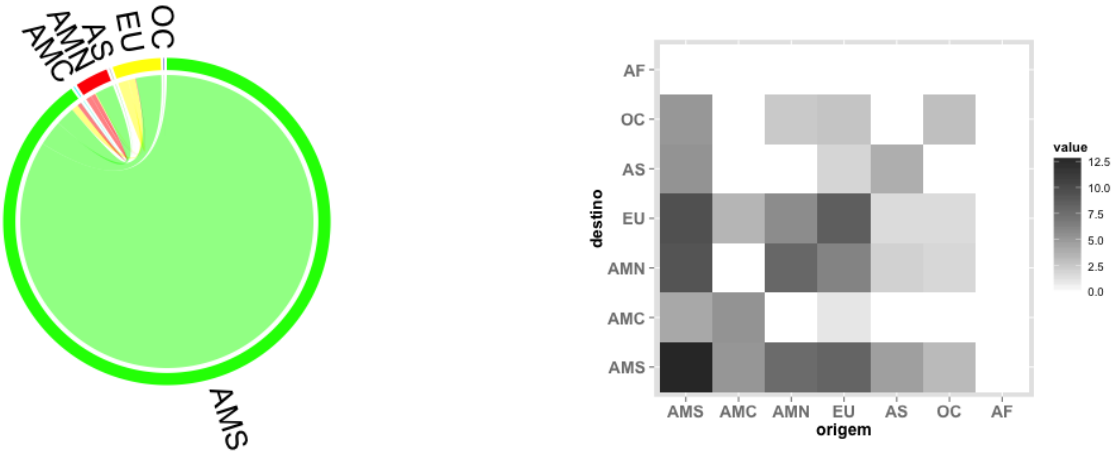
(d) Distribuição dos deslocamentos no Brasil por ano

Figura 5.2: Análise dos deslocamentos no escopo global e nacional.

Analisando a Figura 5.2b também fica evidente que existe dois fluxos intensos, nas partes horizontal inferior e intermediária, e um fluxo menos intenso na parte horizontal superior. Por meio do GM no contexto da cidade, visível na Figura 2.10b, percebe-se que a América do Norte e Europa são destinos intensos no exterior, e que as medidas das DD para essas duas regiões explicam boa parte da densidade de pontos intermediários da Figura 5.2b. A região inferior de baixa DD compreende quase que a totalidade dos fluxos. Esse comportamento se justifica principalmente pelos deslocamentos no Brasil, podendo ser melhor identificado no extrato da Figura 5.2d, com a distribuição apenas dos fluxos brasileiros. Enquanto que o deslocamento superior de alta DD compreende os fluxos entre os hemisférios oeste e leste do globo terrestre, e vice-versa, que são de baixa intensidade e envolvem a Ásia e Oceania.

Uma outa forma bem mais clara de enxergar tal resultado pode ser visto no MFC, na

Figura 5.3a, ou no MC, na Figura 5.3b, em que boa parte dos fluxos entre os continentes[1] está compreendida nas regiões da América do Sul, Europa e América do Norte, nas cores mais intensas do MC e nos fluxos de maior tamanho no MFC. Hipoteticamente, uma possível explicação para a escolha dos dois últimos continentes, para quem vai fazer a formação no exterior, pode ser deduzida pelo fato de que os grandes centros de pesquisa estão localizados nestas regiões.



(a) MFC

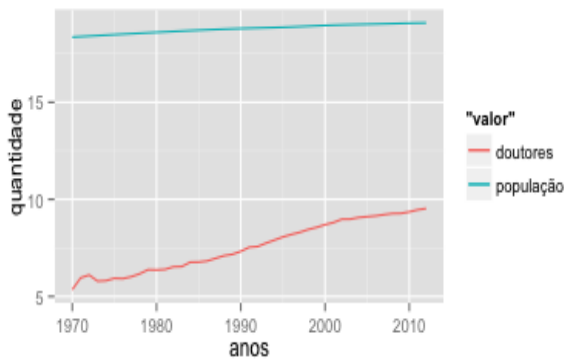
(b) MC (log)

Figura 5.3: Fluxos de formação entre os continentes.

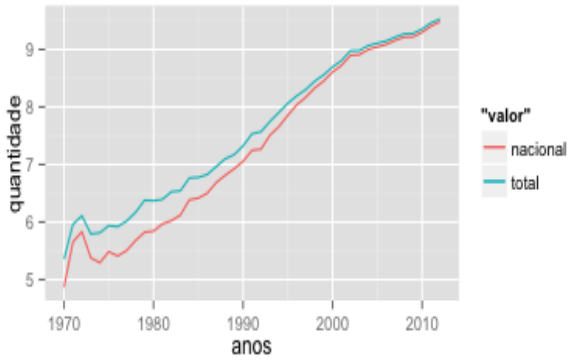
Um fato peculiar é que nenhum registro foi contabilizado na África entre os FFF. Uma possível explicação para essa ausência de deslocamento pode ser dada porque 7% dos registros de localidade não foram usados nesta análise, ou seja, isso não anula a existência de deslocamentos para a África. Pode sugerir, portanto, que possivelmente existam alguns deslocamentos para o continente. Porém, devem possuir valores baixos em relação ao total, devido ao padrão de frequência dos registros de localidade descrito na Seção 4.

No aspecto temporal, destaca-se que o número de formação de doutores vem se caracterizando com uma tendência de crescimento sempre maior do que a escala do crescimento populacional no Brasil, conforme mostra a Figura 5.4a, de tal forma que esse crescimento na maioria das vezes teve a predominância da formação no Brasil, segundo apresenta a Figura 5.4b. Contudo, a meta 14 do PNE define que a linha de crescimento de formação dos doutores deve atingir a marca de 25.000 novos doutores por ano até 2020, o que significa um grande desafio diante da atual margem de crescimento.

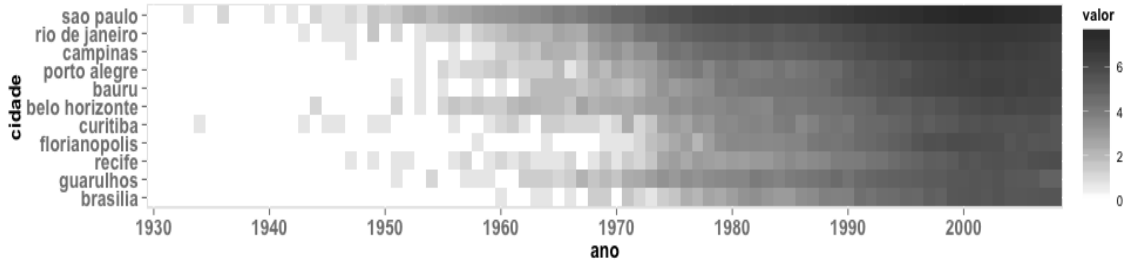
[1]No gráficos da Figura 5.3, os continentes estão indicados por América do Sul(AMS), América do Norte(AMN), América Central(AMC), Europa (EU), Ásia(AS), Oceania(OC) e África(AF)



(a) Formação e crescimento populacional (log).



(b) Formação nacional e total (log).



(c) MC das 11 cidades com maiores formações (log).

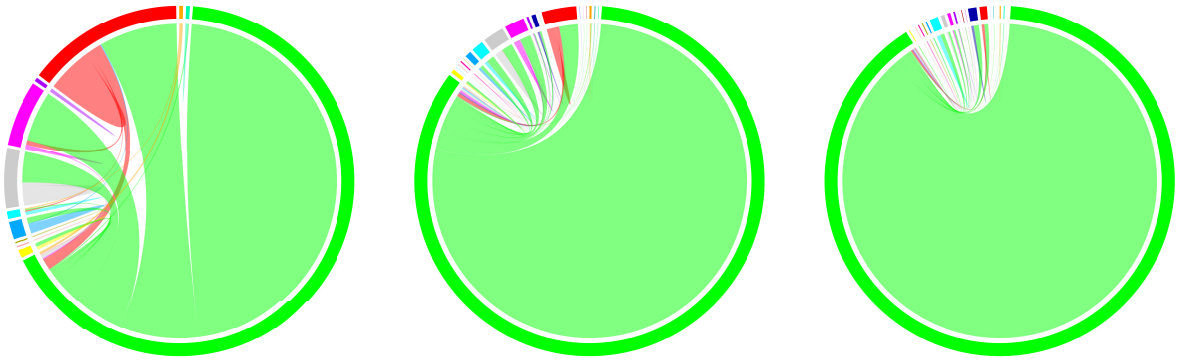
Figura 5.4: Análise temporal da formação de doutores.

