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

Luiz Carlos R. Chaves[1], Alexandre N. Duarte[2]

[1]Unidade de Informática – Instituto Federal de Ciência, Edu. e Tec. da Paraíba – IFPB

[2]Centro de Informática – Universidade Federal da Paraíba – UFPB

luiz.chaves@ifpb.edu.br, alexandre@ci.ufpb.br

Abstract. *Researcher's mobility is one the key factors for science advancement. In this study we analyze the mobility patterns of the researchers active in the Brazilian scientific community or formed by scientific and education institutions through the extraction and treatment of location-dependent metrics recorded in their curriculum vitae. With this analysis we could identify the most influential formation centers for the Brazilian scientific community and the network that connects these formation centers to the researchers workplaces.*

Resumo. *A mobilidade de pesquisadores é um dos fatores que afetam significativamente o progresso da ciência. Neste estudo analisamos os padrões de mobilidade observados entre os pesquisadores atuantes ou capacitados no Brasil em instituições de ensino e pesquisa através da extração e tratamento de métricas dependentes da localidade presentes em seus registros curriculares. Com esta análise pudemos 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.*

1. Introdução

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 obter inovação e melhorias na logística produtiva para assim conquistar melhor competitividade [Seidl 2011].

Neste contexto, inúmeros são os focos de análise que exploram e apontam as correlações e motivações desta evolução [Catalá-López et al. 2012, Kannebley Junior et al. 2013, Lima et al. 2013, Mena-Chalco et al. 2014]. Mas neste estudo pretendemos analisar os padrões de mobilidade observados entre os pesquisadores atuantes ou formados no Brasil de modo a identificar os principais centros de formação e atuação para avaliar como é a influência destes centros, inclusive descrevendo como os deslocamentos variam ao longo do tempo.

Para possibilitar tal análise se faz necessário identificar os locais que fizeram parte da jornada científica do pesquisador através de informações disponíveis em seu currículo vitae. Contudo, a tarefa de obter tais informações sobre os pesquisadores nem sempre é uma questão trivial devido ao padrão e consistência dos dados declarados para

o processamento de localidades, isso quando tais dados são disponibilizados [Mena-Chalco et al. 2014].

No Brasil uma importante base de registro de atividades de pesquisa está acessível na plataforma Lattes do CNPq. 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 informação 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. Tal escolha implicou na elaboração de uma metodologia de extração dos dados disponibilizados na plataforma, com acesso eficiente e eficaz aos dados para assim gerar interpretações dos deslocamentos.

A metodologia de extração prevê a utilização de descrição estatística das métricas espaciais, temporais e de ligação dos fluxos de mobilidade como forma de facilitar a identificação de padrões de deslocamento entre os estados brasileiros e entre diferentes países. Além disso, possibilita identificar o grau de importância das grandes universidades e cidades brasileiras na formação de pesquisadores e profissionais ativos.

O restante deste trabalho encontra-se estruturado da seguinte forma: a Seção 2 descreve uma breve fundamentação teórica. A Seção 3 apresenta a metodologia utilizada no desenvolvimento do estudo. Na Seção 4 é apresentada os trabalhos relacionados. A seção 5 expõe a descrição e tratamento dos dados. Já a Seção 6 expõe os resultados encontrados. Concluindo o trabalho, a Seção 7, apresenta as considerações finais da pesquisa e propostas para trabalhos futuros.

2. Fundamentação

Esta seção apresenta a concepção de mobilidade, abordando os contextos de visualização, algumas possibilidades de extração da informação e métricas utilizadas. E para o escopo desta pesquisa consideram-se como pontos de mobilidade apenas as localidades de nascimento, formação acadêmica e atuação profissional declaradas pelos pesquisadores.

Por meio da ligação destas localidades informadas por todos os pesquisadores é que é possível compor um grafo de arestas direcionadas nomeado de Grafo de Mobilidade (GM), no qual os nós, representando os pesquisadores, podem ser visualizados através do contexto de sua instituição, cidade, estado, região, país ou continente, enquanto que as arestas são representadas pelo agrupamento dos fluxos de cada pesquisador entre os nós de um contexto específico, sendo direcionado e ponderado pela quantidade de fluxos existentes no percurso.

Por exemplo, quando coletamos a mobilidade que cada pesquisador com doutorado declarou no contexto da cidade se compõe o GM da Figura 1(a), no qual os nós são compostos pelas cidades que possuem Registro de Nascimento (RN), Formação (RF) e Trabalho (RT). Já as arestas são compostas pelo agrupamento de 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. Entretanto esses fluxos possuem direcionamento orientado dependendo da ordem cronológica de tais eventos em cada pesquisador, iniciando pela cidade com RN, passando pelas cidades com RF e terminando com a cidade com RT.