Outra forma de quantificar essa evolução pode ser vista na Figura 5.4c através do Número de Instâncias, NI, extraído do MC de doutores formados desde 1930 nas 11 cidades de maior CG. Quanto ao ID dos RF em cada pesquisador, geralmente, é compatível com a duração do curso. Além disso, se considerarmos o Intervalo de Duração, ID, de todos os FFF, existe uma média de duração de 13 anos.

Mas é importante destacar que nos documentos do PNPG fica claro associar alguns fatos históricos com a série de crescimento exposta, como a influência de algumas crises econômicas e políticas da história brasileira na oscilação do ritmo de crescimento da formação de doutores em alguns anos. Também no mesmo plano se encontra outros fatos, ações e metas que atuaram, direta e indiretamente, na evolução do sistema de formação e fomento da CT&I do Brasil (CAPES, 2010).

Através do MMD foi possível exibir de forma discretizada tais fatos, como a da qualificação e reprodução do corpo docente e dos pesquisadores no Brasil. Pela Figura 5.5 é possível perceber que o fluxo de formação dos doutores entre os países na década 80 mostrava que o Brasil, de cor verde, sob a influência do PNPG da época, estava fomentando pesquisadores

a fazerem doutorado em outros países, representados pelas demais cores, como forma de se adquirir cada vez mais uma maior auto suficiência na formação dos doutores no Brasil.



(a) Fluxo de 1983. (b) Fluxo de 1998. (c) Fluxo de 2013.

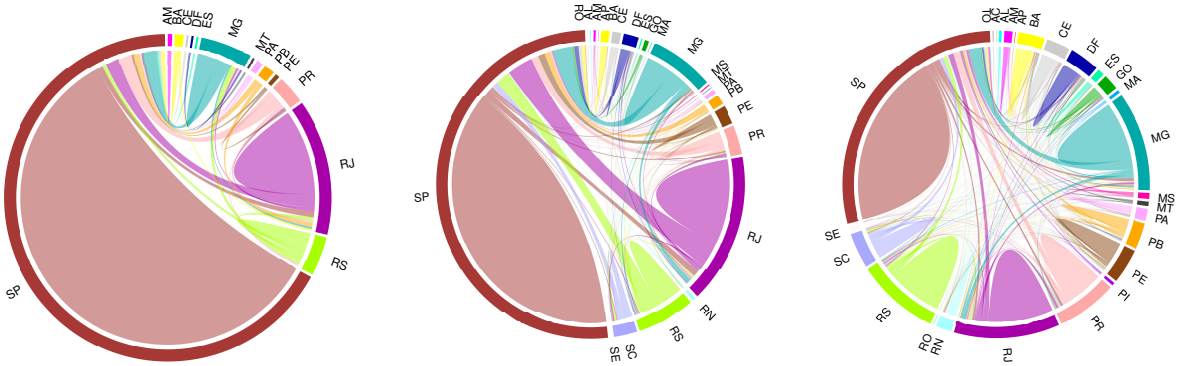
Figura 5.5: Análise temporal da formação de doutores no contexto do país.

Então, por meio da evolução nas Figuras 5.5a, 5.5b e 5.5c, ilustra-se que o Brasil vem adquirindo maior concentração na formação dos doutores, começando com pouco mais da metade dessa formação até ocupar quase que a totalidade. Nesse percurso, identifica-se que três países sempre ocuparam a preferência da formação no exterior, Estados Unidos, Inglaterra e França. Os Estados Unidos, em vermelho, quase sempre ocupava a primeira escolha no exterior entre os pesquisadores, mas hoje divide o favoritismo com Portugal e Espanha que ocupam os primeiros lugares. Contudo, os MFC tendem a induzir que houve uma redução das formações no exterior. Na realidade, o que houve foi um aumento proporcionalmente maior da formação nacional em relação ao exterior, ocorrendo sempre em ambos os casos um crescimento do número de formação ao longo dos anos, conforme ilustra a Figura 5.4b.

Não é o objetivo do PNPG acabar com a formação no exterior. Pelo contrário, segundo as metas do plano, tal formação deve ser incentivada para aumentar o intercâmbio e acesso às tecnologias de ponta dos países desenvolvidos e das instituições de referência mundial (CAPES, 2010). Este fato pode ser comprovado através dos incentivos realizados em programas como o Ciência sem Fronteira, que promovem o fluxo de saída de pesquisadores brasileiros e a entrada de estrangeiros.

Entretanto, o mesmo fluxo de formação descrito na Figura 5.5 quando analisado no contexto dos estados brasileiros, também aponta a visualização de outra meta do PNPG implantada ao longo dos anos, que foi a redução da assimetria regional na formação de doutores visualizada na Figura 5.6. Na sequência dessa figura fica claro identificar que o estado de

São Paulo, em vermelho, ocupava na década de 80 mais de 50% dessa formação, mas que no decorrer dos anos houve uma distribuição na formação dos doutores e a diminuição da responsabilidade dos estados do Sudeste nesta formação. Mesmo assim, todo o Sudeste ainda ocupa em 2013 importante espaço no cenário nacional, sendo responsável por quase 50% da formação dos doutores.



(a) Fluxo de 1983. (b) Fluxo de 1998. (c) Fluxo de 2013.

Figura 5.6: Análise temporal da formação de doutores no contexto do estado.

Quanto ao contexto das instituições, percebe-se que as instituições públicas do sudeste e sul possuem grande impacto na formação dos doutores. Na Figura 5.7 é traçado o fluxo de deslocamento para o doutorado das dez instituição de maior CG, o que ocupa 43% de todo o tráfego dessa formação, e que foi se consolidando ao longo dos anos conforme a Figura 2.12. De todas essas instituições, apenas a PUC-SP não é uma instituição pública e a UFPE é a única que não pertence ao eixo sul e sudeste. Além disso, pode-se perceber que é mais fácil que uma pessoa permaneça na própria instituição para fazer o doutorado do que sair para outra instituição, o que novamente reforça ainda mais a preferência dos deslocamentos curtos, inclusive nessa etapa de ida para o doutorado. Mas também pode-se ressaltar um hábito presente em algumas instituições, que é de favorecer a permanência de seus pesquisadores nos seus processos seletivos de pós-graduação.

5.3 Destino dos Doutores

Os Fluxos de Formação para o Trabalho, os FFT, que representam 19% do fluxo total, mostram que 98% dos RT se localizam no Brasil. A grande maioria dessas ocorrências estão nas regiões sudeste, sul e nordeste. Enquanto que os FFT que envolvem outros países apresen-

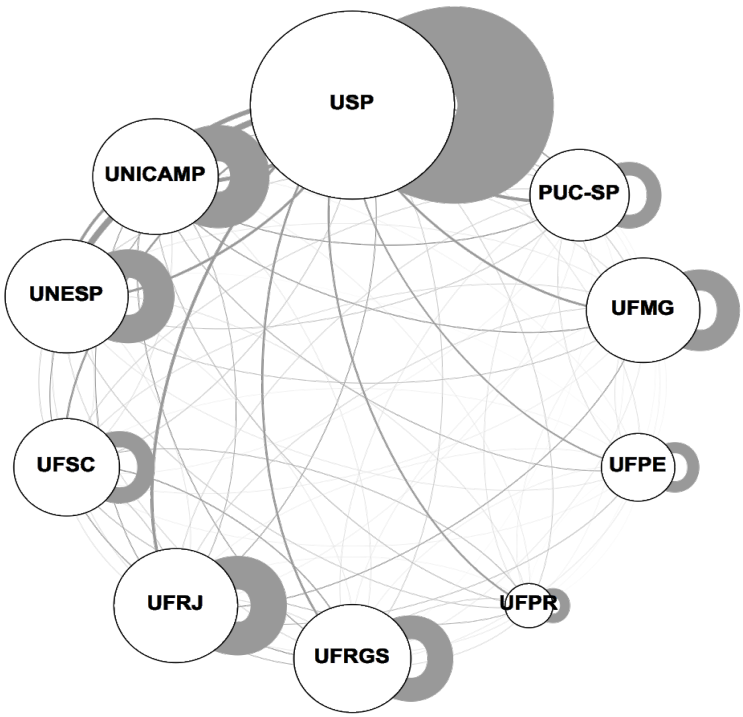


Figura 5.7: Grafo da mobilidade das dez instituições que mais ministram cursos de doutorado.

tam mais uma tendência de entrada para o Brasil, pois boa parte desses fluxos são de pessoas que se formam na América do Norte e Europa e que vêm trabalhar no Brasil.

A USP, UNICAMP, UFRJ, UNESP e UFMG são as universidades que mais fornecem pessoas para atuar profissionalmente, enquanto que as instituições que mais atraem os doutores para trabalhar são a USP, UFRJ, UNESP, UNICAMP e UFRGS. Em termos absolutos representam respectivamente cerca de 20% e 38% do total. Quando se analisa o contexto da cidade, percebe-se que 61% das pessoas saem da cidade de última formação para atuar profissionalmente. Detalhe, no contexto entre as instituições, percebe-se que quase que a totalidade dos fluxos são de instituições que não absorvem seus próprios concluintes com doutorado.

Já quando se considera todos os fluxos de RN, RF e RT, de todos os anos, percebe-se que algumas localidades apresentam algum grau de saída de pesquisadores maior do que se comparado com a entrada. Por exemplo, no contexto dos estados brasileiros, alguns estados do norte apresentam mais saída de pesquisadores, já estados como São Paulo e Rio de Janeiro possuem mais entrada de pesquisadores. No contexto dos principais países com fluxo de

pesquisadores, apresentado na Figura 5.8, os países latino-americanos possuem mais saída, enquanto que países da América do Norte e Europa possuem mais entrada.

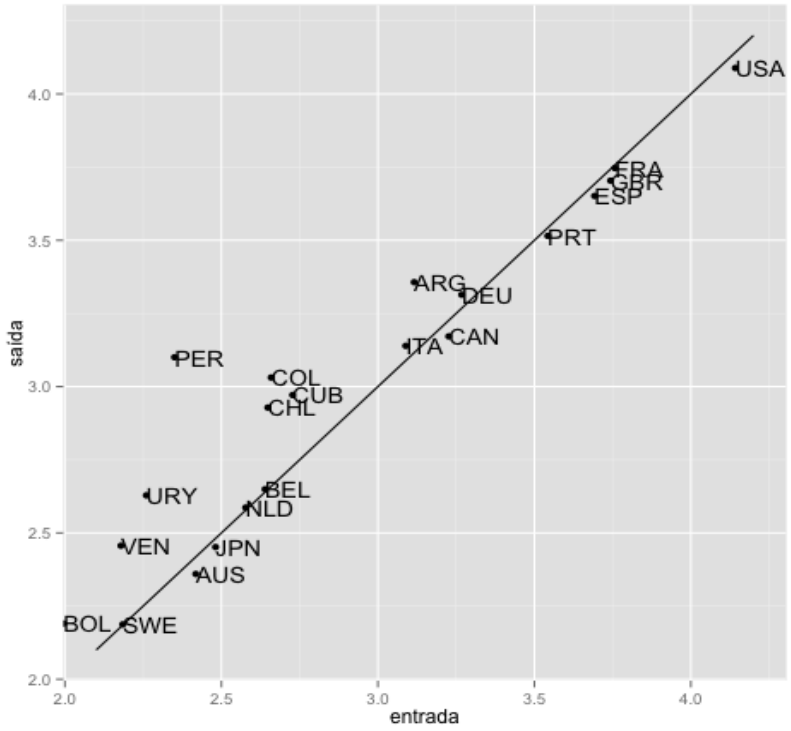


Figura 5.8: Relação de entrada e saída de pesquisadores em alguns países (log).

Considerando o aspecto temporal, se acumularmos todos os fluxos em quatro períodos de tempo, conforme apresentado na Figura 5.9, percebe-se que a mobilidade apresenta sempre uma tendência de crescimento.

5.4 Associação com Deslocamento

Por meio do MMD, foi possível mensurar e identificar alguns padrões na origem, formação e destino dos doutores, como a polarização de algumas instituições e a preferência por caminhos curtos. Contudo, não coube a esta pesquisa aprofundar nas questões de causalidade de tais eventos, como a relação política-econômica nessa polarização, ou no vínculo do baixo deslocamento com a preferência na continuidade das pesquisas já realizadas nas formações iniciais. Contudo, a importância dos artefatos gerados pelo MMD permite potencializar a geração de subsídios para tais pesquisas, inclusive, com a identificação de mais padrões ou associações que possam ser analisados com a mobilidade dos pesquisadores.

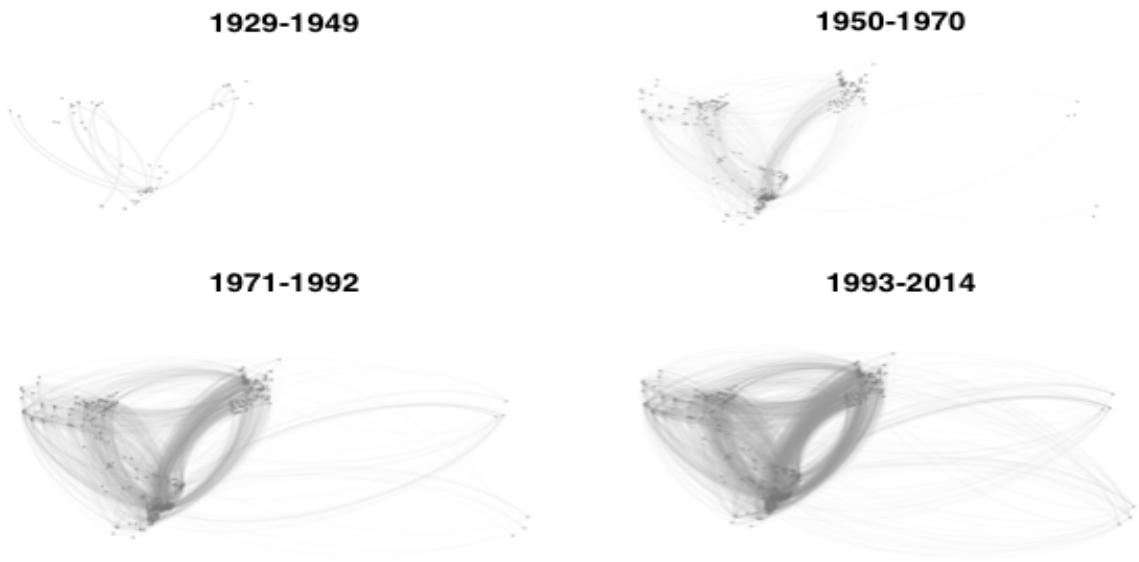


Figura 5.9: Evolução temporal do gráfico de mobilidade agrupada a cada período de tempo.

Então, usando o MMD e seus artefatos de saída seria possível determinar comparações entre os nós usando a ideia do raio de abrangência, calculado pelo GR_S e GR_E. Logo, caso se compare a USP e a UFPE, na Figura 5.10, pelo número de instituições que forneceram pessoas para seus programas de doutorado, percebe-se que a USP atrai mais pessoas de outras instituições do que a UFPE, respectivamente com os raios de abrangência apresentando um GR_E de 295 e 85 instituições.

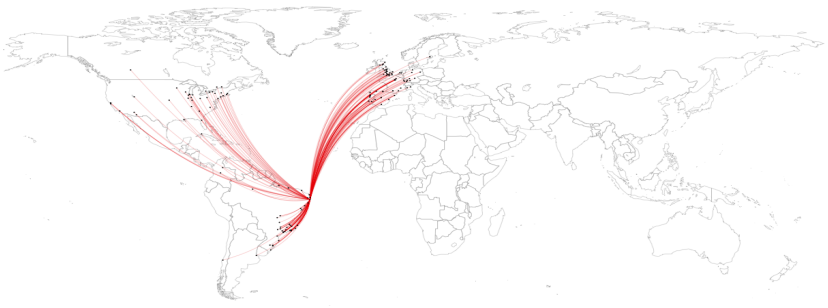
O importante deste número é que ele poderia servir como um possível candidato para identificar o grau de importância de um nó na rede de mobilidade, ou até ser um mecanismo discreto para mensurar conceitos como o grau de internacionalização de uma instituição, que é muito utilizado em programas de pós-graduação no Brasil para determinar seus conceitos.

Outra possibilidade de análise seria indicar associações com outras variáveis. Por exemplo, se compararmos as variáveis extraídas nas métricas de avaliação de instituição usada no Brasil, como o Índice Geral de Cursos (IGC), com a mobilidade de seus envolvidos, percebe-se que existem algumas correlações. Uma das principais seria a relação entre as maiores DD acumuladas dos pesquisadores que atuam em uma instituição e o parâmetro β do IGC[2] possuindo correlação que chega a quase -0,8 na Figura 5.11a.

Entretanto, o IGC não inclui algumas instituições importantes no cenário da mobilidade

[2]Esta métrica é um cálculo da media dos conceitos de mestrandos e o de doutorandos das instituições sob seu quantitativo em termos de graduandos equivalentes.