Portanto, vale salientar que por questões de visualização quando existe múltiplos deslocamentos no sentido de uma cidade para outra é processado apenas uma aresta ponderado com o número de ocorrências de cada um dos três tipos de fluxo existente no percurso. Por exemplo, no GM da Figura 1 (a) existe uma aresta da cidade de São Paulo para o Rio de Janeiro, mas totalizam 72.905 fluxos sendo 775 FNF, 41.805 FFF e 30.325 FFT.

Agora quando se considera o contexto dos estados o GM é computado apenas com fluxos que acontecem entre estas entidades, compondo os nós com os estados que possuem RN, RF e RT e as arestas com FNF, FFF e FFT entre as entidades envolvidas, o que resulta na Figura 1 (b). O mesmo procedimento pode ser aplicado para os demais contexto de país, continente, região e instituição.

A princípio os nós do GM são posicionados geograficamente para facilitar as interpretações junto ao mapa do contexto, mas é possível interpretá-los sem a associação espacial. Além disso, outra possiblidade consiste em correlacionar ou filtrar os fluxos usando algum critério específico. Então seria possível extrair métricas de mobilidade associadas as pessoas de uma área de formação, como ciência da computação; compor o GM apenas com pesquisadores oriundos de estados que possuem baixos índices de PIB e IDH; ou até mesmo limitar que as arestas sejam ponderadas com fluxos de apenas um tipo, como o FFF.

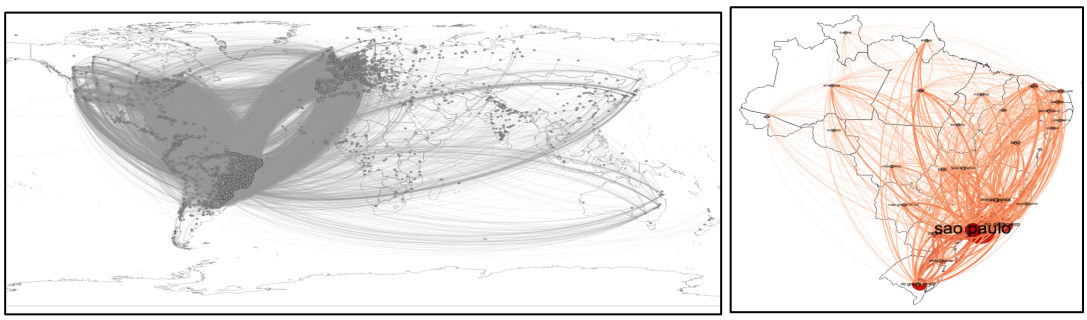

Figura 1 – Mapa da mobilidade dos doutores cadastrados na plataforma Lattes no contexto entre (a) cidades e (b) estados brasileiros.

Para facilitar a visualização do GM, em alguns casos, se faz necessário mudar o tamanho da escala ou opacidade das arestas, ou até mesmo alterar o tamanho e a cor dos nós, arestas e textos descritivos. Por exemplo, o fluxo mais intenso no Brasil se concentra na região sudeste, visível por meio das cores e larguras de aresta mais intensas nesta região na Figura 1 (b), além de sua alta densidade de arestas; o estado que possui mais ligações é São Paulo, indicado pelo nó com maior texto descritivo na mesma figura.

Porém, outras informações só ficam mais aparentes quando extraímos algumas métricas do GM, e.g., no contexto dos estados brasileiros, da Figura 1 (b), apenas 11 destes não possuem conexões de entrada e saída para todos os demais estados e os 9 maiores fluxos entre eles ocorrem sem sair de si mesmo; ou que a cidade de maior fluxo é São Paulo na Figura 1 (a).

Devido a dificuldade de interpretar informações ocultas ou embaralhadas visualmente no GM 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 filtrados por critérios específicos.

Através da aferição da ligação do grafo é possível analisar os fluxos para indicar sua intensidade e padrão. Então é possível se extrair entre os vários contextos de nó o seu Grau (GR) que consiste no número de ligação incidentes em si mesmo, seja ele de entrada (GR_E) ou saída (GR_S), ponderada em alguma escala, como o logaritmo ($GR_{[log]}$), ou não.