(a) USP



(b) UFPE

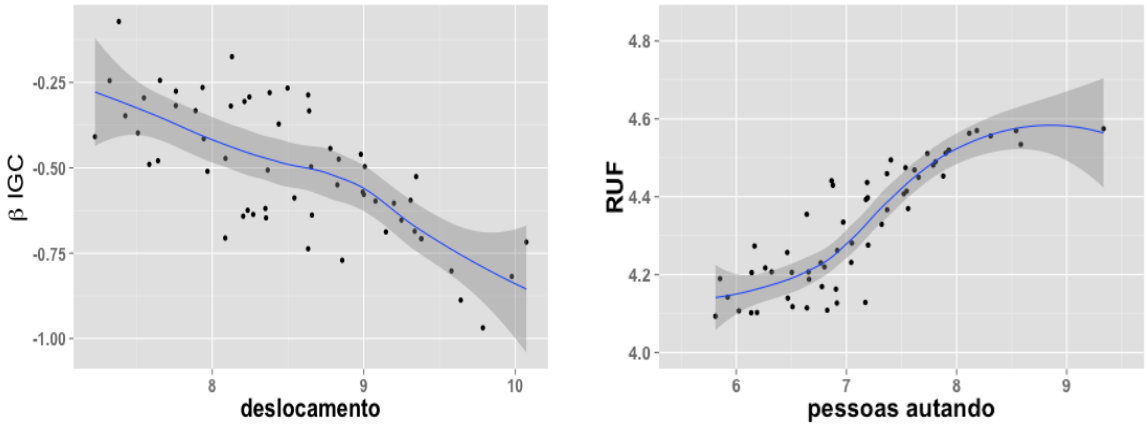
Figura 5.10: Grafo de Mobilidade da origem dos doutores por instituição.

dos doutores, como a USP, devido a ausência na participação do ENADE para avaliar seus alunos de graduação, que é uma das variáveis utilizada no IGC. Então, como a maioria das instituições são nacionais, logo não seria interessante utilizar métricas que incluam tais instituições, como o Webometrics, mas que abrangem poucas instituições nacionais. Por isso, com a utilização da classificação RUF, que focaliza mais as instituições nacionais, é possível extrair outras associações como a relação do NI de pessoas atuantes e as principais notas de tal conceito, que possui uma correlação de quase 0,9 na Figura 5.11b.

5.5 Considerações Finais

O objetivo deste capítulo foi apresentar os resultados obtidos através do processamento dos dados da amostra. Diante disso, foi possível compreender como a amostra se desloca, inclusive exibindo a tendência da amostra em fazer poucas mudanças de instituições na obtenção de seus títulos, além de visualizar como é a atuação destas instituições nesse processo de deslocamentos.

Os inúmeros gráficos, métricas e estatísticas geradas, ajudaram a compreender alguns padrões, como a redução da assimetria regional e independência nacional na formação dos



(a) DD acumulado pelo β IGC

(b) NI de atuação por RUF

Figura 5.11: Associação das métricas geradas pelo MMD.

doutores ao longo dos anos, viabilizando a associação e a quantificação de conceitos que, podem ou não, depender do fluxo de mobilidade dos pesquisadores.

Levando em consideração esses resultados, o próximo capítulo apresenta as conclusões obtidas neste trabalho, bem como as principais contribuições, limitações e perspectivas de trabalhos futuros.

Capítulo 6

Conclusão

Este capítulo apresenta as considerações finais a respeito do trabalho desenvolvido, as contribuições e limitações da pesquisa, bem como as frentes de trabalhos futuros.

Diante do atual cenário fica claro destacar a importância da mineração de dados para ajudar na compreensão da mobilidade. Então, através dos processos de visualização dos dados, muitas informações, que em alguns casos estão sob domínio público, passam a ser evidentes, garantindo maior aplicabilidade e importância para os dados. Portanto, as pesquisas apresentadas exibiram vários tratamentos e aplicações sobre os dados, cujo o propósito inicial não era exibir as informações apresentadas, mas que devido a extração, análise e principalmente a visualização, ajudaram a descrever melhor alguns fenômenos, por exemplo, a Plataforma Lattes tem como propósito o armazenamento dos registros curriculares dos pesquisadores, mas que devido à mineração foi possível exibir a mobilidade das pessoas em vários contextos.

Outro fator crucial é o tempo de acesso a bases com grande volume de informação, presente no Lattes. Pois, o tempo de leitura e geração das informações praticamente se tornam inviáveis na escala global, por causa do alto tempo de resposta gerado pelas inúmeras requisições na Web e devido às barreiras impostas para se acessar os dados, e tal forma que ainda pode se agravar dependendo das condições de acesso à Web.

Portanto, levando em consideração essa necessidade de acesso eficaz a grandes bases, essa pesquisa idealizou o Modelo de Mineração de Dados (MMD), que propôs uma metodologia para copiar os dados curriculares do Lattes aplicando técnicas de mineração. A corrente metodologia de análise de mobilidade através dos registros de localidade, também

permitiu caracterizar algumas questões relevantes. Essa característica expõe alguns aspectos peculiares da formação dos pesquisadores brasileiros desde o escopo local até o nível global.

Na literatura, foi possível constatar que em outros países os pesquisadores chegam a conseguir o acesso ao doutorado de forma mais direta, sem necessariamente possuir o que equivale ao mestrado brasileiro. Também demonstra que em muitas áreas e locais do Brasil o mestrado passa a ser o título máximo, acarretando na complexidade de obtenção do título, pois os esforços de alguns programas de pós-graduação se concentram exclusivamente nessa formação, gerando durações mais prolongadas comparadas a de outros países ou programas.

Acreditava-se também que em um país de grande extensão territorial como Brasil, seria comum que pessoas que nasceram nas regiões mais inóspitas do país tivessem que se deslocar bastante para se atingir uma formação específica. Entretanto, os resultados do MMD apontaram de forma quantitativa e graficamente que existe uma preferência por deslocamentos curtos pela maioria da amostra analisada, mesmo considerando a amostra dos doutores, que possui o maior número médio de formações e, por consequência, maior passagem por instituição para a obtenção do título.

Outro fato importante no contexto nacional é que a maioria dos registros de instituição e origem de pessoas se concentrava nas capitais brasileiras, pois quase 50% vinham destes locais. Porém, em termos proporcionais, o sul e sudeste indiscutivelmente emplacam a maioria da formação de doutores do Brasil, concentrando as principais rotas de deslocamento e instituições.

Segundo o Plano Nacional de Pós-Graduação (PNPG), algumas de nossas instituições possuem conceitos que nos projetam entre importantes centros de pesquisa a nível mundial. Muito embora ainda possuímos um número baixo de pessoas com pós-graduação em comparação com países mais desenvolvidos, fora a grande dependência do setor público para criar e controlar os rumos da pesquisa nacional.

Quanto aos vários contextos analisados, entende-se que no cenário global a África, Ásia e Oceania possuem pouca atuação na mobilidade dos pesquisadores nesse estudo de caso, de tal forma que na África existem apenas registros de pessoas que nasceram lá, mas sem nenhuma formação ou atuação profissional. Já a Europa e América do Norte apresentam tanto pessoas que nasceram em tais locais, como registros de instituições com pessoas que se formaram ou trabalharam lá. Vale destacar que essas instituições estrangeiras nem sem-

pre são as primeiras colocadas em conceitos de qualidade em nível mundial, conforme o Webmetrics, mas ocupam importante destaque no âmbito global.

Alguns aspectos positivos foram encontrados na mobilidade. No fluxo de interação das instituições de formação e atuação, cada vez mais, o eixo sul-sudeste distribui sua atuação com os demais centros. Além disso, o aumento crescente nos investimentos, retratados no PNPG, podem ter sido de extrema valia para o atual ritmo de crescimento na formação dos pesquisadores brasileiros, que é extremamente necessário para preencher as lacunas existentes no Brasil e para aumentar o desenvolvimento da CT&I nacional de um modo geral.

Por fim, o MMD demonstrou seu potencial na associação de seus resultados com conceitos que estão relacionados com a pesquisa. Na prática, foi demonstrada a aplicação da Centralidade de Grau, CG, na identificação da influência de nós na rede e assim gerar o grau de internacionalização; e a Distância do Deslocamento, DD, e do Número de Instâncias, NI, na correlação com alguns parâmetros de qualidade praticados no contexto nacional. Contudo, o potencial da utilização de tais métricas ainda pode ser mais explorado.

6.1 Contribuições

O presente trabalho forneceu várias contribuições importantes. Inicialmente se destaca quanto ao método de extração de dados nas bases de registros públicos na Web, por propor uma abordagem de coleta genérica de dados e acesso independente da fonte num tempo ágil, podendo ser aplicado a diversos tipos de dados e bases. Em outras palavras, o MMD não se restringe apenas a extração de localidades nos currículos do Lattes, mas pode ser configurado para extrair outros dados de outras bases de registros públicos, provendo interfaces de consulta local independente da base original.