Mas apesar da composição coletiva das arestas do GM existe um aspecto espacial e de individualidade nos fluxos, pois cada pessoa contribui com localidades em todo o seu trajeto. Logo é possível analisar a partir de um indivíduo específico o Número de Deslocamento (ND) e a Distância do Deslocamento (DD).

O ND conta quantos saltos foram contabilizados no fluxo de um indivíduo podendo também ser exibido em termos de ND distintos (ND_d) que exclui da contagem deslocamentos já realizados. Implicitamente o ND_d revela o Número de Locais (NL) distintos envolvido no fluxo quando se adiciona mais um ao ND_d. Já a DD mede a extensão em quilômetros (Km) do deslocamento sendo contabilizado de forma Instantânea (DD_I) ou Total (DD_T), ou seja, que considera respectivamente um percurso específico ou todo o trajeto de deslocamento do pesquisador.

Por fim, existe o aspecto temporal aplicada as formações, pois os curso no Lattes geralmente são registrados com sua duração, portanto 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), que neste caso é Instantâneo (ID_I). Mas também se poderia calcular o ID Total (ID_T) que consiste no intervalo da primeira formação até a última. Outra possibilidade de medida temporal consiste no Número de Instância (NI) contabilizados num intervalo de tempo. E pode ser expresso de forma Instantânea (NI_I) ou Total (NI_T). Por exemplo, é possível determinar o NI de doutores que se formam num ano específico ou nas últimas décadas.

3. Metodologia

Para possibilitar a utilização de métricas e visualização do GM primeiro é necessário pensar num modelo de extração e tratamento de dados existentes na plataforma Lattes.

Claro que o foco é a extração de dados contento localização, mas o modelo deve suportar a extração de qualquer dado do currículo, já pensando em suportar futuras necessidades de outros dados. Além disso, eles devem ser acessíveis de forma local sem a dependência dos servidores do Lattes, precavendo qualquer mudança da estrutura e acesso da plataforma no momento da análise, e até dependência de disponibilidade do serviço.

No sentido de idealizar algum formato de armazenamento dos dados curriculares foi analisado a estrutura do Lattes. Todos os currículos estão sob domínio do CNPq, disponíveis de forma pública em páginas HTML e arquivos XML na Web.

Atualmente o acesso direto aos dados só é possível mediante um ofício à presidência do CNPq, 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. Sendo este um processo burocrático e potencialmente demorado, pesquisadores têm buscado alternativas para extrair dados da plataforma lattes, copiando os dados a partir dos formatos públicos existentes.

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. Contudo 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.

Portanto através desta problemática é que foi definido o Modelo de Mineração de Dados (MMD), conforme a estrutura e fluxo da Figura 2 organizado em 3 etapas inspirado nos conceitos de mineração do *Knowledge Discovery In Databases* que servem para extrair e auxiliar nas interpretações de mobilidade realizando:

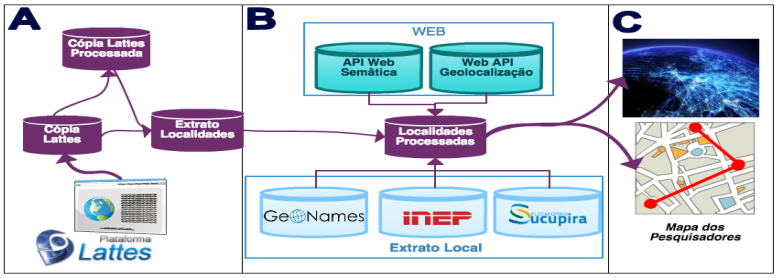

Figura 2 – Fluxo de extração de informação do MMD.

A. Obtenção dos dados: consiste na extração dos dados de localidade contidos nos XMLs dos pesquisadores na plataforma Lattes. Nesta etapa foram criadas as bases "Cópia Lattes" contento a cópia de todos os XMLs da plataforma Lattes, "Cópia Lattes Processada" que possui os XMLs distribuídos em tabelas estruturadas, e "Extrato de localidades" que contém apenas os registros de localidades;

B. Tratamento dos dados: consiste no ajuste dos dados extraídos realizando possíveis remoções de variáveis, valores ausentes e redundantes. Também foi realizado a normalização e inferência das localidades através de bases remotas, de geolocalização e ontológicas como o Google Maps e Freebase, e do Extrato Local do MMD, que possui nomes de cidades advindos do GeoNames[1] e IBGE, além de instituições por meio do INEP[2] e Sucupira[3]. Por fim se atribuiu a coordenada a cada registro de localidade dos pesquisadores que foi normalizado, o que resultou na base "Extrato Localidades";

C. Análise dos dados: avalia os dados extraídos juntamente com a descrição estatística. E também gera as métricas e visualizações gráficas sobre os

[1] http://www.geonames.org/

[2] http://portal.inep.gov.br/basica-levantamentos-acessar

[3] https://sucupira.capes.gov.br/sucupira/

deslocamentos. Todos os possíveis dados e resultados poderão ser disponibilizados futuramente numa aplicação Web pública.