Uma das maiores relevâncias desta proposta consistiu em definir um mecanismo de tratamento e inferência de localidades utilizando simples processos de comparação, devido à identificação da distribuição de frequência dos registros de localidade. Graças a essa eficiência na interpretação das localidades é que foi factível o presente estudo sobre a mobilidade dos pesquisadores.

Contudo, mais importante que as possibilidades de análise propostas são as visualizações e métricas geradas neste trabalho, pois foi através delas que foi possível elaborar uma boa

estratégia para descrever e quantificar, de forma simplificada, a mobilidade entre os pesquisadores. Esta estratégia permitiu discretizar e entender melhor como os doutores saíram de suas cidades de origem até encontrarem um local de atuação profissional.

6.2 Desafios e Limitações

Inúmeras foram as limitações e os desafios encontrados nesta pesquisa, iniciando pelo processo de obtenção e preparação dos dados. Isso porque para obter os dados primeiro foi preciso selecionar algumas fontes de dados para entender a estrutura e o acesso aos mesmos, para assim realizar a coleta e prosseguir com o fluxo do MMD. Todos os dados estavam presentes em arquivos textuais, mas na maioria das vezes, disponíveis em arquivos HTML ou XML na Web. Neste cenário o maior desafio foi superar a burocracia da obtenção dos dados públicos do Lattes para acessar os dados curriculares dos pesquisadores cadastrados. Foi necessária uma boa análise sobre os fluxos HTTP e sobre algoritmos de reconhecimento de padrão para ultrapassar os CAPTCHA, além da elaboração de um algoritmo concorrente e tolerante a falhas para copiar todos os currículos. Tudo isso porque a plataforma Lattes ainda não disponibilizou um serviço aberto e sem restrições, que periodicamente poderia distribuir seus dados, como o que acontece na base de currículos japonês KAKEN.

Em seguida, os desafios prosseguiram com o tratamento e formato dos dados, devido aos problemas recorrentes de localização sem coordenada e com inserção manual, acompanhado de possíveis erros de digitação. Vários processos de inferência e comparação de localidades foram analisados, pois o grande número de localizações inviabilizava o fornecimento de coordenadas de modo manual, bem como a identificação de erros de digitação e normalização. Então, graças as análises de frequência e utilização de técnicas de geolocalização, é que foi possível determinar quase que integralmente as localidades.

Diante da validação das coordenadas geradas, percebeu-se que, até o contexto do estado, as localidades determinadas pelo MMD possuíam uma boa definição. Em alguns casos, porém, optou-se por desconsiderar alguns registros da amostra por conta da generalização dos nomes informados para a localidade. Por exemplo, se alguém informava que trabalhava no Departamento de Informática, dificilmente se obtinha sua localização com exatidão. Em outros casos, instituições que ultrapassam o escopo do estado, como a localidade de nome

EMBRAPA, que possui sede em quase todo o território brasileiro, seria muito difícil determinar uma definição genérica no contexto da cidade e estado devido a sua capilaridade. Já para algumas instituições que possuem vários *Campi* e não informava qual seria seu *Campus* efetivo, a alternativa foi considerar sua definição na sede, maximizando seu posicionamento no contexto do estado. Além disso, houveram outros problemas com os registros de localidade, como na formação sanduíche que não detalhava de forma adequada as instituições envolvidas.

A próxima limitação foi no armazenamento e acesso dos dados, porque o arquivo XML pode até ser um formato de distribuição de dados de modo semântico e estruturado, mas esta longe de ser o formado ideal para acesso segmentado de suas informações em grande escala, como nos quase 4 milhões de currículos coletados. Algumas estratégias e arquiteturas de armazenamento foram analisadas, como através da utilização dos bancos NoSQL, BDR, banco de dados de XML e Neo4J, com o enfoque principal na possibilidade de receber um arquivo XML do Lattes e assim poder disponibilizar informações granulares em tempo hábil, como a lista de formações de um pesquisador específico. Superado o problema do XML, o outro desafio consistiu em como armazenar o grafo de mobilidade, pois a solução implantada foi usando um BDR, mas que se mostrou não ser um boa forma de armazenar e consultar conteúdos do grafo de modo *ad hoc*, gerando um importante trabalho futuro na continuidade de tal estudo junto com a análise dos bancos NoSQL, para fornecer tais informações de modo escalável.

Mas o maior desafio consistiu no tempo de resposta de cada ação, pois devido a proporção dos dados e a complexidade em seu tratamento foi necessário analisar e testar inúmeras práticas de extração, armazenamento, leitura e processamento dos dados, para atingir um tempo viável para execução de tal pesquisa. E vale salientar que o tempo é um fator crucial na mineração de dados, exigindo bastante cuidado e atenção no momento de planejar e executar as tarefas, pois um erro fatalmente pode desperdiçar dias de trabalho e ser de difícil diagnóstico. Principalmente por causa das atividades minuciosas envolvidas na transformação dos dados e devido ao grande volume de dados.

6.3 Trabalhos Futuros

Diante dos resultados gerados no MMD, foi possível compreender e visualizar a mobilidade da amostra selecionada. Nesse contexto, outros trabalhos futuros poderiam seguir para explorar e aprofundar a aplicação do modelo usando outras amostras ou até mesmo toda a base de currículos.

Uma das alternativas para a amostragem seria a seleção de bolsistas em áreas distintas, como os bolsistas de produtividade da CAPES disponíveis em seu portal. Até mesmo a seleção de profissionais que atuam em áreas específicas, como os professores que ensinam em programas de pós-graduação, disponíveis no portal do Sucupira. Tal seleção seria interessante para tentar identificar algum padrão na mobilidade da amostra selecionada, ou até mesmo averiguar se existe alguma semelhança na mobilidade dos grupos, ou correlação com seus indicadores de desempenho. Outra possibilidade de trabalho pode consistir em propor o inverso, ou seja, tentar realizar agrupamentos e classificações sobre a mobilidade dos pesquisadores para induzir a criação de tais grupos.

Outra possibilidade de trabalho seria observar se a evolução do número de formação de doutores nas cidades possui associação com seus indicadores socioeconômicos, como o IDH-M e PIB, aprofundando em questões de causa e efeito de tal situação. Ou até mesmo investigar os padrões de mobilidade na escala temporal para indicar tendências.

Também existe a possibilidade de adaptar o MMD para utilizar outras localidades presentes nos registros curriculares, pois por motivos de simplificação do modelo, foram considerados apenas RN, RF e RT, eliminando os demais registros de localidades e até outras variáveis do currículo. Portanto, o MMD poderia considerar mais localidades, como as publicações em eventos, que têm sido bem exploradas pela literatura atualmente com as filiações nos artigos, para ampliar a ideia de mobilidade.

Quanto a cópia dos currículos realizada no final de 2014, pode-se haver uma estratégia de atualização mais frequente e otimizada dos dados da cópia, usando seus atuais valores na plataforma com atualizações incrementais, onde concentrará seu maior esforço na cópia inicial, que já foi realizada, para depois fazer pequenas atualizações. Para otimizar a periodicidade da atualização dos currículos, pode-se levar em conta os registros de atualização de cada currículo para estimar melhores intervalos de atualizações, porque não faz sentido intensificar

a verificação de atualização curricular dos pesquisadores que não costumam atualizar seus dados com frequência, garantindo assim uma melhora no fluxo das atualizações.

Finalmente, ainda como sugestão de trabalho futuro, pretende-se desenvolver uma aplicação para disponibilizar todos os artefatos gerados através de uma cópia da base MMP, ou por meio de uma API Web para facilitar o consumo de tais dados de forma pública.

Bibliografia

ABEL, G. J.; SANDER, N. Quantifying global international migration flows. *Science*, American Association for the Advancement of Science, v. 343, n. 6178, p. 1520–1522, 2014.

AGARWAL, R.; DHAR, V. big data, data science, and analytics: The opportunity and challenge for is research. *Information Systems Research*, INFORMS, v. 25, n. 3, p. 443–448, 2014.

ANDRIS, C. Weighted radial variation for node feature classification. *arXiv preprint arXiv:1102.4873*, 2011.

BALCAN, D. et al. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 106, n. 51, p. 21484–21489, 2009.

BARKER, L. E. et al. Bayesian small area estimates of diabetes incidence by united states county, 2009. *Journal of Data Science*, v. 11, n. 2, p. 249–267, 2013.

BARUFFALDI, S. H.; LANDONI, P. Return mobility and scientific productivity of researchers working abroad: The role of home country linkages. *Research Policy*, Elsevier, v. 41, n. 9, p. 1655–1665, 2012.