4. Trabalhos Relacionados

Como forma de comparação com os trabalhos relacionados foram pesquisados alguns aspectos correlatos. Entre os mais evidentes se pode elencar os trabalhos que abordam a forma como se visualiza a mobilidade, as técnicas de extração dos registros curriculares do Lattes e as principais analises realizadas sobre esta base.

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 de TI, 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 [Dhar 2013].

Para que o processo de engenharia e manipulação dos dados, e extração de informação se torne possível muitas áreas da computação acabam se integrando a outras áreas como matemática e estatística.

Já em relação a análise de dados que envolvem localização existe um desafio devido a relação geoespacial das informações, que exige uma exibição por meio de mapas, uma vez que informações tabulares e sem organização podem não permitir boas interpretações neste cenário. Além dos mapas outros formatos de visualização podem auxiliar na compreensão da mobilidade como o apresentado por Abel e Sander (2014) através do diagrama de corda para caracterizar o fluxo da migração internacional, Balcan et al. (2009) com diagramas de Vonoroi, ou até junção de mapas com gráficos convencionais, como o gráfico em barra ou pizza localizado em pontos do mapa.

Alguns trabalhos já envolvem a análise de dados com localidades, como em 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. Seu resultado possibilita estimar a probabilidade de habitantes de um determinado local possuírem ou não a doença, contribuindo com o processo de intervenção precoce. Já Kumar et al. (2011) utiliza dados de redes sociais públicas interpoladas em mapas para auxiliar socorristas e a defesa civil no amparo de desastres naturais.

O foco deste trabalho se fundamenta na questão de mobilidade dos pesquisadores, e apresenta aspectos semelhantes aos encontrados nos mapas de análise de densidade e conflito de rotas da aviação civil elaborados por Koblin (2009), na escolha das melhores rotas de avião por Frey e Dueck (2007) e na análise do deslocamento entre o local de nascimento e morte de um conjunto de pessoas importante da história mundial feito por Schich et al. (2014).

Vale salientar que tais visualizações de mobilidade podem extrapolar a essência do deslocamento fornecendo outros tipos de informações correlatas. Por exemplo, Schich et al. (2014) ao rastear o deslocamento de algumas pessoas notáveis na história mundial exibiu que fatos históricos, como a formação dos grandes centros europeus e a colonização norte americana, estavam envolvidos com a mobilidade destas pessoas. Portanto, este trabalho inicialmente consiste em descrever e visualizar a mobilidade dos pesquisadores, mas inevitavelmente isto poderá ajudar a identificar outros fatos

associados como o destaque dos grandes centros políticos e econômicos na polarização das pesquisas. Contudo, analisar a causalidade destes fatos não será alvo desta pesquisa.

Em relação a extração e interpretação de dados da plataforma Lattes existem alguns trabalhos correlatos que disponibilizam publicamente suas ferramentas de extração, como o scriptLattes [Mena-Chalco et al. 2014] e o LattesCrawler [Wanderley et al 2014], que possibilitam a geração de relatórios ou coleta de dados curriculares para fins específicos. Já outras alternativas são mais fechadas porém com finalidades semelhantes, como a de Araújo et al. (2014), que consegue extrair dados para analisar a rede de coautoria e métricas de produtividade. Em muitos casos estas redes de coautoria extraídas a partir do Lattes tentam correlacionar suas métricas de produção com outros parâmetros para indicar a relevância de suas produções [et al. 2013] ou probabilidade de se adquirir bolsa [Wainer et al. 2013].

Considerando outras bases curriculares ou até mesmo bases de publicações de pesquisa percebe-se que existem ideias de extração e análises semelhantes às que já são feitas tomando o Lattes como fonte de dados. Em Mendonça Júnior et al. (2014) existe a utilização de um programa para extração de publicações de portais de periódicos para identificar publicações relevantes. Ponds et al. (2007) utiliza dados de publicações para analisar a relação das rede de colaboração com a proximidade geográfica dos seus envolvidos, já Yan (2012) investiga as similaridades entre diferentes redes acadêmicas. Em alguns casos tenta-se explorar o percurso evolutivo do pesquisador mas ainda usando métricas de produtividade como em Wu et al. (2013), ou em Yang et al. (2010) que investiga a relação entre algumas instituições de pesquisa porém sem considerar a relação geográfica.

5. Descrição e Tratamento dos dados

Esta seção descreve o conteúdo extraído da base do Lattes, além do tratamento e inferência das localidades obtidas a partir dos currículos. E como forma de maximizar o êxito nas interpretações e hipótese 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 em novembro 2014 foi possível contabilizar 3.911.585 indivíduos, sendo 99% destes de brasileiro. Já o tamanho total da base acumula 149 GB, mas metade dos currículos têm tamanho inferior a 8,6 KB. Neste aspecto percebe-se que currículos pequenos podem ser associados às pessoas no início de formação. Aproximadamente 75% dos currículos possuem até o título de especialização e o número de formações aumenta de modo gradativo com a titulação, ou seja, os pós-doutores são as pessoas que possuem a maior média de registro de formação.

A princípio se utilizou o número de formação e tamanho dos dados coletados para selecionar uma possível amostra com significância para as análises de mobilidade do MMD. Diante da característica da concentração de formação pela titulação optou-se por considerar apenas os currículos de doutores e pós-doutores, isso porque esta amostra, de 221.898 currículos, concentra mais localidades se comparada com outras possíveis amostras possuindo média de 6,3 Registros de Localidades (RL) por currículo, 67% da amostra com algum RT e 89% com RN. No entanto, mediante ao êxito das análises desta amostra é que seria possível ampliar o estudo para toda a base.

Usando esta amostra foi possível extrair aproximadamente 1,3 milhão de RL. Porém como já foi posto, estas informações não determinam o posicionamento geográfico com coordenadas, seja ela nome da cidade ou instituição. Logo, a princípio, o MMD para resolver esta indefinição utilizou o Extrato Local com coordenadas de 27.729 cidades e 2.374 instituições para comparar com os RL extraídos. Contudo a inserção manual dos RL no cadastro gera falta de padronização das localidades exigindo que se aplique alguma normalização antes do processo de comparação. No MMD a normalização funcionou com um simples processo de remoção de caracteres especiais.

Contudo mediante a ausência de associação literal do RL no Extrato Local se faz necessário uma consulta às bases geográficas ou semânticas para determinar as coordenadas, como descrito na metodologia. Mas mesmo assim ainda existe a possibilidade de não se encontrar o posicionamento para todo RL devido a eventuais erros de grafia. Tal fato exigiu uma outra estratégia para otimizar a convergência das coordenadas.

Finalmente, com a análise da distribuição de frequência dos nomes das instituições, sem a normalização, percebeu-se que a sua distribuição apresentava um padrão no qual 80% dos RL se concentravam nos primeiros 1.000 nomes de maior frequência. Portanto, diante dessa circunstância, incialmente, foi decido inferir no MMD apenas 1% dos nomes de maior frequência. O mesmo padrão se repetiu para os nomes de cidades exigindo a mesma estratégia. Resultado, 93% dos RL obtiveram suas coordenadas e foram usadas na analise de mobilidade do presente estudo.

6. Resultados e Discussões

Por meio das localidades inferidas e sua aplicação entre os vários contextos e tipos de fluxos citados é que foi viável compreender algumas características de mobilidade dos pesquisadores. A seguir são exibidos alguns padrões relevantes identificados através das análises realizadas.

O FNF mostra que a origem dos doutores é 95% de brasileiros, o que se aproxima da proporção do valor geral, e a maioria se localiza entre partes das regiões Sul, Sudeste e Nordeste do Brasil, ao passo que no exterior a maioria das origens advêm da Europa e América do Norte. Também foi possível identificar que 40% das primeiras formações estão presentes na cidade de origem do doutor.

A USP, UFRJ, UNESP, UNICAMP e UnB são as universidades que mais atraem os doutores no início de formação em termos absolutos, atraem cerca de 53% dos doutores. Coincidentemente estas instituições estão localizadas em regiões de grande concentração habitacional em relação as demais e possui grande relevância no contexto da pesquisa nacional. Mas se analisarmos a origem das pessoas atraídas por essas instituições vemos que nem todas elas vieram da cidade da instituição, e.g., no caso da 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 atinge-se um valor de 40%.