BHATTACHARYA, K. et al. The international trade network: weighted network analysis and modelling. *Journal of Statistical Mechanics: Theory and Experiment*, IOP Publishing, v. 2008, n. 02, p. P02002, 2008.

BOLDI, P. et al. Ubicrawler: A scalable fully distributed web crawler. *Software: Practice and Experience*, Wiley Online Library, v. 34, n. 8, p. 711–726, 2004.

BURZAŃSKA, M. et al. Recursive queries using object relational mapping. In: *Future Generation Information Technology*. [S.l.]: Springer, 2010. p. 42–50.

CAÑIBANO, C.; BOZEMAN, B. et al. Curriculum vitae method in science policy and research evaluation: the state-of-the-art. *Research Evaluation*, v. 18, n. 2, p. 86, 2009.

CAÑIBANO, C.; OTAMENDI, F. J.; SOLÍS, F. International temporary mobility of researchers: a cross-discipline study. *Scientometrics*, Akadémiai Kiadó, co-published with Springer Science+ Business Media BV, Formerly Kluwer Academic Publishers BV, v. 89, n. 2, p. 653–675, 2011.

CAÑIBANO, C.; OTAMENDI, J.; ANDÚJAR, I. Measuring and assessing researcher mobility from cv analysis: the case of the ramón y cajal programme in spain. *Research Evaluation*, Oxford University Press, v. 17, n. 1, p. 17–31, 2008.

CAPES. *PLANO NACIONAL DE PÓS-GRADUAÇÃO (PNPG) 2011-2020*. 2010. Disponível em: <https://www.capes.gov.br/images/stories/download/Livros-PNPG-Volume-I-Mont.pdf>. Acesso em: 25 jun.2015.

CAPES. *GeoCapes*. 2015. Disponível em: <http://geocapes.capes.gov.br/geocapesds/>. Acesso em: 25 jun.2015.

CHANG, F. et al. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, ACM, v. 26, n. 2, p. 4, 2008.

CHEN, C.-h.; HÄRDLE, W. K.; UNWIN, A. *Handbook of data visualization*. [S.l.]: Springer Science & Business Media, 2007.

CHRISTEN, P. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. [S.l.]: Springer Science & Business Media, 2012.

CHRISTEN, P.; BELACIC, D. Automated probabilistic address standardisation and verification. In: *Australasian Data Mining Conference (AusDM'05)*. [S.l.: s.n.], 2005. p. 53–67.

CHRISTEN, P.; CHURCHES, T.; WILLMORE, A. A probabilistic geocoding system based on a national address file. In: CITESEER. *Proceedings of the 3rd Australasian Data Mining Conference, Cairns*. [S.l.], 2004.

CHUN, T. Y. World wide web robots: an overview. *Online and CD-Rom Review*, MCB UP Ltd, v. 23, n. 3, p. 135–142, 1999.

CNPQ. *Plataforma Lattes*. 2015. Disponível em: <http://lattes.cnpq.br/>. Acesso em: 25 jun.2015.

CONCHI, S.; MICHELS, C. *Scientific mobility: An analysis of Germany, Austria, France and Great Britain*. [S.l.], 2014.

CORMEN, T. H. *Introduction to algorithms*. [S.l.]: MIT press, 2009.

CRILLY, T. *50 Mathematical Ideas You Really Need to Know*. Book Sales, 2007. ISBN 9781847241474. Disponível em: <http://books.google.com.br/books?id=XfYSKAAACAAJ>.

DEAN, J.; GHEMAWAT, S. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, ACM, v. 51, n. 1, p. 107–113, 2008.

DHAR, V. Data science and prediction. *Communications of the ACM*, ACM, v. 56, n. 12, p. 64–73, 2013.

DIETZ, J. S. et al. Using the curriculum vita to study the career paths of scientists and engineers: An exploratory assessment. *Scientometrics*, Springer, v. 49, n. 3, p. 419–442, 2000.

DIGIAMPIETRI, L. et al. Minerando e caracterizando dados de currıculos lattes. In: *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. [S.l.: s.n.], 2012.

DIGIAMPIETRI, L. et al. Dinâmica das relações de coautoria nos programas de pós-graduação em computação no brasil. In: *2012 Brazilian Workshop on Social Network Analysis and Mining*. [S.l.: s.n.], 2012.

DIGIAMPIETRI, L. A. et al. Análise da rede dos doutores que atuam em computação no brasil. 2014.

DIGIAMPIETRI, L. A. et al. Brax-ray: an x-ray of the brazilian computer science graduate programs. *PloS one*, Public Library of Science, v. 9, n. 4, p. e94541, 2014.

DING, L. et al. An ontology based prototype for geocoding offset addresses. In: IEEE. *Software Engineering, 2009. WCSE'09. WRI World Congress on*. [S.l.], 2009. v. 2, p. 239–243.

DIONISIO, J. D. N.; DAHLQUIST, K. D. Improving the computer science in bioinformatics through open source pedagogy. *ACM SIGCSE Bulletin*, ACM, v. 40, n. 2, p. 115–119, 2008.

DUKOVICH, A. et al. Joxm: Java object xml mapping. In: IEEE. *Web Engineering, 2008. ICWE'08. Eighth International Conference on*. [S.l.], 2008. p. 332–335.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, ACM, v. 39, n. 11, p. 27–34, 1996.

FEDORSCHAK, K. et al. Data analytics and human trafficking. In: *Advancing the Impact of Design Science: Moving from Theory to Practice*. [S.l.]: Springer, 2014. p. 69–84.

FILIPPO, D. D.; CASADO, E. S.; GÓMEZ, I. Quantitative and qualitative approaches to the study of mobility and scientific performance: a case study of a spanish university. *Research Evaluation*, Oxford University Press, v. 18, n. 3, p. 191–200, 2009.

FLORESCU, D.; KOSSMANN, D. A performance evaluation of alternative mapping schemes for storing xml data in a relational database. 1999.

FU, Z.; CHRISTEN, P.; BOOT, M. Automatic cleaning and linking of historical census data using household information. In: IEEE. *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. [S.l.], 2011. p. 413–420.

FURUKAWA, T.; SHIRAKAWA, N.; OKUWADA, K. Quantitative analysis of collaborative and mobility networks. *Scientometrics*, Akadémiai Kiadó, co-published with Springer Science+ Business Media BV, Formerly Kluwer Academic Publishers BV, v. 87, n. 3, p. 451–466, 2011.

GAUGHAN, M. Using the curriculum vitae for policy research: an evaluation of national institutes of health center and training support on career trajectories. *Research evaluation*, Oxford University Press, v. 18, n. 2, p. 117–124, 2009.

GIBSON, J.; MCKENZIE, D. Scientific mobility and knowledge networks in high emigration countries: Evidence from the pacific. *Research Policy*, Elsevier, v. 43, n. 9, p. 1486–1495, 2014.

GIMENEZ, X. *Flujos de emigración en España en la actualidad*. 2015. Disponível em: <http://www.xavigimenez.net/innovadata/>. Acesso em: 25 jun.2015.

GOEKEN, R. et al. New methods of census record linking. *Historical methods*, Taylor & Francis, v. 44, n. 1, p. 7–14, 2011.

GOLDBERG, D. W. A geocoding best practices guide. *Springfield, IL: North American Association of Central Cancer Registries*, 2008.

GOLDBERG, D. W. Improving geocoding match rates with spatially-varying block metrics. *Transactions in GIS*, Wiley Online Library, v. 15, n. 6, p. 829–850, 2011.

GOVAERTS, S.; DUVAL, E. A web-based approach to determine the origin of an artist. In: *Proceedings of ISMIR2009: 10th International Society for Music Information Retrieval Conference*. [S.l.: s.n.], 2009. p. 261–266.

GUBLER, D. J. Dengue, urbanization and globalization: the unholy trinity of the 21st century. *Tropical medicine and health*, Japanese Society of Tropical Medicine, v. 39, n. 4 Suppl, p. 3, 2011.

GUERRA, G. N. Modelo de reputação e ontologia aplicados à rede social científica do observeunb. 2012.

HAHMANN, S.; BURGHARDT, D. Connecting linkedgeodata and geonames in the spatial semantic web. In: *6th International GIScience Conference*. [S.l.: s.n.], 2010.

HAN, J. et al. Survey on nosql database. In: IEEE. *Pervasive computing and applications (ICPCA), 2011 6th international conference on*. [S.l.], 2011. p. 363–366.

HEER, J.; BOYD, D. Vizster: Visualizing online social networks. In: IEEE. *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*. [S.l.], 2005. p. 32–39.

HERSH, W. R. Healthcare data analytics. *Health Informatics: Practical Guide for Healthcare*, 2014.

HEYDON, A.; NAJORK, M. Mercator: A scalable, extensible web crawler. *World Wide Web*, Springer, v. 2, n. 4, p. 219–229, 1999.