Já o FFT mostra que 98% dos RT estão no Brasil, no qual a grande maioria dessas ocorrências se localizam nas regiões Sudeste, Sul e Nordeste. Agora no FFT que envolve outros países, existe mais uma tendências das pessoas que fizeram formação

fora virem trabalhar aqui no Brasil do que o inverso, pois boa parte desse fluxo são de pessoas que se formaram na América do Norte e Europa e que vem 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 é a USP, UFRJ, UNESP, UNICAMP e UFRGS. Em termos absolutos representa 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 instituição percebe-se que quase que a totalidade dos fluxos são de instituições que não absorvem seus próprios doutores concluintes.

O total de fluxos entre os doutores contabiliza 520.397 deslocamento entre 1.779 instituições distintas, porém quando analisamos apenas os fluxos de formação no contexto das instituições, excluindo fluxos entre os RN e RT, percebe-se que o número de fluxos reduz em 53,6%. No entanto o fluxo da maioria das pessoas, 65%, chega a passar por no máximo duas instituições distintas entre as várias formações, com NL igual a 2. Já 27% tem o NL igual a 1, *i.e.*, ausência de deslocamento com DD_T de 0 Km.

Na Figura 3 (a) boa parte do deslocamento, em metros, é feito de pequenas distâncias, *e.g.*, 67% dos fluxos apresentam DD_I igual a zero, e 87% não ultrapassam 1.000 Km. Se analisarmos a distribuição do DD_I na escala temporal, por meio da Figura 3 (b), percebe-se que a partir da década 70 os deslocamentos mantiveram um padrão, sendo alterado apenas sua intensidade entre os fluxos. Porém, na figura a diminuição de pontos nos últimos anos ocorrer porque esse período não computou a formação pré-doutorado de futuros doutores, constando apenas a formação dos concluintes até 2013.

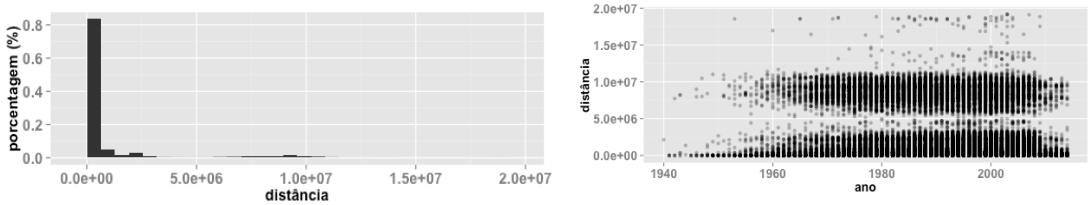

Figura 3 – (a) Histograma em valores percentuais de todas distâncias de deslocamento e (b) Dispersão da distância do deslocamento das formações através do tempo.

Analisando a Figura 3 (b) também fica evidente que existe dois fluxos intensos, na parte horizontal inferior e intermediário, e um fluxo menos intenso, na parte superior. Por meio da DD_I no contexto do país, visível na Figura 1(a), percebe-se que a América do Norte e Europa são destinos intensos, e que as medidas de DD_I para essas duas regiões explicam a densidade de pontos intermediários da Figura 3 (b). A região inferior, a de baixa DD_I, compreende quase que a totalidade dos fluxos e é justificado pelos deslocamentos feitos principalmente no Brasil. Enquanto que o deslocamento superior, de alta DD_I, compreendo os fluxos para o hemisfério leste que são de baixa intensidade.

Outra possibilidade de visualização da mobilidade, especificamente para analisar o GR_S e GR_E entre os nós, pode ser visto através do Mapa de Calor (MC) ilustrado na Figura 4 (a). Esta visualização que está no contexto do continente entre os FFF intensifica o número de deslocamento de saída de um nó de uma forma mais direta e aparente, usando variação de cores na matriz de associação entre os continentes sendo o sentido do fluxo expresso do eixo da abscissa à ordenada.

O GM pode até exibir esta mesma informação, mas é difícil definir qual é o sentido e intensidade das arestas devido ao seu formato com ausência de setas em sua extremidade, e dependência de posicionamento espacial gerando nos casos de alta densidade de fluxo impossibilidade de localização das arestas.

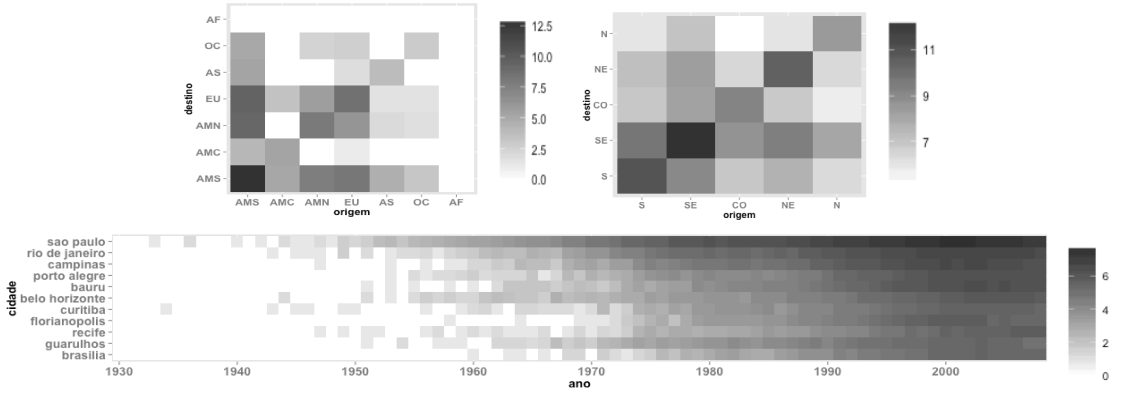

Figura 4 - MC do log do número de (a) fluxos de formação entre os continentes, (b) fluxos de formação entre as regiões do Brasil e (c) doutores formados deste 1930 até 2008 em algumas cidades com maiores registros de fluxo.

Então a Figura 4 (a) contorna essa dificuldade do GM e exibe de forma clara que boa parte dos fluxos entre os continentes se compreende na América do Sul (AMS), Europa (EU) e América do Norte (AMN) pelas cores mais intensas. Uma possível associação para a escolha dos dois últimos continentes, para quem vai fazer a formação no exterior, pode ser associado pelo fato de que os grandes centros de pesquisa mundialmente conhecidos, da maioria das áreas, estão localizados nestas regiões.

Mas um fato peculiar é que nenhum registro foi contabilizado na África (AF). Uma possível explicação para essa ausência de deslocamento pode ser dada porque 7% dos RL não foram usados nesta análise, ou seja, isso não anula diretamente deslocamentos para AF, mas pode sugerir que possivelmente existe algum deslocamento para ele contudo deve possuir valores baixos em relação ao total devido ao padrão de frequência dos RL descrito na seção 5.

O FFF estrangeiro apresenta uma centralização no Brasil com maior GR_S em relação ao GR_E, e totaliza 9% dos fluxos envolvendo os 30% de instituições do Extrato Local que são estrangeiras. Já o FFF nacional, que afeta a maioria dos RL, possui uma característica em comum. Independente do contexto da região, estado ou cidade, geralmente, os fluxo concêntricos, de DD nulo, se destacam mais que os externos, com DD maior que zero. Por exemplo, se comprarmos a mobilidade entre as regiões brasileiras na Figura 4 (b) percebe-se que o deslocamento na própria região, representada pela coluna secundária, é maior do que entre regiões distintas. Fato que também ocorre para todos os estados e na maioria das cidades com grande GR, e reforça mais ainda que a preferência das pessoas pelos menores percursos representa uma característica relevante na mobilidade dos pesquisadores.

Outra questão no aspecto temporal é que o número de doutores vem se caracterizando com uma tendência de crescimento sempre maior do que a escala do crescimento população do Brasil. E uma forma de quantificar essa evolução pode ser vista na Figura 4 (c) através do NI_I de doutores formados desde 1930 nas 11 cidades de maior GR. Quanto ao ID_I de cada pesquisador geralmente é compatível com a duração do curso, e se considerarmos o ID_T existe uma média de duração de 13 anos.

Por fim, outra possibilidade de análise de mobilidade é no sentido de permitir comparações entre nós usando a ideia do raio de abrangência calculado pelo GR_S ou GR_E. Por exemplo, se compararmos a USP e a UFPE pelo número de instituições que forneceram pessoas para seus programas de doutorado percebe-se que a USP atrai mais pessoas de outros instituições do que a UFPE, respectivamente com os raios de abrangência apresentando um GR_E de 295 e 85. O legal deste número é que ele poderia servir como possível candidato a identificar o grau de importância de um nó na rede de mobilidade, ou 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.

7. Conclusão

A corrente metodologia de análise de mobilidade dos pesquisadores através dos registros de localidade permitiu a caracterização de algumas questões relevantes, como a preferência de deslocamento curtos pela maioria dos doutores. Quanto ao método de extração de dados nas bases de registros públicos na Web, o presente trabalho forneceu uma importante contribuição por propor uma abordagem de coleta genérica de dados e o disponibilizando independente da fonte, podendo ser aplicado a diversos tipos de dados e bases. Ou seja, 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 provento interfaces de consulta local independente da base original.