HIRST, T. *Reshaping Horse Import/Export Data to Fit a Sankey Diagram*. 2013. Disponível em: <http://www.r-bloggers.com/arc-diagrams-in-r-les-miserables/>. Acesso em: 25 jun.2015.

HUANG, Y.-W. et al. Web application security assessment by fault injection and behavior monitoring. In: ACM. *Proceedings of the 12th international conference on World Wide Web*. [S.l.], 2003. p. 148–159.

JAGADISH, H. V. et al. Timber: A native xml database. *The VLDB Journal—The International Journal on Very Large Data Bases*, Springer-Verlag New York, Inc., v. 11, n. 4, p. 274–291, 2002.

JAIN, P. et al. Linked data is merely more data. In: *AAAI Spring Symposium: linked data meets artificial intelligence*. [S.l.: s.n.], 2010. v. 11.

JONES, M. T. Data science and open source. IBM, 2013. Disponível em: <http: //www.ibm.com/developerworks/library/os-datascience/>.

JONKERS, K.; TIJSSEN, R. Chinese researchers returning home: Impacts of international mobility on research collaboration and scientific productivity. *Scientometrics*, Springer, v. 77, n. 2, p. 309–333, 2008.

JÚNIOR, S. K.; CAROLO, M. D.; NEGRI, F. de. Impacto dos fundos setoriais sobre a produtividade acadêmica de cientistas universitários. *Estudos Econômicos (São Paulo)*, SciELO Brasil, v. 43, n. 4, p. 647–685, 2013.

KAWASHIMA, H.; TOMIZAWA, H. Accuracy evaluation of scopus author id based on the largest funding database in japan. *Scientometrics*, Springer, v. 103, n. 3, p. 1061–1071, 2015.

KLETTKE, M.; MEYER, H. Xml and object-relational database systems enhancing structural mappings based on statistics. In: *The World Wide Web and Databases*. [S.l.]: Springer, 2001. p. 151–170.

KOBLIN, A. Flight patterns. In: ACM. *ACM SIGGRAPH ASIA 2009 Art Gallery & Emerging Technologies: Adaptation*. [S.l.], 2009. p. 29–29.

KUMAR, S. et al. Tweettracker: An analysis tool for humanitarian and disaster relief. In: *ICWSM*. [S.l.: s.n.], 2011.

LANE, J. Let's make science metrics more scientific. *Nature*, Nature Publishing Group, v. 464, n. 7288, p. 488–489, 2010.

LEPORI, B.; PROBST, C. et al. Using curricula vitae for mapping scientific fields: A small-scale experience for swiss communication sciences. *Research Evaluation*, v. 18, n. 2, p. 125, 2009.

LEWISON, G.; KUNDRA, R. The internal migration of indian scientists, 1981–2003, from an analysis of surnames. *Scientometrics*, Akadémiai Kiadó, co-published with Springer Science+ Business Media BV, Formerly Kluwer Academic Publishers BV, v. 75, n. 1, p. 21–35, 2008.

MAYERES, I.; OCHELEN, S.; PROOST, S. The marginal external costs of urban transport. *Transportation Research Part D: Transport and Environment*, Elsevier, v. 1, n. 2, p. 111–130, 1996.

MCAFEE, A. et al. Big data. *The management revolution. Harvard Bus Rev*, v. 90, n. 10, p. 61–67, 2012.

MENA-CHALCO, J. P. et al. Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology*, Wiley Online Library, v. 65, n. 7, p. 1424–1445, 2014.

MENEZES, L. C. de et al. Dyscs: A platform to build geographically and semantically enhanced social content sites. *Journal of Systems and Software*, Elsevier, v. 94, p. 39–49, 2014.

MOED, H. F.; HALEVI, G. A bibliometric approach to tracking international scientific migration. *Scientometrics*, Springer, v. 101, n. 3, p. 1987–2001, 2014.

MOED, H. F.; PLUME, A. et al. Studying scientific migration in scopus. *Scientometrics*, Springer, v. 94, n. 3, p. 929–942, 2013.

MOROSINI, M. C.; SOUZA, A. A pós-graduação no brasil: formação e desafios. *Revista Argentina de Educación Superior*, v. 1, n. 1, p. 125–152, 2009.

OEC. *Learn More About Trade in Brazil*. 2015. Disponível em: <https://atlas.media.mit.edu/en/profile/country/bra/>. Acesso em: 25 jun.2015.

PALMER, S. *Data Science For The C-Suite*. [s.n.], 2015. Disponível em: <http://www.shellypalmer.com/shelly-palmer-bio/>.

PONDS, R.; OORT, F. V.; FRENKEN, K. The geographical and institutional proximity of research collaboration*. *Papers in regional science*, Wiley Online Library, v. 86, n. 3, p. 423–443, 2007.

POPESCU, A.; GREFENSTETTE, G.; MOËLLIC, P. A. Gazetiki: automatic creation of a geographical gazetteer. In: ACM. *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*. [S.l.], 2008. p. 85–93.

PROVOST, F.; FAWCETT, T. Data science and its relationship to big data and data-driven decision making. *Big Data*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 1, n. 1, p. 51–59, 2013.

RAE, A. From spatial interaction data to spatial interaction information? geovisualisation and spatial structures of migration from the 2001 uk census. *Computers, Environment and Urban Systems*, Elsevier, v. 33, n. 3, p. 161–178, 2009.

RATTI, C. et al. Redrawing the map of great britain from a network of human interactions. *PloS one*, Public Library of Science, v. 5, n. 12, p. e14248, 2010.

RESEARCH, E. *How Obama Won Re-election*. 2012. Disponível em: <http://www.nytimes.com/interactive/2012/11/07/us/politics/obamas-diverse-base-of-support.html?_r=0>. Acesso em: 25 jun.2015.

ROBERGE, G.; CAMPBELL, D. Canadian researchers migration analysis based on scopus author ids. In: *17th international conference on science and technology indicators (STI 2012), Montreal. http://sticonference. org/Proceedings/vol2/Roberge_Canadian_884. pdf*. [S.l.: s.n.], 2012.

ROBINSON, W. D. et al. Integrating concepts and technologies to advance the study of bird migration. *Frontiers in Ecology and the Environment*, Eco Soc America, v. 8, n. 7, p. 354–361, 2009.

ROMÊO-UNU, J. R. M.; IBMEC, C. I. M. R.-F. Estudos de pós-graduação no brasil. 2004.

ROONGPIBOONSOPIT, D.; KARIMI, H. A. Comparative evaluation and analysis of online geocoding services. *International Journal of Geographical Information Science*, Taylor & Francis, v. 24, n. 7, p. 1081–1100, 2010.

RUSHTON, G. et al. Geocoding in cancer research: a review. *American journal of preventive medicine*, Elsevier, v. 30, n. 2, p. S16–S24, 2006.

SANCHEZ, G. *Arc Diagrams in R: Les Miserables*. 2013. Disponível em: <http://www.r-bloggers.com/arc-diagrams-in-r-les-miserables/>. Acesso em: 25 jun.2015.

SANTOS, A. L. F. dos; AZEVEDO, J. M. L. de. A pós-graduação no brasil, a pesquisa em educação e os estudos sobre a política educacional: os contornos da constituição de um campo acadêmico. *Revista Brasileira de Educação*, SciELO Brasil, v. 14, n. 42, p. 535, 2009.

SANTOS, C. M. dos. Tradições e contradições da pós-graduação no brasil. *Educ. Soc*, SciELO Brasil, v. 24, n. 83, p. 627–641, 2003.

SCHICH, M. et al. A network framework of cultural history. *science*, American Association for the Advancement of Science, v. 345, n. 6196, p. 558–562, 2014.

SCOTT, J. *Social network analysis*. [S.l.]: Sage, 2012.

SEIDL, W. On the importance of scientific research in relation to humanities. In: *Drawing a Hypothesis*. [S.l.]: Springer, 2011. p. 215–219.

SENGAR, V. et al. Robust location search from text queries. In: ACM. *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*. [S.l.], 2007. p. 24.

SHAH, S. *Stegosploit: Hacking With Pictures*. HITB2015AMS Conference, 2015. Disponível em: <https://conference.hitb.org/hitbsecconf2015ams/sessions/stegosploit-hacking-with-pictures/>.

STEELE, J.; ILIINSKY, N. *Beautiful visualization: looking at data through the eyes of experts*. [S.l.]: "O'Reilly Media, Inc.", 2010.

STEINER, R. et al. Improving geocoding of traffic crashes using a custom arcgis address matching application. In: *22nd Environmental Systems Research Institute International User Conference: July 7-11 2003: San Diego, CA*. [S.l.: s.n.], 2003.