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

Com isso várias possibilidades de trabalho surgiram permitindo: correlacionar e visualizar as relações de deslocamento espacial da formação e atuação dos pesquisadores com métricas tradicionais de avaliação da qualidade das instituições de ensino e pesquisa no Brasil, e no Mundo; observar se a evolução do número de formação de doutores nas cidades possui correlação com seus indicadores socioeconômicos, como o IDH-M e PIB; ou até mesmo investigar os padrões de mobilidade na escala temporal para indicar tendências. Finalmente existe também a possibilidade de adaptar o MMD para utilizar outros localidades presentes nos registros curriculares, tal como as publicação em eventos, para ampliar a ideia de mobilidade.

Contudo mais importante que essas possibilidades de análise são as visualizações e métricas geradas neste trabalho, pois concedem uma boa estratégia para descrever e quantificar a mobilidade entre os pesquisadores.

Referência

Abel, G. J., Sander, N. (2014) “Quantifying global international migration flows” Science, v. 343, n. 6178.

Araújo, E. B. et al. (2014) “Collaboration Networks from a Large CV Database: Dynamics, Topology and Bonus Impact” PLoS ONE, v. 9, n. 3.

Balcan, D. et al. (2009) “Multiscale mobility networks and the spatial spreading of infectious diseases” Proceedings of the National Academy of Sciences, v. 106, n. 51.

Barker, L. E. et al. (2013) “Bayesian small area estimates of diabetes incidence by United States county” Journal of Data Science, v. 11, n. 2.

Catalá-López, F. et al. (2012) “Coauthorship and Institutional Collaborations on Cost-Effectiveness Analyses: A Systematic Network Analysis” PLoS ONE, v. 7, n. 5.

Dhar, V. (2013) “Data science and prediction” Communications ACM, v. 56, n.12.

Frey, B. J., Dueck, D. (2007) “Clustering by passing messages between data points” Science, v. 315, n. 5814.

Kannebley Junior, S. et al. (2013) “Impacto dos Fundos Setoriais sobre a produtividade acadêmica de cientistas universitários” Estudos Econômicos, v. 43, n. 4.

Koblin, A. (2009) “Flight patterns” ACM SIGGRAPH ASIA 2009 Art Gallery & Emerging Technologies: Adaptation.

Kumar, S et al. (2011) “TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief” ICWSM.

Lima, H. et al. (2013) “Aggregating productivity indices for ranking researchers across multiple áreas” JCDL, v. 13.

Mena-Chalco et al. (2014) “Brazilian bibliometric coauthorship networks” JASIST, v. 65, n. 7.

Mendonça Junior, M. L. et al. (2014) “Uma Ferramenta para Extração Semiautomática e Análise de Relevância de Artigos Científicos” Brasnam.

Schich, M. et al (2014)“A network framework of cultural history”Science,v.345,n.6196.

Seidl, W. (2011) “On the Importance of Scientific Research in Relation to Humanities” Drawing a Hypothesis, Springer.

Ponds, R., Oort, F. V., Frenken, K. (2007) “The geographical and institutional proximity of research collaboration” Papers in Regional Science, v. 86, n. 3.

Vieira, P. V. M., Wainer, J. (2013) “Correlações entre a contagem de citações de pesquisadores brasileiros, usando o Web of Science, Scopus e Scholar” Perspectiva em Ciência da Informação, v. 18, n. 3.

Wainer, J., , P. (2013) “Avaliação de bolsas de produtividade em pesquisa do CNPq e medidas bibliométricas: correlações para todas as grandes áreas” Perspectiva em Ciência da Informação, v. 18, n. 2.

Wanderley, A. J. et al. (2014) “Identificando correlações entre métricas de Análise de Redes Sociais e o h-index de pesquisadores de Ciência da Computação” BraSNAM

Wu, M. Q. Y. et al. (2013) “Visual exploration of academic career paths” ASONAM.

Yang, B., Liu, Z., Meloche, J. A. (2010) “Visualization of the Chinese academic web based on social network analysis” JIS, v. 36, n. 2.

Yan, E. (2012) “Scholarly Network Similarities: How Bibliographic Coupling Networks, Citation Networks, Cocitation Networks, Topical Networks, Coauthorship Networks, and Coword Networks Relate to Each Other” JASIST, v. 63, n. 7.