TATARINOV, I. et al. Storing and querying ordered xml using a relational database system. In: ACM. *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*. [S.l.], 2002. p. 204–215.

THELWALL, M. A web crawler design for data mining. *Journal of Information Science*, Sage Publications, v. 27, n. 5, p. 319–325, 2001.

UHLIR, P. F.; SCHRÖDER, P. Open data for global science. *Data Science Journal*, CODATA, v. 6, p. OD36–OD53, 2007.

VELLOSO, J. et al. *A pós-graduação no Brasil: formação e trabalho de mestres e doutores no país*. [S.l.]: Capes, 2003. v. 2.

VERHINE, R. E. Pós-graduação no brasil e nos estados unidos: uma análise comparativa. *Educação*, v. 31, n. 2, 2008.

VIEIRA, P. V. M.; WAINER, J. Correlações entre a contagem de citações de pesquisadores brasileiros, usando o web of science, scopus e scholar. *Perspect. ciênc. inf.[online]. Belo Horizonte*, SciELO Brasil, v. 18, n. 3, p. 45–60, 2013.

WALLER, M. A.; FAWCETT, S. E. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, Wiley Online Library, v. 34, n. 2, p. 77–84, 2013.

WANG, A. The shazam music recognition service. *Communications of the ACM*, ACM, v. 49, n. 8, p. 44–48, 2006.

WEBBER, J. A programmatic introduction to neo4j. In: ACM. *Proceedings of the 3rd annual conference on Systems, Programming, and Applications: Software for Humanity*. [S.l.], 2012. p. 217–218.

WINKLER, W. Overview of research linkage and current research directions. *US Bureau of the Census, Statistical Research Report Series RRS2006/02*, 2006.

WINKLER, W. E. Matching and record linkage. In: *Business Survey Methods*. [S.l.]: Wiley, 1995. p. 355–384.

YAMASHITA, Y.; YOSHINAGA, D. Influence of researchers' international mobilities on publication: a comparison of highly cited and uncited papers. *Scientometrics*, Springer, v. 101, n. 2, p. 1475–1489, 2014.

YU, A. Z. et al. Pantheon: Visualizing historical cultural production. In: IEEE. *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*. [S.l.], 2014. p. 289–290.

ZANDBERGEN, P. A. A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems*, Elsevier, v. 32, n. 3, p. 214–232, 2008.

ZELLNER, C. The economic effects of basic research: evidence for embodied knowledge transfer via scientists' migration. *Research Policy*, Elsevier, v. 32, n. 10, p. 1881–1895, 2003.

Apêndice A

Ferramentas Utilizadas para Geração de Gráficos

Existem muitas ferramentas com o propósito de gerar análises e visualizações de dados. No escopo desta pesquisa, sobre a mobilidade dos pesquisadores, algumas ferramentas foram selecionadas devido o seu potencial e agilidade em atender este propósito. Dentre estas as principias foram o QGis[1], o Gephi[2] e algumas bibliotecas da linguagem R[3].

A princípio, os dados com geolocalização eram exibidos através da ferramenta de geo-processamento QGis, devido a sua facilidade em carregar arquivos *shapefile*[4] e registros de localidade. A própria Figura 4.8 é um exemplo de utilização desta ferramenta, no qual a lista de registros foi importada na ferramenta através do formato csv[5] e visualizada juntamente com o *shapefile* dos estados brasileiro disponível no site do IBGE.

Também é possível identificar por meio da Figura 4.8 que o QGis permite gerar algumas personalizações quanto ao tamanho e cor dos pontos de acordo com alguma métrica associada. Este mesmo recurso também permite que se crie mapas do tipo *choropleth* no QGis. Contudo, a representação do grande volume de fluxos presente no Multigrafo de Mobilidade dos Pesquisadores (MMP) se demonstrou ser mais complexo na ferramenta.

[1] http://www.qgis.org/en/site/

[2] https://gephi.org/

[3] https://www.r-project.org/

[4] Este tipo de arquivo é utilizado para armazenar pontos, linhas e polígonos que geralmente compõe um mapa.

[5] Este tipo de arquivo tem como finalidade geral armazenar dados usando registros separados por quebra de linha e colunas por vírgula.

Então a primeira alternativa utilizada para exibição dos fluxos em escala foi a ferramenta Gephi, por causa do seu potencial em: processar e visualizar grafo de grande volume; dispor os nós de um gráfico que possui atributos de geolocalização em sua projeção efetiva[6]; além das suas funcionalidades de geração de métricas e visualizações personalizadas.

Para gerar as visualizações no Gephi, a princípio, era necessário importar dois arquivos no formato csv para informar os nós e arestas, que no MMP era fornecido pelas listas de registros de localidades e de fluxos agrupados com ponderação para evitar a visualização de arestas múltiplas. Depois da importação, o Gephi permitiu a geração de gráficos como as das Figuras 2.9, 5.7 e 5.9.

Entretanto, devido à problemas de integração do Gephi com arquivos *shapefile* é que se tentou utilizar uma outra alternativa de ferramenta que contemplasse tanto os recursos de tratamento de dados com geolocalização quanto a dos dados disponibilizados em grafo. Dentre as opções disponível a linguagem R se mostrou ser mais atraente, inicialmente, pela possibilidade de tratar grafos com o uso da biblioteca igraph[7], e dados georeferenciados com *shapefiles* através das bibliotecas: maptools[8], rgdal[9], geosphere[10], maps[11] e ggplot2[12]. A Figura 5.10 é um exemplo desta viabilidade.

Além disso, o uso de uma linguagem, e não um programa, permite um nível de personalização bem maior que as telas dos programas citados, devido a extensão de suas funcionalidades e vasta quantidade de bibliotecas. Por fim, através da linguagem R foi possível gerar outros inúmeros gráficos na pesquisa.

A título de demonstração, o Código Fonte A.1 exibe como que a linguagem R foi utilizada para gerar o Mapa de Carlos (MC) da Figura 2.11a. No código fica claro que existe o uso da biblioteca ggplot2 para exibir os dados extraídos dos registros de localidades coletados na linha 5, no qual: a função $ggplot()$ recebe um $dataframe$ junto com a indicação das colunas e linhas; a função $geom\_tile()$ define de onde será obtido a métrica que pondera a associação entra as colunas, ou seja, a essência do MC; a função $theme()$ estiliza o títlulo e as legendas; a função $xlab()$ 3 $ylab()$ define o nome das legendas; já $scale\_fill\_gradient()$ determina a

[6]Este recurso é possível graças ao GeoLayout (https://marketplace.gephi.org/plugin/geolayout/)
[7]http://igraph.org/r/
[8]https://cran.r-project.org/web/packages/maptools/index.html
[9]https://cran.r-project.org/web/packages/rgdal/index.html
[10]https://cran.r-project.org/web/packages/geosphere/index.html
[11]https://cran.r-project.org/web/packages/maps/index.html
[12]http://ggplot2.org/

escala de cores; $geom\_text()$ aponta o valor da associação; enquanto que $ggtitle()$ apresenta o título; por fim, a função $print()$ exibe o gráfico criado.

Código Fonte A.1: Geração do gráfico MC usando ggplot2.

```
1  library(ggplot2)
2
3  title <- getitle()
4  label <- c('origem','destino')
5  df <- getdata()
6
7  gg <- ggplot(df, aes(x=from, y=to))+
8    geom_tile(aes(fill = valor)) +
9    theme(
10     title=element_text(size=14,face="bold"),
11     axis.text=element_text(size=14,face="bold"),
12     axis.title=element_text(size=14,face="bold"),
13     axis.text.x=element_text(angle=-90)
14   ) +
15   xlab(label[1])+ylab(label[2]) +
16   scale_fill_gradient(low='white', high='grey20') +
17   geom_text(aes(fill = valor, label = round(valor, 1))) +
18   ggtitle(title)
19
20 print(gg)
```

Mas a biblioteca ggplot2, utilizada na demonstração, também foi utilizado para criar os seguintes gráficos:

- *Boxplot* na Figura 4.5b através da função $geom\_boxplot()$;
- Gráfico de barra na Figura 4.6a usando $geom\_bar()$;
- Histograma na Figura 5.2a com a função $stat\_bin()$;
- Mapa de dispersão na Figura 5.2b por meio da função $geom\_point()$.

Mais bibliotecas também foram utilizadas para geração de outros tipos de gráficos que não são contemplados no ggpplot, como o circlize[13] para gerar o Mapa de Fluxo Circular (MFC) e o treemap[14] para fazer mapas do tipo $treemap$.

[13]https://cran.r-project.org/web/packages/circlize/index.html
[14]https://cran.r-project.org/web/packages/treemap/index.html